European Community on Computational Methods in Applied Sciences

ECCOMAS Congress 2016

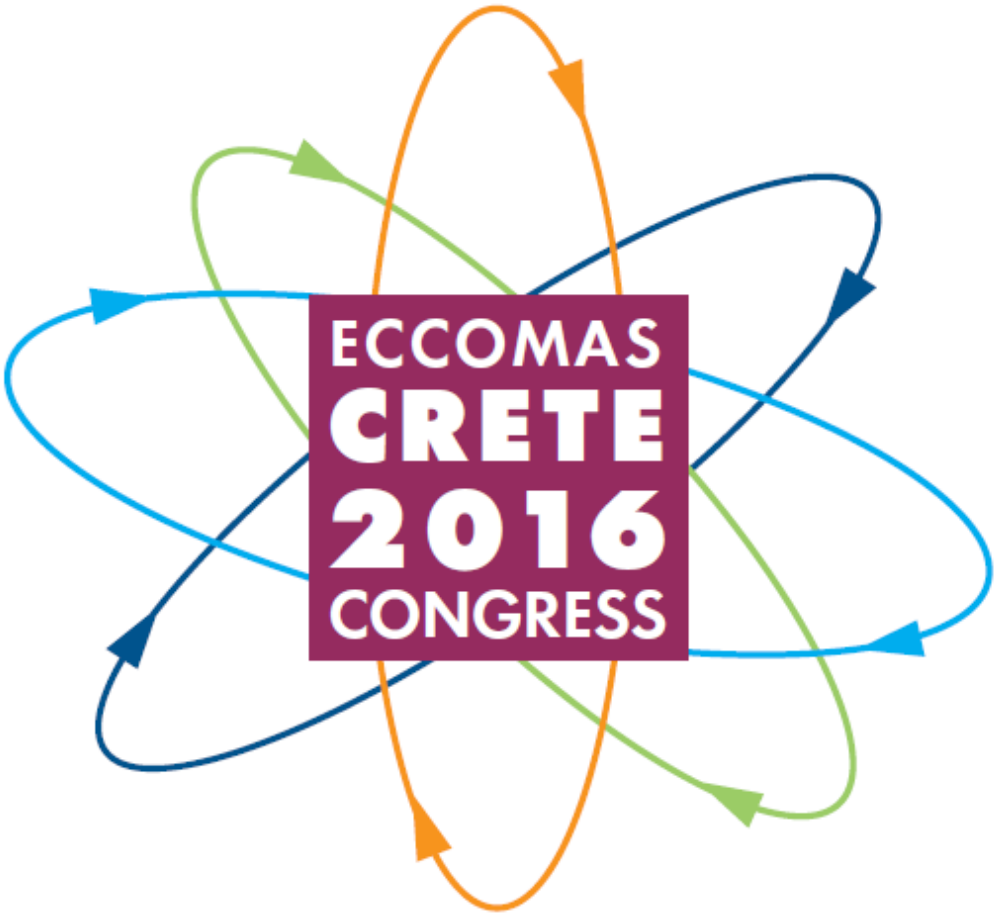
VII European Congress

on Computational Methods in Applied Sciences and Engineering

PROCEEDINGS

Volume II

M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds)



ECCOMAS Congress 2016

Proceedings of the VII European Congress on Computational Methods
in Applied Sciences and Engineering
Held in Crete, Greece
5-10 June 2016

Edited by:

M. Papadrakakis
National Technical University of Athens, Greece

V. Papadopoulos
National Technical University of Athens, Greece

G. Stefanou
Aristotle University of Thessaloniki, Greece

V. Plevris
Oslo and Akershus University College of Applied Sciences, Oslo, Norway

A publication of:

Institute of Structural Analysis and Antiseismic Research
School of Civil Engineering
National Technical University of Athens (NTUA)
Greece

ECCOMAS Congress 2016
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds)

First Edition, September 2016

© The authors

ISBN: **978-618-82844-0-1**

PREFACE

This volume contains the full-length papers presented at the **VII European Congress on Computational Methods in Applied Sciences and Engineering** (ECCOMAS Congress 2016) that was held on June 5-10, 2016 on the Crete Island, Greece.

The main objective of the quadrennial ECCOMAS Congresses is to provide a forum for presentation and discussion of state-of-the-art advances in computational methods in applied sciences and engineering, including basic methodologies, scientific developments and industrial applications and to serve as a platform for establishing links between research groups of academia and industry with common as well as complementary activities. About 2,200 papers were presented at the ECCOMAS Congress by authors from 53 countries around the world. This volume, consists of 667 full length accepted papers which will be indexed by SCOPUS database with access to the pdf file of the paper.

The ECCOMAS Congress 2016 is organized by the Institute of Structural Analysis and Antiseismic Research of the National Technical University of Athens under the support of the Greek Association for Computational Mechanics (GRACM), the Institute of Research and Development for Computational Methods in Engineering Sciences (ICMES) and the Computer Applications and Education in Engineering Sciences (CAEES).

The editor of this volume would like to thank all authors for their contributions. Special thanks go to the colleagues who contributed to the organization of the Minisymposia and to the reviewers who, with their work, contributed to the scientific quality of this e-book.

M. Papadrakakis
National Technical University of Athens, Greece

V. Papadopoulos
National Technical University of Athens, Greece

G. Stefanou
Aristotle University of Thessaloniki, Greece

V. Plevris
Oslo and Akershus University College of Applied Sciences, Oslo, Norway

ACKNOWLEDGEMENTS

The conference organizers acknowledge the support towards the organization of the 7th European Congress on Computational Methods in Applied Sciences and Engineering, to the following organizations:

- Greek Association for Computational Mechanics (GRACM)
- Institute of Structural Analysis and Antiseismic Research, National Technical University of Athens
- John Argyris Foundation
- Institute of Research and Development for Computational Methods in Engineering Sciences (ICMES)
- Computer Applications and Education in Engineering Sciences (CAEES)
- Region of Crete
- Municipality of Hersonissos

Plenary Speakers

We would also like to thank the Plenary and Semi-Plenary Speakers for their help in the setting up of a high standard Scientific Programme.

Plenary Speakers: Alexandre Ern, Charbel Farhat, Gerhard A. Holzapfel, Antonio Huerta, Thomas J.R. Hughes, Alfio Quarteroni, Ole Sigmund, Theodosios P. Tasios

Semi-Plenary Speakers: Assyr Abdulle, Olivier Allix, Lourenco Beirao da Veiga, Andreas G. Boudouvis, Annalisa Buffa, Ramon Codina, Laura De Lorenzis, Bernard Geurts, Kyriakos Giannakoglou, Antonio J. Gil, Dan Givoli, Anthony Gravouil, George Karniadakis, Mats G. Larson, Kim Meow Liew, Wing Kam Liu, Federico Negri, Shinji Nishiwaki, Ursula Rasthofer, Ernst Rank, Cord-Christian Rossow, Giancarlo Sangalli, Jörg Schröder, Bert Sluys, Christian Soize, Roberto Verzicco, Barbara Wohlmuth

SUMMARY

Preface..... iii

Acknowledgements..... iv

Contents..... xvii

VOLUME I

SEMI-PLENARY LECTURES 1

MINISYMPOSIA

MS 102 - COMPUTATIONAL MODELS IN BIOMECHANICS AND MECHANOBIOLOGY 27
Organized by Estefania Peña, Renato Natal Jorge, Miguel A. Martínez, Pedro S. Martins

MS 103 - MECHANICS OF BIOLOGICAL TISSUES 55
Organized by Markus Böl, Gerhard A. Holzapfel

MS 105 - SIMULATION OF CARDIOVASCULAR PROCEDURES AND DEVICES 81
Organized by Ferdinando Auricchio, Michele Conti, Simone Morganti, Alessandro Reali, Alessandro Veneziani

MS 106 - DIRECT AND INVERSE METHODS FOR CARDIOVASCULAR AND PULMONARY BIOMECHANICS 98
Organized by Wolfgang A. Wall, C. Alberto Figueroa, Marek Behr

MS 108 - NUMERICAL METHODS FOR COUPLED PROBLEMS IN BIOMEDICAL APPLICATIONS 105
Organized by Martina Bukac, Annalisa Quaini

MS 111 - POPULATION BALANCE MODELING: CURRENT STATUS, FUTURE PROSPECTS AND NOVEL APPLICATIONS FROM NANOPARTICLES' SYNTHESIS TO (LUNG) CANCER 120
Organized by Georgios Lolas, Georgios Bourantas, Panagiotis Gavriliadis, Konstantinos Syrigos

MS 112 - ANEURYSMS: SOLID MECHANICS, FLUID MECHANICS, AND MECHANOBIOLOGY 137
Organized by Christian J. Cyron, Sven Hirsch, Philippe Bijlenga, Roland C. Aydin, Anne M. Robertson, Gerhard A. Holzapfel

MS 113 - MATHEMATICAL AND NUMERICAL MODELING OF THE HEART 155
Organized by Luca Dede', Luca Pavarino, Alfio Quarteroni

MS 115 - TUMOR GROWTH MODELING AND THE MECHANICAL ASPECTS OF CANCER 178
Organized by Hector Gomez, Assad Oberai, Krishna Garikipati, Kristen Mills, Thomas J.R. Hughes

MS 201 - MICROSTRUCTURE-DRIVEN DEFORMATION AND FAILURE IN CRYSTALLINE MATERIALS 188
Organized by Pilar Ariza, Lucia Nicola, Angelo Simone

MS 202 - CIVIL ENGINEERING MATERIALS AND STRUCTURES UNDER EXTREME LOADINGS 216
Organized by Fabrice Gatuingt, Frédéric Dufour, Panagiotis Kotronis

MS 203 - COMPUTATIONAL METHODS FOR MODELLING INSTABILITIES IN SOLIDS & STRUCTURES 254
Organized by Spyros A. Karamanos

MS 204 - IMPACT AND CRASH MECHANICS 273
Organized by Manfred Bischoff, Fabian Duddeck

MS 205 - PROBABILISTIC APPROACH TO NUMERICAL SIMULATION OF FRACTURE 323
Organized by Alexander V. Gerasimov

MS 301 - METHODS FOR CUT AND COMPOSITE MESHES: THEORY, ALGORITHMS AND APPLICATIONS 373
Organized by Mats G. Larson, André Massing

MS 302 - MESH GENERATION AND ADAPTION 384
Organized by Josep Sarrate, Xevi Roca, Rafael Montenegro, Eloi Ruiz

MS 303 - CURVED MESH GENERATION FOR HIGH-ORDER METHODS 417
Organized by Xevi Roca, Josep Sarrate

MS 304 - COMPUTATIONAL MODELLING OF ADDITIVE PRODUCTION PROCESSES 469
Organized by Dirk Hartmann, Stefan Kollmannsberger, Ernst Rank, Utz Wever

MS 305 - ADVANCED MESHING METHODS FOR INDUSTRIAL APPLICATIONS 480
Organized by Frederic Alauzet, Thierry Coupez, Alain Dervieux, Adrien Loseille

MS 306 - LATTICE SPRING METHODS FOR LINEAR AND NONLINEAR CONTINUA 511
Organized by Ioannis Doltsinis

MS 403 - PARTICLE-BASED METHODS IN FLUID MECHANICS 539
Organized by Sergio Idelsohn, Eugenio Oñate

MS 404 - SIMULATION OF ENVIRONMENTAL FLOWS 650
Organized by Pablo Ortiz, Piotr K. Smolarkiewicz, Joanna Szmelter

MS 405 - COMPUTATIONAL MODELING OF MULTIPHASE FLOWS: ADVANCED METHODS, INTERFACE PHENOMENA AND ENVIRONMENTAL APPLICATIONS 750
Organized by Adeline Montlaur, Santiago Arias Calderón, Martin Kronbichler

MS 406 - ADVANCES IN COMPUTATIONAL METHODS FOR GAS-LIQUID TWO-PHASE FLOW 788
Organized by Byeong Rog Shin, Takeo Kajishima

MS 408 - MANIPULATION AND CONTROL OF TURBULENT FLOW 831
Organized by Markus Rütten, Christina Voß

MS 409 - CURRENT TRENDS IN MODELLING AND SIMULATION OF TURBULENT FLOWS 865
Organized by Suad Jakirlić

MS 410 - COMPLEX FLUID FLOWS IN ENGINEERING: MODELLING, SIMULATION AND OPTIMIZATION 949
Organized by Stefanie Elgeti, Philipp Knechtges

MS 411 - NON-NEWTONIAN HEAT AND FLUID FLOW SUBJECTED TO MAGNETIC FORCES 1061
Organized by Laszlo Konozsy, Dimitris Drikakis

MS 412 - NUMERICAL METHODS FOR WAVES AND FLOWS IN COASTAL AND DEEP WATER HYDRODYNAMICS 1093
Organized by Nina Shokina, Yuri Shokin, Leonid Chubarov, Gayaz Khakimzyanov, Vadym Aizinger, Denys Dutykh

MS 414 - NEW TRENDS IN NUMERICAL METHODS FOR MULTI-MATERIAL COMPRESSIBLE FLUID FLOWS 1139
Organized by Andy Barlow, Michael Dumbser, Raphaël Loubère, Pierre-Henri Maire, Rob Rieben, Mikhail Shashkov, François Vilar

MS 415 - COMPUTATIONAL NON-NEWTONIAN FLUID MECHANICS 1205
Organized by Georgios Georgiou, John Tsamopoulos

MS 501 - ALGORITHMIC ASPECTS OF HIGH-PERFORMANCE COMPUTING FOR MECHANICS AND PHYSICS 1237
Organized by Santiago Badia, Victor Calo, Javier Principe

MS 503 - HPC-BASED SIMULATIONS FOR THE ENGINEERING REALM AND INDUSTRIAL APPLICATIONS 1277
Organized by Makoto Tsubokura, Mariano Vázquez, Takayuki Aoki

MS 504 - NUMERICAL METHODS AND TOOLS FOR KEY EXASCALE COMPUTING CHALLENGES IN ENGINEERING AND APPLIED SCIENCES 1352
Organized by Eugenio Oñate, Manolis Papadrakakis, Peter Wriggers

MS 505 - INTERACTIVE SIMULATIONS IN COMPUTATIONAL ENGINEERING 1392
Organized by Adrian Harwood, Petra Wenisch

MS 506 - ACCURACY AND EFFICIENCY OF APPROXIMATE COMPUTATIONS IN SCIENCE AND ENGINEERING 1409
Organized by Aram Soroushian

MS 601 - SHOCK WAVE-BOUNDARY LAYER INTERACTION AND ITS CONTROL 1455
Organized by Piotr Doerffer, George Barakos

MS 602 - INNOVATIVE METHODS FOR FLUID-STRUCTURE-INTERACTION 1498
Organized by E. Harald van Brummelen, Roger Ohayon, Trond Kvamsdal

MS 603 - COMPUTATIONAL METHODS IN FLUID-STRUCTURE INTERACTION WITH IMPACT ON INDUSTRIAL APPLICATIONS 1582
Organized by Elisabeth Longatte, Yannick Hoarau, Marianna Braza

MS 605 - FRICTIONAL CONTACTS WITH LUBRICATION – BASICS AND APPLICATIONS 1623
Organized by Michael Müller, Thomas Hagemann

MS 606 - COMPUTATIONAL MODELING OF HYDRAULIC FRACTURING 1637
Organized by Gianluca Cusatis, Gilles Pijaudier-Cabot, Günther Meschke

MS 607 - ADVANCES IN COMPUTATIONAL METHODS FOR LIQUID-VAPOR FLOWS WITH PHASE TRANSFER PROCESSES 1672
Organized by Rémi Abgrall, Pietro M. Congedo, Tore Flåtten, Bernhard Müller, Marica Pelanti, Maria Giovanna Rodio

MS 608 - ADVANCES IN TIME INTEGRATION FOR SOLID, FLUID AND COUPLED SYSTEMS 1709
Organized by Ilinca Stanciulescu, Peter Betsch

MS 609 - ADVANCED COMPUTATIONAL MODELING OF BATTERIES AND FUEL CELLS 1838
Organized by Edwin Knobbe, Wolfgang A. Wall

MS 610 - NUMERICAL METHODS TO STUDY THE CONTACT MECHANICS OF DRY, ADHESIVE AND LUBRICATED ROUGH SURFACES 1850
Organized by Carmine Putignano, Daniele Dini

MS 612 - NUMERICAL SIMULATIONS FOR SMART-CITY APPLICATIONS 1876
Organized by Julien Waeytens, Rachida Chakir

MS 615 - COMPUTATIONAL MODELS IN MAGNETOHYDRODYNAMICS 1886
Organized by Oleg Zikanov

MS 701 - ADVANCED MATERIALS: COMPUTATIONAL ANALYSIS OF PROPERTIES AND PERFORMANCE 1893
Organized by Vadim V. Silberschmidt, Valery P. Matveenko

MS 702 - MODELING OF NANOFILLED COMPOSITES 1971
Organized by Konstantinos I. Tserpes

MS 705 - IDENTIFICATION OF MATERIAL MODELS 2003
Organized by Danuta Szeliga, Wacław Kuś, Tadeusz Burczyński, Jan Kusiak

MS 706 - MODELING OF FIBER-BASED STRUCTURES - TEXTILES AND TEXTILE REINFORCED COMPOSITES 2010
Organized by Yordan Kyosev, Philippe Boisse, Nahiene Hamila, Damien Durville

MS 707 - MICROMECHANICAL MODELLING: COMPETITION BETWEEN ANALYTICAL AND NUMERICAL METHODS 2039
Organized by Siegfried Schmauder, Vera Petrova

MS 711 - FOURIER-BASED METHODS FOR COMPUTING THE BEHAVIOR OF HETEROGENEOUS MATERIALS DEVELOPMENTS, EXTENSIONS AND APPLICATIONS 2099
Organized by Lionel Gélébart, Hervé Moulinec, Franz Roters, François Willot

MS 712 - SMART MATERIAL SYSTEMS AND STRUCTURES 2110
Organized by Mieczysław Kuczma, Pavel Krejči, Jörg Schröder, Georgios E. Stavroulakis, Gwidon Szefer

MS 713 - MICROSTRUCTURE-BASED MODELLING OF HETEROGENEOUS MATERIALS 2134
Organized by Jan Zeman, Jan Novak, Guillermo Díaz

MS 714 - STRENGTH, FATIGUE AND STABILITY OF COMPOSITE STRUCTURES 2182
Organized by Raimund Rolfes, Martin Ruess, Kai-Uwe Schröder

VOLUME II

MINISYMPOSIA

MS 715 - COMPUTATIONAL ANALYSIS OF COMPOSITE STRUCTURES 2210
Organized by Efstathios E. Theotokoglou

MS 801 - MULTISCALE COMPUTATIONAL HOMOGENIZATION FOR BRIDGING SCALES IN THE MECHANICS AND PHYSICS OF COMPLEX MATERIALS 2266
Organized by Julien Yvonnet, Kenjiro Terada, Peter Wriggers, Marc Geers

MS 802 - ADVANCES IN THE MODELLING OF MULTI-SCALE, MULTI-PHYSICS AND MULTI-UNCERTAINTY PROBLEMS 2323
Organized by Francisco M. Andrade Pires, Chengfeng Li

MS 805 - ADVANCED MULTI-PHYSICS AND MULTI-SCALE TECHNIQUES FOR MODELING INELASTIC PROCESSES IN SOLIDS: DAMAGE, FRACTURE AND CONTACT MECHANICS 2353
Organized by Mauro Corrado, Marco Paggi, José Reinoso

MS 806 - MULTISCALE MODELLING OF MATERIALS AND STRUCTURES 2363
Organized by Tadeusz Burczyński, Xavier Oliver, Maciej Pietrzyk, Alfredo Huespe

MS 807 - ADVANCED COMPUTATIONAL STRATEGIES FOR MODELLING, SIMULATION AND CHARACTERISATION OF MULTI-SCALE HETEROGENEOUS MATERIALS 2449
Organized by Stéphane Bordas, Daniel Dias-da-Costa, Fabrice Pierron, Timon Rabczuk, Pierre Kerfriden, Pascal Lava

MS 808 - MULTISCALE AND MULTIPHYSICS MODELING OF CEMENTITIOUS MATERIALS 2460
Organized by Jörg F. Unger, Thomas Titscher

MS 809 - MULTISCALE STOCHASTIC FINITE ELEMENT METHODS 2471
Organized by George Stefanou, Xi Frank Xu, Yu Ching Wu

MS 811 - MULTISCALE MODELING OF CONCRETE AND CONCRETE STRUCTURES 2501
Organized by Herbert Mang, Yong Yuan

MS 901 - ISOGEOMETRIC METHODS 2514
Organized by Yuri Bazilevs, David J. Benson, Rene De Borst, Thomas J.R. Hughes, Trond Kvamsdal, Alessandro Reali, Giancarlo Sangalli, Clemens V. Verhoosel

MS 902 - INNOVATIVE NUMERICAL APPROACHES FOR MULTI-PHYSICS PROBLEMS 2558
Organized by Anna Pandolfi, Laurent Stainier, Kerstin Weinberg

MS 903 - ADVANCES IN FICTITIOUS DOMAIN METHODS FOR SOLID MECHANICS 2566
Organized by Alexander Düster, Ernst Rank, Stefan Kollmannsberger, Andreas Schröder

MS 905 - DISCONTINUOUS GALERKIN METHODS: NEW TRENDS AND APPLICATIONS 2614
Organized by Sonia Fernández-Méndez, Nicoletta Franchina

MS 906 - MATHEMATICAL ADVANCES IN ISOGEOMETRIC ANALYSIS 2715
Organized by Annalisa Buffa, John A. Evans, Thomas J.R. Hughes, Giancarlo Sangalli

MS 907 - REGULARIZED ENRICHED APPROXIMATIONS AND QUADRATURE FOR DISCONTINUITIES, SINGULARITIES AND CONTINUOUS-DISCONTINUOUS TRANSITION 2726
Organized by Elena Benvenuti, Giulio Ventura, José M.A. César de Sá

MS 908 - VERIFICATION AND VALIDATION OF STRUCTURAL MECHANICS SIMULATION MODELS 2734
Organized by George Lampeas

MS 910 - HIGH ORDER CFD METHODS: CONCLUSIONS AND OUTLOOK 2843
Organized by Koen Hillewaert, John Ekaterinaris, Peter Vincent, Norbert Kroll, Norbert Huynh, Z.J. Wang

MS 911 - NUMERICAL METHODS IN THE MECHANICS OF GENERALIZED CONTINUA 2869
Organized by Elena Atroshchenko, Jack S. Hale, George Bourantas, Stéphane P.A Bordas

MS 912 - HIGH-ORDER METHODS, SENSITIVITY ANALYSIS AND ADAPTATION FOR THE NAVIER STOKES EQUATIONS 2883
Organized by Vincent Couaillier, Rémi Abgrall, Eusebio Valero

MS 913 - HIGH-ORDER METHODS FOR POLYGONAL AND POLYHEDRAL MESHES 2930
Organized by Lourenço Beirao da Veiga, Franco Brezzi, Donatella Marini, Alessandro Russo

MS 916 - DIRECT METHODS FOR LIMIT STATES OF MATERIALS AND STRUCTURES 2971
Organized by Konstantinos V. Spiliopoulos, Dieter Weichert

MS 917 - MESOSCOPIC METHODS FOR COMPLEX FLUIDS AND SOFT MATTER 3026
Organized by Zhen Li, Wenxiao Pan, Igor V. Pivkin

MS 918 - COMPUTER ALGEBRA SYSTEMS IN MODELLING STATIC AND DYNAMIC PROBLEMS IN MECHANICS OF SOLIDS 3045
Organized by Alexander V. Matrosov, Dmitriy P. Goloskokov

MS 919 - RECENT ADVANCES IN NUMERICAL SIMULATION AND ANALYSIS OF KINETIC MODELS 3113
Organized by E. Harald van Brummelen, Manuel Torrilhon

MS 921 - RECENT ADVANCES IN BOUNDARY ELEMENT METHODS 3139
Organized by Gernot Beer, Luiz Wrobel, Martin Schan

MS 922 - HIGH-ORDER METHODS FOR ELASTIC WAVES AND THEIR APPLICATION 3211
Organized by Thomas Hagstrom, Daniel Appelo

MS 923 - NOVEL DISCRETIZATION METHODS – MATHEMATICAL AND MECHANICAL ASPECTS 3224
Organized by Jörg Schröder, Peter Wriggers, Ferdinando Auricchio, Carsten Carstensen

MS 1001 - STRUCTURAL AND MULTIDISCIPLINARY OPTIMIZATION 3239
Organized by J.F. Aguilar Madeira, Helder C. Rodrigues

MS 1002 - EVOLUTIONARY ALGORITHMS AND METAHEURISTICS IN CIVIL ENGINEERING AND CONSTRUCTION MANAGEMENT 3473
Organized by Jorge Magalhaes-Mendes, David Greiner

MS 1003 - ADVANCES IN DESIGN OPTIMIZATION OF STRUCTURES AND MATERIALS 3534
Organized by Zhen Luo, Zhan Kang

MS 1004 - AERODYNAMIC STRATEGIES FOR THE GLOBAL OPTIMIZATION OF FLYING CONFIGURATIONS IN SUPERSONIC FLOW 3596
Organized by Adriana Nastase, Catalin Nae

MS 1005 - MONITORING AND CONTROL OF STRUCTURES 3630
Organized by Resat Oyguc

MS 1006 - PARAMETER IDENTIFICATION IN SOLID MECHANICS 3652
Organized by A. Gil Andrade-Campos, Marco Rossi, Sandrine Thuillier, Franck Toussaint, Marta C. Oliveira

MS 1007 - ADDITIVE MANUFACTURING AND OPTIMIZATION 3689
Organized by Ekkehard Ramm, Ole Sigmund, Pierre Duysinx, Wing Kam Liu

MS 1008 - ULTRASONIC GUIDED WAVES TESTING AND MONITORING 3753
Organized by Yaacoubi Slah, Nico Declercq

MS 1009 - ADJOINT METHODS FOR STEADY & UNSTEADY OPTIMIZATION 3770
Organized by Kyriakos C. Giannakoglou, Jens Dominik Mueller

MS 1010 - INVERSE PROBLEMS, DESIGN AND OPTIMIZATION 4061
Organized by Marcelo J. Colaço, Helcio R. B. Orlande, George S. Dulikravich

MS 1011 - SURROGATE-ASSISTED EVOLUTIONARY ALGORITHMS IN AERODYNAMIC DESIGN/OPTIMIZATION 4082
Organized by Varvara Asouti, Esther Andrés, Emiliano Iuliano

MS 1013 - SOLUTION OF LARGE-SCALE INVERSE PROBLEMS 4164
Organized by Clint Dawson, Steve Mattis, Troy Butler, Lindley Graham

MS 1014 - DESIGN OPTIMIZATION AND INVERSE PROBLEMS FOR WAVE PROPAGATION PROBLEMS 4178
Organized by Martin Berggren

MS 1101 - REDUCED BASIS, POD AND PGD MODEL ORDER REDUCTION TECHNIQUES 4187
Organized by

MS 1103 - MATHEMATICAL SURROGATE MODELLING IN ELECTROMAGNETICS 4284
Organized by Francisco Chinesta, Elias Cueto, Antonio Huerta, Pierre Ladeveze, Gianluigi Rozza

MS 1201 - COMPUTATIONAL STRUCTURAL DYNAMICS 4351
Organized by
Evangelos J. Sapountzakis, Andreas E. Kampitsis

VOLUME III

MINISYMPOSIA

MS 1202 - ADVANCED BEAM MODELS 4445
Organized by Dinar Camotim, Zuzana Dimitrovová, Rodrigo Gonçalves

MS 1203 - THE MODELS AND INVESTIGATIONS METHODS OF DYNAMICS OF THE SOLIDS SYSTEMS WITH DRY FRICTION 4553
Organized by Alexey A. Kireenkov, Alexander V. Karapetyan

MS 1204 - NONLINEAR DYNAMICS OF ROTATING STRUCTURES 4611
Organized by Evangeline Capiez-Lernout, Marc P. Mignolet, Christian Soize

MS 1206 - ADVANCES IN NUMERICAL METHODS FOR LINEAR AND NON-LINEAR DYNAMICS AND WAVE PROPAGATION 4674
Organized by Alexander Idesman

MS 1207 - COMPUTATIONAL SIMULATION OF SMART STRUCTURES AND MATERIALS 4728
Organized by Ruediger Schmidt, Kai-Uwe Schröder

MS 1208 - BIFURCATIONS AND STABILITY 4741
Organized by Pekka Neittaanmäki, Nikolay Banichuk, Juha Jeronen, Tero Tuovinen

MS 1209 - DYNAMIC ANALYSIS OF BEAMS UNDER MOVING VEHICLES: APPLICATION TO RAILWAY TRACK MODELLING, DESIGN AND REHABILITATION 4780
Organized by Fernardo Simões, Antonio Pinto da Costa

MS 1210 - ADVANCES IN MODELING AND ANALYSIS OF FGM STRUCTURES 4810
Organized by Justin Murin, Stephan Kugler, Mehdi Aminbaghai

MS 1211 - COMPUTATIONAL STRATEGIES FOR STRUCTURAL ROBUSTNESS ASSESSMENT 4900
Organized by Domenic Asprone, Fulvio Parisi

MS 1212 - DYNAMICS AND SEISMIC RESPONSE OF ROCKING AND SELF-CENTERING STRUCTURES..... 4959
Organized by Matthew DeJong, Elias Dimitrakopoulos, Michalis Fragiadakis

MS 1213 - INNOVATIVE STRUCTURAL SYSTEMS FOR SEISMIC RESISTANT BUILDINGS 5142
Organized by Carlo Castiglioni

MS 1214 - HISTORIC MASONRY STRUCTURES: MODELLING, ASSESSMENT & RETROFIT 5172
Organized by Panagiotis Asteris, Charilaos Maniatakis, Constantine Spyrakos

MS 1215 - NONLINEAR VIBRATIONS OF CONSERVATIVE AND NONCONSERVATIVE SYSTEMS: PHENOMENA AND ADVANCED NUMERICAL METHODS 5396
Organized by Malte Krack, Ludovic Renson, Gaëtan Kerschen

MS 1216 - STRUCTURAL ANALYSIS AND VIBRATIONS 5444
Organized by Diana V. Bambill, Carlos A. Rossit

MS 1217 - COMPUTATIONAL METHODS IN EARTHQUAKE ENGINEERING AND STRUCTURAL DYNAMICS 5464
Organized by Vagelis Plevris, Georgia Kremmyda, Yasin Fahjan

MS 1218 - STABILITY AND CONTROL OF FLEXIBLE STRUCTURES 5646
Organized by Ilaria Venanzi, Marco Lepidi

MS 1220 - DYNAMICS OF STRUCTURES SUBJECT TO SEISMIC EXCITATIONS 5712
Organized by Michel Géradin, Evtim Zahariev, Evangelos J. Sapountzakis

MS 1221 - COMPUTATIONAL STRUCTURAL STABILITY 5768
Organized by Herbert A. Mang, Yeon-Bin Yang

MS 1222 - INFLUENCE OF LIQUEFIABLE SOIL ON SINGLE AND CLOSELY CLUSTERED STRUCTURES 5817
Organized by Nawawi Chouw, Rolly Orense, Tam Larkin

MS 1224 - INNOVATIVE SOLUTIONS FOR THE SEISMIC PROTECTION OF INDUSTRIAL BUILDINGS 5834
Organized by Walter Salvatore, Carlo Castiglioni, Francesco Morelli, Nikolaos Bakas

MS 1225 - SEISMIC PERFORMANCE ASSESSMENT OF STRUCTURES AND SEISMIC RISK MITIGATION STRATEGIES 5934
Organized by Marco Vona

MS 1301 - THE STOCHASTIC COMPUTER METHODS IN MECHANICS 6029
Organized by Marcin Kamiński, Takahiko Kurahashi

MS 1303 - ANALYSIS AND DESIGN OF SAFETY CRITICAL SYSTEMS UNDER UNCERTAINTY 6091
Organized by Edoardo Patelli, Michael Beer, Matteo Broggi, Francisco Alejandro Díaz De la O

MS 1304 - STOCHASTIC MODELING AND IDENTIFICATION OF UNCERTAINTIES IN COMPUTATIONAL MECHANICS 6136
Organized by Johann Guilleminot, Maarten Arnst, Christian Soize

MS 1306 - ERCOFTAC SIG-45: UNCERTAINTY QUANTIFICATION IN CFD AND FLUID STRUCTURE INTERACTION 6196
Organized by Didier Lucor, Sunetra Sarkar

MS 1307 - NON-INTRUSIVE SURROGATE MODELS FOR UNCERTAINTY QUANTIFICATION IN HIGH DIMENSIONS 6250
Organized by Bruno Sudret, Eleni Chatzi, Jean-Marc Bourinet, Nicolas Gayton

MS 1308 - MODELLING AND INVERSE METHODS IN NONLINEAR DYNAMICAL SYSTEMS 6316
Organized by Sotirios Natsiavas, Costas Papadimitriou, Eleni Chatzi, Dimitrios Giagopoulos

MS 1309 - SCALABLE MULTI-FIDELITY MODELING FOR DESIGN, UNCERTAINTY QUANTIFICATION, AND INVERSE PROBLEMS 6383
Organized by Paris Perdikaris, George Em. Karniadakis

MS 1310 - COMPUTATIONAL METHODS FOR THE SOLUTION OF STOCHASTIC DIFFERENTIAL EQUATIONS 6398
Organized by Jianbing Chen, Ioannis Kougioumtzoglou, Vissarion Papadopoulos

MS 1401 - TOICA: THERMAL OVERALL INTEGRATED CONCEPT AIRCRAFT 6420
Organized by Pierre Arbez, Jean-Claude Dunyach

STS 3 - INNOVATIVE DESIGN OPTIMIZATION TOOLS LINKED TO INDUSTRIAL AERONAUTICAL APPLICATIONS: TARGETING GREENER PERFORMANCES 6471
Organized by Jacques Periaux, Gabriel Bugeda

STS 5 - TRANSITION LOCATION EFFECT ON SHOCK WAVE BOUNDARY LAYER INTERACTION 6485
Organized by Piotr Doerffer, Pawel Flaszynski

STS 7 - MORPHING TECHNOLOGIES FOR AIRCRAFT WINGS 6496
Organized by Hans Peter Monner

STS 9 - ADVANCED WING HIGH-LIFT SYSTEMS 6505
Organized by Jochen Wild

YOUNG INVESTIGATORS MINISYMPOSIUM 6515
Organized by Jaan-Willem Simon, Alexander Popp, Joan Baiges

ECCOMAS OLYMPIAD MINISYMPOSIUM 6526

VOLUME IV

CONTRIBUTED SESSIONS

CS 110 - NUMERICAL MODELS IN BIOMECHANICS 6539

CS 210 - NUMERICAL SIMULATION OF COMPOSITE MATERIALS 6637

CS 211 - CRACK PROPAGATION 6688

CS 212 - NUMERICAL MODELING OF DAMAGE, FAILURE AND FRACTURE 6699

CS 230 - DYNAMIC FAILURE AND FRACTURE 6799

CS 310 - CAD, CAM AND CAE 6847

CS 320 - GRID GENERATION AND ADAPTIVE TECHNIQUES 6911

CS 410 - COMPUTATIONAL FLUID MECHANICS 6955

CS 420 - MULTI-PHASE AND CHEMICALLY REACTING FLOWS 7138

CS 450 - NUMERICAL METHODS AND CONVERGENCE ACCELERATION IN CFD 7274

CS 460 - UNSTEADY FLOW COMPUTATION 7390

CS 500 - HIGH PERFORMANCE COMPUTING 7572

CS 610 - AERO-ACOUSTICS 7589

CS 620 - COMPUTATIONAL CONTACT MECHANICS 7628

CS 630 - SIMULATION OF FLUID-STRUCTURE INTERACTION 7640

CS 720 - COMPUTATIONAL MATERIALS SCIENCE 7711

CS 750 - COMPUTATIONAL MODELING OF COMPOSITES 7765

CS 751 - SMART MATERIALS AND STRUCTURES 7853

CS 830 - COMPUTATIONAL NANOTECHNOLOGY 7875

CS 840 - MULTI-SCALE COMPUTATIONAL METHODS FOR SOLIDS AND FLUIDS 7886

CS 930 - HIGH-ORDER DISCRETIZATION METHODS 7956

CS 940 - EXTENDED DISCRETIZATION METHODS 8066

CS 960 - MESHLESS METHODS 8115

CS 980 - NUMERICAL AND SYMBOLIC COMPUTATION 8162

CS 990 - PARTICLE-BASED METHODS 8196

CS 1010 - COMPUTATIONAL INVERSE PROBLEMS AND OPTIMIZATION 8247

CS 1020 - EVOLUTIONARY AND DETERMINISTIC METHODS FOR DESIGN, OPTIMIZATION AND CONTROL 8391

CS 1100 - REDUCTION ORDER METHODS 8539

CS 1200 - STRUCTURAL DYNAMICS 8597

CS 1201 - COMPUTATIONAL SOIL MECHANICS 8708

CS 1202 - STRUCTURAL ANALYSIS AND MULTIBODY DYNAMICS 8739

CS 1300 - UNCERTAINTY QUANTIFICATION AND ERROR ESTIMATION 8769

CONTENTS

Semi-plenary Lectures

NONPARAMETRIC PROBABILISTIC APPROACH OF MODEL UNCERTAINTIES INTRODUCED BY A PROJECTION-BASED NONLINEAR REDUCED-ORDER MODEL 1
Christian Soize, Charbel Farhat

Minisymposia

100 BIOLOGICAL SYSTEMS

MS 102: COMPUTATIONAL MODELS IN BIOMECHANICS AND MECHANOBIOLOGY

THE INFLUENCE OF MECHANICAL STIMULUS ON NUTRIENT TRANSPORT AND CELL GROWTH IN ENGINEERED CARTILAGE: A FINITE ELEMENT APPROACH 27
S. Cortez, A. Completo, J.L. Alves

HYBRID CELL CENTRED/VERTEX MODEL FOR LARGE TISSUE DEFORMATIONS 37
José J. Muñoz, Payman Mosaffa, Yanlan Mao, Rob Tetley, Nina Asadipour, Antonio Rodríguez-Ferran

SENSITIVITY OF TEMPERATURE FIELD IN THE SYSTEM PROTECTIVE CLOTHING - FOREARM WITH RESPECT TO PERTURBATIONS OF EXTERNAL HEATING CONDITIONS 46
Bohdan Mochnacki, Mariusz Ciesielski

MS 103: MECHANICS OF BIOLOGICAL TISSUES

INFLUENCE OF MUSCLE COMPRESSION ON DYNAMIC MUSCLE PERFORMANCE 55
Tobias Siebert, Norman Stutzig, Olaf Till, Christian Rode

A 3D ELECTROMECHANICAL FEM-BASED MODEL FOR CARDIAC TISSUE 63
Minh Tuan Duong, Alexander Jung, Ralf Frotscher, Manfred Staat

BIOMECHANICAL PROPERTIES OF CORNEAL EXTRACELLULAR MATRIX USING EXPERIMENTAL AND COMPUTATIONAL METHODS 74
Hamed Hatami-Marbini

MS 105: SIMULATION OF CARDIOVASCULAR PROCEDURES AND DEVICES

MODELLING OF PATIENT-SPECIFIC CASES OF ATHEROSCLEROSIS IN CAROTID ARTERIES 81
Timur Gamilov, Roman Pryamonosov, Sergey Simakov

PERSONALIZED COMPUTATION OF FRACTIONAL FLOW RESERVE IN CASE OF TWO CONSECUTIVE STENOSES ... 90
Yuri Vassilevski, Timur Gamilov, Philip Kopylov

MS 106: DIRECT AND INVERSE METHODS FOR CARDIOVASCULAR AND PULMONARY BIOMECHANICS

BLOOD FLOW IN THE COMMON CAROTID ARTERY WITH STENOSIS 98
Helena Henriques, Luisa Sousa, Catarina Castro, Carlos António, Rosa Santos, Pedro Castro, Elsa Azevedo

MS 108: NUMERICAL METHODS FOR COUPLED PROBLEMS IN BIOMEDICAL APPLICATIONS
A FENICS-HPC FRAMEWORK FOR MULTI-COMPARTMENT BLOCH-TORREY MODELS 105
Dang Van Nguyen, Johan Jansson, Johan Hoffman

MS 111: POPULATION BALANCE MODELING: CURRENT STATUS, FUTURE PROSPECTS AND NOVEL APPLICATIONS FROM NANOPARTICLES' SYNTHESIS TO (LUNG) CANCER
THERMODYNAMICS OF IRREVERSIBLE AGGREGATION 120
Themis Matsoukas

MS 112: ANEURYSMS: SOLID MECHANICS, FLUID MECHANICS, AND MECHANOBIOLOGY
RELATIVE ROLES OF MECHANICS AND BIOCHEMISTRY IN THE INITIATION AND PROGRESSION OF CEREBRAL ANEURYSM THROMBOSIS 137
Malebogo Ngoepe, Yiannis Ventikos

TECHNIQUES TO INTEGRATE PATIENT-SPECIFIC SIMULATION OF ANEURYSMAL BLOOD FLOW INTO THE CLINICAL WORKFLOW 147
Gábor Závodszky, Roland Joó-Kovács, György Paál, István Szikora

MS 113: MATHEMATICAL AND NUMERICAL MODELING OF THE HEART
PROXIMAL ISOVELOCITY SURFACE FOR DIFFERENT MITRAL VALVE HOLE GEOMETRIES 155
Alexandre This, Hernán G. Morales, Odile Bonnefous

COMPARISON OF TETRAHEDRAL AND HEXAHEDRAL MESHES FOR FINITE ELEMENT SIMULATION OF CARDIAC ELECTRO-MECHANICS 164
Bernardo Lino De Oliveira, Joakim Sundnes

MS 115: TUMOR GROWTH MODELING AND THE MECHANICAL ASPECTS OF CANCER
A PARAMETRIC STUDY OF A MULTIPHASE POROUS MEDIA MODEL FOR TUMOR SPHEROIDS AND ENVIRONMENT INTERACTIONS 178
P. Mascheroni, D.P. Boso, C. Stigliano, M. Carfagna, L. Preziosi, P. Decuzzi, B.A. Schrefler

200 DAMAGE, FRACTURE AND FAILURE

MS 201: MICROSTRUCTURE-DRIVEN DEFORMATION AND FAILURE IN CRYSTALLINE MATERIALS
STABILITY OF ASYMMETRIC GRAIN BOUNDARIES IN GRAPHENE 188
J.P. Mendez, F. Arca, M.P. Ariza

DEFORMATION-DIFFUSION COUPLED ANALYSIS OF LONG-TERM HYDROGEN DIFFUSION IN NANOFILMS 197
Xingsheng Sun, Pilar Ariza, Kevin Wang

MESH OBJECTIVE DAMAGE MODELING OF DUCTILE FRACTURE AT VISCO-PLASTIC CONTINUUM RESPONSE ... 209
Senad Razanica, Ragnar Larsson, B. Lennart Josefson

MS 202: CIVIL ENGINEERING MATERIALS AND STRUCTURES UNDER EXTREME LOADINGS

EQUIVALENT SDOF MODEL FOR ESTIMATING BLAST-INDUCED DYNAMIC REACTIONS OF EQUILATERAL TRIANGULAR HARDENED WALL ELEMENTS 216
Sebastian Mendes, Liling Cao, Douglas Heinze, Elisabeth Malsch

MODELLING STRATEGIES OF PRESTRESSING TENDONS AND REINFORCEMENT BARS IN CONCRETE STRUCTURES 234
Antoine Llau, Ludovic Jason, Frédéric Dufour, Julien Baroth

MODELING AND SIMULATION FOR AN OPTIMIZED DESIGN OF A DYNAMIC BEND TEST. 242
Rana Akiki, Fabrice Gatuingt, Cédric Giry, Nicolas Schmitt, Lavinia Stéfan

MS 203: COMPUTATIONAL METHODS FOR MODELLING INSTABILITIES IN SOLIDS & STRUCTURES

ON STABILITY LOSS OF A THIN-WALLED SPHERICAL SHELL SUBJECTED TO EXTERNAL PRESSURE AND INTERNAL HOMOGENEOUS CORROSION 254
Emmanuel Gutman, Rudolf Bergman, Semyon Levitski

EFFICIENCY OF DAMAGE-PLASTICITY MODELS IN CAPTURING COMPACTION-EXPANSION TRANSITION OF CONCRETE UNDER DIFFERENT COMPRESSION LOADING CONDITIONS 261
Reza Mousavi, Masoud D. Champiri, Kaspar J. Willam

MS 204: IMPACT AND CRASH MECHANICS

A SIMPLIFIED MODEL OF A STEEL COLUMN SUBJECTED TO IMPACT 273
Piseth Heng, Mohammed Hjiaj, Jean-Marc Battini

MATERIAL MODELLING OF CAST ALUMINIUM BY APPLICATION OF THE WILKINS DAMAGE MODEL 287
Christian Mühlstätter, Matthias Hartmann

FINITE ELEMENT SIMULATION OF CRACK PROPAGATION AND DELAMINATION IN LAYERED SHELLS DUE TO BLADE CUTTING 295
Federica Confalonieri, Aldo Ghisi, Umberto Perego

EVOLUTIONARY LEVEL SET METHOD FOR CRASHWORTHINESS TOPOLOGY OPTIMIZATION 309
Mariusz Bujny, Nikola Aulig, Markus Olhofer, Fabian Duddeck

MS 205: PROBABILISTIC APPROACH TO NUMERICAL SIMULATION OF FRACTURE

NUMERICAL SIMULATION OF HIGH-VELOCITY INTERACTION ELEMENTS WITH THIN-WALLED STRUCTURES ... 323
Alexander V. Gerasimov, Sergey V. Pashkov

NUMERICAL SIMULATION OF MULTILAYER COMPOSITES FAILURE UNDER DYNAMIC LOADING 334
Sergey A. Zelepugin, Aleksey S. Zelepugin, Vladimir F. Tolkachev

SIMULATION OF MECHANICAL PROPERTIES OF CERAMIC PARTS PRODUCED BY ADDITIVE TECHNOLOGIES IN WIDE RANGE OF LOADING RATES 341
Vladimir A. Skripnyak, Evgeniya Skripnyak, Vladimir V. Skripnyak, Irina K. Vaganova

FRACTURE OF THIN METAL SHEETS WITH DISTRIBUTION OF GRAIN SIZES IN THE LAYERS 355
Natalia V. Skripnyak, Vladimir V. Skripnyak, Vladimir A. Skripnyak

BRITTLE OR QUASI-BRITTLE FRACTURE OF CERAMIC NANOCOMPOSITES UNDER DYNAMIC LOADING 366
Evgeniya G. Skripnyak, Vladimir A. Skripnyak, Vladimir V. Skripnyak, Natalia V. Skripnyak, Irina K. Vaganova

300 DISCRETIZATION METHODS, GRID, MESH AND SOLID GENERATION

MS 301: METHODS FOR CUT AND COMPOSITE MESHES: THEORY, ALGORITHMS AND APPLICATIONS
REMOVING THE STABILIZATION PARAMETER IN FITTED AND UNFITTED SYMMETRIC NITSCHE FORMULATIONS 373
Christoph Lehrenfeld

MS 302: MESH GENERATION AND ADAPTION
A SEMI-UNSTRUCTURED TURBOMACHINERY MESHING LIBRARY WITH FOCUS ON MODELING OF SPECIFIC GEOMETRICAL FEATURES 384
Marco Stelldinger, Thomas Giersch, Felix Figaschewsky, Arnold Kühhorn

CARTESIAN MESHES WITH DYNAMIC LOCAL WAVELET-BASED REFINEMENT FOR FLOW AROUND BODY PROBLEMS. 402
Kirill Merkulov, Igor Menshov, Andrew Plenkin

A MESH GENERATION METHOD FOR HISTORICAL MONUMENTAL BUILDINGS: AN INNOVATIVE APPROACH ... 409
Giovanni Castellazzi, Antonio Maria D'Altri, Stefano De Miranda, Francesco Ubertini, Gabriele Bitelli, Alessandro Lambertini, Ilenia Selvaggi, Antonio Tralli

MS 303: CURVED MESH GENERATION FOR HIGH-ORDER METHODS
HIGH-ORDER CURVED MESH GENERATION BY USING A FINE LINEAR TARGET MESH 417
Verena Schmid, Hadrien Beriot, Onur Atak, Gwenael Gabard

AUTOMATIC GENERATION OF 3D UNSTRUCTURED HIGH-ORDER CURVILINEAR MESHE 428
Michael Turner, David Moxey, Spencerj.Sherwin, Joaquim Peiro

MESH GENERATION FOR THE 2D NURBS-ENHANCED FINITE ELEMENT METHOD 444
Ruben Sevilla, Luke Rees, Oubay Hassan

IMAGE SEGMENTATION TECHNIQUES FOR BIOMEDICAL MODELING: ELECTROPHYSIOLOGY AND HEMODYNAMICS 454
Alexander Danilov, Roman Pryamonosov, Alexandra Yurova

A CONSISTENTLY LINEARISED SOLID MECHANICS BASED MESH DEFORMATION TECHNIQUE FOR HIGH ORDER CURVED ELEMENTS 462
Roman Poya, Ruben Sevilla, Antonio Gil

MS 304: COMPUTATIONAL MODELLING OF ADDITIVE PRODUCTION PROCESSES
VALIDATION OF MODELLING ASSUMPTIONS FOR THE BUILDUP SIMULATION OF LASER BEAM MELTING ON THE BASIS OF THE RESIDUAL STRESS DISTRIBUTION 469
Fabian Bayerlein, Christian Zeller, Martin Wunderer, Johannes Weirather, Michael F. Zäh, Markus Schmid, Georg Schlick, Roland Hessert, Michael Hofmann, Thomas Uihlein

MS 305: ADVANCED MESHING METHODS FOR INDUSTRIAL APPLICATIONS
MESHING OCEAN DOMAINS FOR COASTAL ENGINEERING APPLICATIONS 480
Alexandros Avdis, Christian Jacobs, Simon Mouradian, Jon Hill, Matthew Piggott

HIGH-ORDER TRACKING METHOD FOR THE SIMULATION OF BURNING FRONTS IN 3D 493
Denis Gueyffier, Bastien Andrieu

SINGULAR MESH GENERATION FROM MULTIPLE OVERSET MESHES: A TOOL FOR INDUSTRIAL APPLICATIONS...501
Gennaro Abbruzzese, Marta Cordero Gracia, Mariola Gómez López, Nuno Vinha

MS 306: LATTICE SPRING METHODS FOR LINEAR AND NONLINEAR CONTINUA
GENERATING TRIANGULAR LATTICES FOR SURFACES WITH IRREGULAR BOUNDARY 511
Tatiana Sá Marques, Vitor Dias Da Silva

WAVE TRANSMISSION THROUGH NONLINEAR IMPACTING METAMATERIAL UNIT 526
Arnab Banerjee, Raj Das, Emilio Calius

400 FLOW PROBLEMS

MS 403: PARTICLE-BASED METHODS IN FLUID MECHANICS
A PARTICLE FINITE ELEMENT METHOD FOR MACHINING SIMULATIONS 539
Juan Manuel Rodríguez Prieto, Pär Jonsén, Ales Svoboda

DYNAMICS OF DUST PARTICLES NEAR THE ROUND BELL SUCTION INLET 554
Olga Averkova, Konstantin Logachev, Artur Logachev, Elena Tolmacheva

RECOVERY OF DIFFERENTIATION/INTEGRATION COMPATIBILITY OF MESHLESS OPERATORS VIA LOCAL ADAPTATION OF THE POINT CLOUD IN THE CONTEXT OF NODAL INTEGRATION 568
Gabriel Fougeron, Guillaume Pierrot, Denis Aubry

MODELING OF NON-SPHERICAL, ELONGATED PARTICLES FOR INDUSTRIAL SUSPENSION FLOW SIMULATION ... 586
Jakob Dominik Redlinger-Pohn, Lisa Maria König, Christoph Kloss, Christoph Goniva, Stefan Radl

A LAGRANGIAN FINITE ELEMENT METHOD FOR THE SIMULATION OF 3D COMPRESSIBLE FLOWS 600
Massimiliano Cremonesi, Attilio Frangi

A LAGRANGIAN PFEM APPROACH TO THE NUMERICAL SIMULATION OF 3D LARGE SCALE LANDSLIDES IMPINGING IN WATER RESERVOIRS 608
Massimiliano Cremonesi, Francesco Ferri, Umberto Perego

A NUMERICAL INVESTIGATION OF FLOW DYNAMICS OVER A TRAPEZOIDAL SMOOTH OPEN CHANNEL ……..... 619
Giacomo Viccione

A NUMERICAL INVESTIGATION OF LIQUID IMPACT ON PLANAR SURFACES. …….. 627
Giacomo Viccione, Vittorio Bovolin, Eugenio Pugliese Carratelli

AERODYNAMICS OF AN UNSTEADY PARTICLE FLOW IN THE CHUTE …….. 638
Olga Averkova, Ivan Logachev, Konstantin Logachev

MS 404: SIMULATION OF ENVIRONMENTAL FLOWS
PROJECTION-BASED MODEL REDUCTION FOR FINITE ELEMENT APPROXIMATION OF SHALLOW WATER FLOWS …….. 650
Matthew Farthing, Alexander Lozovskiy, Christopher Kees

EFFICIENT TWO-DIMENSIONAL SIMULATION MODELS FOR HYDRAULIC AND MORPHODYNAMIC TRANSIENTS..670
Pilar Garcia-Navarro, Javier Murillo, Mario Morales-Hernandez, Carmelo Juez, Asier Lacasta

ADDRESSING THE CHALLENGES OF IMPLEMENTATION OF HIGH-ORDER FINITE-VOLUME SCHEMES FOR ATMOSPHERIC DYNAMICS ON UNSTRUCTURED MESHES …….. 684
Panagiotis Tsoutsanis, Dimitris Drikakis

MODELLING OF LONG WAVES (SEICHES) IN CASCAIS BAY …….. 709
Vera Bras, Antonio Trigo-Teixeira

A MESHLESS METHOD FOR SIMULATION OF PARTICLE-DRIVEN GRAVITY CURRENTS WITH OBSTACLES ……. 720
Karel Kovářík, Jozef Mužík, Dana Sitányiová

APPLICATION OF THE SHALLOW WATER EQUATIONS TO REAL FLOODING CASE …….. 735
Hani Ali, Pierre-Yves Lagrée, Jose-Maria Fullana

MS 405: COMPUTATIONAL MODELING OF MULTIPHASE FLOWS: ADVANCED METHODS, INTERFACE PHENOMENA AND ENVIRONMENTAL APPLICATIONS
A MULTIPHASE MODEL FOR THE NUMERICAL SIMULATION OF ICE-FORMATION IN SEA-WATER …….. 750
Vanessa Covello, Antonella Abbà, Luca Bonaventura, Alessandro Della Rocca, Lorenzo Valdettaro

A 3D CFD NUMERICAL STUDY OF THE BUBBLE GENERATION PROCESS INTO A BUBBLE T-JUNCTION GENERATOR AND ITS COMPARISON WITH EXPERIMENTAL DATA: PART I …….. 772
Santiago Arias, Adeline Montlaur

MS 406: ADVANCES IN COMPUTATIONAL METHODS FOR GAS-LIQUID TWO-PHASE FLOW
COUPLING OF FLUID STUCTURE INTRACTION SOLVER WITH A VOF METHOD FOR MULTIPHASE STRUCTURE INTERACTION …….. 788
Daniele Cerroni, Roberto Da Via', Sandro Manservisi, Filippo Menghini

A METHOD OF CHAINED ANALYTICAL WAVE STRUCTURES FOR LARGE SCALE STRATIFIED TWO-PHASE PIPE FLOWS …….. 806
Andreas Holm Akselsen

MS 408: MANIPULATION AND CONTROL OF TURBULENT FLOW
MODELING OF EXCITATION OF CONTROLLING DISTURBANCES IN SWEPT WING BOUNDARY LAYER BY MEANS OF PLASMA ACTUATORS 831
Sergey Chernyshev, Andrey Kiselev, Aleksandr Kuryachii

NUMERICAL SIMULATION OF INTERACTION PROCESS BETWEEN DIELECTRIC BARRIER DISCHARGE AND DUCT FLOW 844
Pavel Semenev, Dmitriy Pudovikov, Pavel Toktaliev

IMPLEMENTATION OF POD AND DMD METHODS IN APACHE SPARK FRAMEWORK FOR SIMULATION OF UNSTEADY TURBULENT FLOW IN THE MODEL COMBUSTOR 857
Mikhail Kalugin, Sergei Strijhak

MS 409: CURRENT TRENDS IN MODELLING AND SIMULATION OF TURBULENT FLOWS
NUMERICAL VALIDATION OF A FOUR PARAMETER LOGARITHMIC TURBULENCE MODEL 865
Daniele Cerroni, Roberto Da Vià, Sandro Manservisi, Filippo Menghini

LARGE-EDDY SIMULATION AND FAR FIELD ACOUSTICS OF A SUBSONIC HOT JET 878
O. Labbé

ON THE APPLICATION OF MODERN TURBULENCE MODELS IN THE FLOW LAMINARIZATION PROBLEMS 893
Sergey Chernyshev, Aleksander Ivanov, Andrey Kiselev, Dmitriy Sboev, Leonid Teperin, Petr Vorotnikov, Valeriy Vozhdaev

ASSESSMENT OF RANS AND DES METHODS FOR THE AHMED BODY 912
Emmanuel Guilmineau, Gan Bo Deng, Alban Leroyer, Patrick Queutey, Michel Visonneau, Jeroen Wackers

A WAY TO IMPROVE JET MODELING WITHIN RANS EQUATION SYSTEM 922
Alexey Troshin

A MODIFIED SSG/LLR-ω REYNOLDS STRESS MODEL FOR PREDICTING BLUFF BODY AERODYNAMICS 936
Csaba Klajbár, László Könözsy, Karl W. Jenkins

MS 410: COMPLEX FLUID FLOWS IN ENGINEERING: MODELLING, SIMULATION AND OPTIMIZATION
NUMERICAL ANALYSIS OF RAREFIED GAS FLOWS USING THE ACADEMIC CFD CODE GALATEA 949
Angelos Klothakis, Georgios Lygidakis, Ioannis Nikolos

TOPOLOGY OPTIMIZATION FOR FLUID FLOW EMPLOYING LOCAL OPTIMALITY CRITERIA 967
Philip Sarstedt, Gerhard Kachel, Jörg Ettrich, Karl Bühler

NUMERICAL SIMULATION OF THE CONJUGATE HEAT TRANSFER IN THE COOLING SYSTEM OF THE COMBUSTION CHAMBERS OF THE AVIATION RAMJET ON THE ENDOTHERMIC FUELS 979
Vadim Volokhov, Pavel Toktaliev, Sergei Martynenko

SENSITIVITY STUDY OF THE NUMERICAL SETUP FOR AN AUTOMATIC OPTIMIZATION PROCEDURE FOR A HYDRAULIC MACHINE 992
Alexander Tismer, Markus Schlipf, Stefan Riedelbauch

ADVANCES IN GEOMETRICAL PARAMETRIZATION AND REDUCED ORDER MODELS AND METHODS FOR COMPUTATIONAL FLUID DYNAMICS PROBLEMS IN APPLIED SCIENCES AND ENGINEERING: OVERVIEW AND PERSPECTIVES 1013
F. Salmoiraghi, F. Ballarin, G. Corsi, A. Mola, M. Tezzele, G. Rozza

NEAR WALL BEHAVIOR OF IMPLICIT LARGE EDDY SIMULATIONS 1032
Ioannis Kokkinakis, Dimitris Drikakis

VALIDATION AND VERIFICATION OF A 2D LATTICE BOLTZMANN SOLVER FOR INCOMPRESSIBLE FLUID FLOW 1046
Tamás István Józsa, Máté Szőke, Tom-Robin Teschner, László Könözsy, Irene Moulitsas

MS 411: NON-NEWTONIAN HEAT AND FLUID FLOW SUBJECTED TO MAGNETIC FORCES
VALIDATION OF A MAGNETO- AND FERRO-HYDRODYNAMIC MODEL FOR NON-ISOTHERMAL FLOWS IN CONJUNCTION WITH NEWTONIAN AND NON-NEWTONIAN FLUIDS 1061
László Könözsy, Pietro Scienza, Dimitris Drikakis

NUMERICAL STUDY OF MAGNETIC PARTICLES CONCENTRATION IN BIOFLUID (BLOOD) UNDER THE INFLUENCE OF HIGH GRADIENT MAGNETIC FIELD IN MICROCHANNEL 1084
Vassilios Loukopoulos, George Bourantas, Dimitrios Labropoulos, Vassilios-Martin Nikiforidis, Stéphane Bordas, George Nikiforidis

MS 412: NUMERICAL METHODS FOR WAVES AND FLOWS IN COASTAL AND DEEP WATER HYDRODYNAMICS
NEW ALGORITHM FOR NUMERICAL SIMULATION OF SURFACE WAVES WITHIN THE FRAMEWORK OF THE FULL NONLINEAR DISPERSIVE MODEL 1093
Zinaida Fedotova, Oleg Gusev, Gayaz Khakimzyanov

NUMERICAL MODELLING OF SURFACE WAVES IN THE FRAMEWORK OF THE SHALLOW WATER MODEL 1104
Nina Shokina, Gayaz Khakimzyanov

ON NUMERICAL METHODS FOR SOLVING RUN-UP PROBLEMS. COMPARATIVE ANALYSIS OF NUMERICAL ALGORITHMS AND NUMERICAL RESULTS. 1127
Leonid Chubarov, Alexandr Rychkov, Gayaz Khakimzyanov, Yurii Shokin

MS 414: NEW TRENDS IN NUMERICAL METHODS FOR MULTI-MATERIAL COMPRESSIBLE FLUID FLOWS
EULERIAN CALCULATIONS OF MULTIMATERIAL FLOWS WITH SUB-CELL RECONSTRUCTION OF INTERFACES ... 1139
Igor Menshov, Pavel Zakharov

SOME ACOUSTIC-TRANSPORT SPLITTING SCHEMES FOR TWO-PHASE COMPRESSIBLE FLOWS 1151
Simon Peluchon, Gérard Gallice, Pierre-Henri Maire

LAGRANGE-FLUX EULERIAN SCHEMES FOR COMPRESSIBLE MULTIMATERIAL FLOWS 1165
Florian De Vuyst, Thibault Gasc, Renaud Motte, Mathieu Peybernes, Raphael Poncet

ISOTROPY–PRESERVING SLOPE LIMITERS FOR FINITE VOLUME METHODS ON SQUARE MESHES 1179
Jan Velechovsky, Marianne Francois

BUILDING A MORE EFFICIENT LAGRANGE-REMAP SCHEME THANKS TO PERFORMANCE MODELING 1191
Thibault Gasc, Florian De Vuyst, Mathieu Peybernes, Raphaël Poncet, Renaud Motte

MS 415: COMPUTATIONAL NON-NEWTONIAN FLUID MECHANICS
DEVELOPMENT OF CONFINED VISCOPLASTIC FLOWS WITH HETEROGENEOUS WALL SLIP 1205
Pandelitsa Panaseti, Maria Philippou, Zacharias Kountouriotis, Georgios Georgiou

SQUEEZE FLOW OF VISCOPLASTIC BINGHAM MATERIAL 1215
Larisa Muravleva

500 HIGH PERFORMANCE COMPUTING

MS 501: ALGORITHMIC ASPECTS OF HIGH-PERFORMANCE COMPUTING FOR MECHANICS AND PHYSICS
PSEUDOMULTIGRID GAUSS–SEIDEL METHOD FOR LARGE SCALE AND HIGH PERFORMANCE COMPUTING 1237
Sergey Martynenko, Vadim Volokhov, Pavel Toktaliev

ALGORITHM FOR FAST SIMULATIONS OF SPACE-TIME FINITE ELEMENT METHOD 1260
Marcin Skotniczny, Anna Paszynska, Maciej Paszynski

HYBRID PARALLELISATION OF AN ALGORITHMICALLY DIFFERENTIATED ADJOINT SOLVER 1268
Pavanakumar Mohanamuraly, Jan Christian Huckelheim, Jens-Dominik Mueller

MS 503: HPC-BASED SIMULATIONS FOR THE ENGINEERING REALM AND INDUSTRIAL APPLICATIONS
EFFICIENT PARALLEL GEOMETRY DISTRIBUTION FOR THE SIMULATION OF COMPLEX FLOWS 1277
Andreas Lintermann

LARGE-SCALE FREE-SURFACE FLOW SIMULATION USING LATTICE BOLTZMANN METHOD ON MULTI-GPU CLUSTERS 1294
Naoyuki Onodera, Kunihide Ohashi,

LARGE-SCALE LES ANALYSIS OF AUTOMOTIVE ENGINE COOLING FAN 1305
Yuji Kobayashi, Kenji Yoshida, Itsuhei Kohri, Masaharu Sakai, Hideo Asano

DIRECT AEROACOUSTIC SIMULATION RELATED WITH MODE CHANGE IN A RECORDER 1315
Hiroshi Yokoyama, Ryoma Hamasuna, Akira Miki, Hirofumi Onitsuka, Akiyoshi Iida

NUMERICAL ANALYSIS OF RADIO FREQUENCY BLACKOUT FOR ATMOSPHERIC REENTRY VEHICLE USING CFD-CEM COMBINED METHOD 1325
Yusuke Takahashi, Reo Nakasato, Nobuyuki Oshima

DIRECT FEM LARGE SCALE COMPUTATION OF TURBULENT MULTIPHASE FLOW IN URBAN WATER SYSTEMS AND MARINE ENERGY 1339
Ezhilmathi Krishnasamy, Johan Hoffman, Johan Jansson

MS 504: NUMERICAL METHODS AND TOOLS FOR KEY EXASCALE COMPUTING CHALLENGES IN ENGINEERING AND APPLIED SCIENCES
KINETIC MODELS AND ALGORITHMS FOR SOLUTION OF THE MAGNETOGASDYNAMIC PROBLEMS ON THE MODERN SUPERCOMPUTING SYSTEMS 1352
Boris Chetverushkin, Nicola D'Ascenzo, A. Saveliev, Valeri Saveliev

ERROR ANALYSIS AND QUANTIFICATION IN NEURON SIMULATIONS 1366
Francesco Casalegno, Francesco Cremonesi, Stuart Yates, Michael L. Hines, Felix Schürmann, Fabien Delalondre

EVALUATION OF SPARSE LINEAR ALGEBRA OPERATIONS IN TRILINOS 1381
Mohammad Siahatgar, Gabriele Von Voigt

MS 505: INTERACTIVE SIMULATIONS IN COMPUTATIONAL ENGINEERING
AUTOMATIC VISUALIZATION AND CONTROL OF ARBITRARY NUMERICAL SIMULATIONS 1392
Jan P. Springer, Helen Wright

MS 506: ACCURACY AND EFFICIENCY OF APPROXIMATE COMPUTATIONS IN SCIENCE AND ENGINEERING
DIRECT TIME-DOMAIN INTEGRATION APPROACH FOR VISCOELASTIC SYSTEMS INVOLVING VARIOUS DAMPING MODELS 1409
Zhe Ding, Li Li, Yujin Hu

ON THE PERFORMANCE OF A COMPUTATIONAL COST REDUCTION TECHNIQUE APPLIED TO COOLING TOWERS TRANSIENT ANALYSIS 1421
Aram Soroushian, Amin Jahani Mehrnoosh, Yalda Zarabimanesh, Mohammad Hadi Ghondaghsaz, Ali Baiani, Ali Zakizade

A DIFFERENT LOOK AT THE RICHARDSON EXTRAPOLATION LEADING TO A NEW PROPOSITION 1432
Aram Soroushian

ON THE FREQUENCY CONTENT OF ERRORS ORIGINATED IN A TIME INTEGRATION COMPUTATIONAL COST REDUCTION TECHNIQUE 1443
Aram Soroushian, Mahmood Hosseini, Seyed Mohammad Khalkhali

600 INTERDISCIPLINARY COUPLED AND CONTACT PROBLEMS

MS 601: SHOCK WAVE-BOUNDARY LAYER INTERACTION AND ITS CONTROL
NUMERICAL SIMULATION OF TRANSONIC BUFFET AND ITS CONTROL USING TANGENTIAL JET BLOWING 1455
Ksenia Abramova, Kamil Khairullin, Alexander Ryzhov, Vitaly Soudakov

SHOCK WAVE BOUNDARY LAYER INTERACTION IN PROFILE CASCADES REPRESENTING ROTOR BLADINGS OF THE LAST STAGE OF LARGE OUTPUT STEAM TURBINES 1464
Martin Luxa, David Simurda, Jaromir Prihoda, Jana Vachova

SHOCK WAVE BOUNDARY LAYER INTERACTION INVESTIGATION: CONTROL BY ROD VORTEX GENERATORS AND THEIR APPLICATION ON HELICOPTER ROTOR BLADES 1474
Fernando Tejero Embuena, Piotr Doerffer, Pawel Flaszynski

ANALYSIS OF EFFECTS OF SHAPE AND LOCATION OF MICRO - TURBULATORS ON UNSTEADY SHOCKWAVE - BOUNDARY LAYER INTERACTION IN TRANSONIC FLOW 1484
Janusz Sznajder, Tomasz Kwiatkowski

MS 602: INNOVATIVE METHODS FOR FLUID-STRUCTURE-INTERACTION
ASSESSMENT OF THE FLUID-STRUCTURE INTERACTION CAPABILITIES FOR AERONAUTICAL APPLICATIONS OF THE OPEN-SOURCE SOLVER SU2. 1498
Ruben Sanchez, Heather L. Kline, David Thomas, Anil Variyar, Marcello Righi, Thomas D. Economon, Juan J. Alonso, Rafael Palacios, Grigorios Dimitriadis, Vincent Terrapon

CONVERGENCE ANALYSIS OF COUPLING ITERATIONS FOR THE UNSTEADY TRANSMISSION PROBLEM WITH MIXED DISCRETIZATIONS 1530
Azahar Monge, Philipp Birken

USING IMPROVED RADIAL BASIS FUNCTIONS METHODS FOR FLUID-STRUCTURE COUPLING AND MESH DEFORMATION 1545
Giorgos Strofylas, Georgios Mazanakis, Sotirios Sarakinos, Georgios Lygidakis, Ioannis Nikolos

A MONOLITHIC FLUID-STRUCTURE INTERACTION METHOD, APPLICATION TO A PISTON PROBLEM 1564
Felix Ischinger, Martijn Anthonissen, Barry Koren

MS 603: COMPUTATIONAL METHODS IN FLUID-STRUCTURE INTERACTION WITH IMPACT ON INDUSTRIAL APPLICATIONS
SLAMMING IMPACT SIMULATION OF 2D WATER ENTRY FOR RIGID STRUCTURES 1582
Omar Hashim Hassoon, Mostapha Tarfaoui, A. El Malki

A MESH MORPHING BASED FSI METHOD USED IN AERONAUTICAL OPTIMIZATION APPLICATIONS 1594
Matej Andrejašič, David Eržen, Emiliano Costa, Stefano Porziani, Marco Evangelos Biancolini, Corrado Groth

A PATH-CONSERVATIVE OSHER-TYPE SCHEME FOR AXIALLY SYMMETRIC COMPRESSIBLE FLOWS IN FLEXIBLE TUBES 1613
Julia Leibinger, Michael Dumbser, Uwe Iben

MS 605: FRICTIONAL CONTACTS WITH LUBRICATION – BASICS AND APPLICATIONS
LIMITING SHEAR STRESS FORMULATION FOR TEHL SIMULATION 1623
Thomas Lohner, Klaus Michaelis, Karsten Stahl

MS 606: COMPUTATIONAL MODELING OF HYDRAULIC FRACTURING
THE XFEM FOR A SIMPLIFIED MODEL IN HYDRAULIC FRACTURING 1637
Markus Schätzer, Thomas-Peter Fries

SENSITIVITY ANALYSIS OF HYDRAULIC FRACTURING USING AN EXTENDED FINITE ELEMENT METHOD FOR THE PKN MODEL 1647
Hasini Garikapati, Clemens V. Verhoosel, Harald Van Brummelen, Pedro Diez

NUMERICAL MODELLING OF COUPLED HYDRO-THERMAL PROCESSES OF THE SOULTZ HETEROGENEOUS GEOTHERMAL SYSTEM 1659
Musa Dahiru Aliyu, Hua-Peng Chen

MS 607: ADVANCES IN COMPUTATIONAL METHODS FOR LIQUID-VAPOR FLOWS WITH PHASE TRANSFER PROCESSES
A LOW DIFFUSIVE APPROACH FOR TWO PHASE FLOWS 1672
Alberto Beccantini

A LARGE TIME STEP ROE SCHEME APPLIED TO TWO-PHASE FLOW 1693
Sofia Lindqvist, Halvor Lund

MS 608:ADVANCES IN TIME INTEGRATION FOR SOLID, FLUID AND COUPLED SYSTEMS
ENERGY CONSERVING TIME INTEGRATION BASED ON GALERKIN-VARIATIONAL INTEGRATORS WITH CONSTRAINTS 1709
Matthias Bartelt, Michael Groß

MIXED INTEGRATORS FOR STRUCTURE-PRESERVING SIMULATIONS IN NONLINEAR STRUCTURAL DYNAMICS 1736
Alexander Janz, Peter Betsch, Christian Hesch

ENERGY AND MOMENTUM CONSERVING VARIATIONAL BASED TIME INTEGRATION OF ANISOTROPIC HYPERELASTIC CONTINUA 1765
Michael Groß, Rajesh Ramesh, Julian Dietzsch

STRUCTURE-PRESERVING OPTIMAL CONTROL OF DISCRETE MECHANICAL SYSTEMS 1789
Peter Betsch, Christian Becker

VARIATIONAL INTEGRATORS OF MIXED ORDER FOR DYNAMICAL SYSTEMS WITH MULTIPLE TIME SCALES AND SPLIT POTENTIALS 1818
Theresa Wenger, Sina Ober-Blöbaum, Sigrid Leyendecker

CO-SIMULATION OF LARGE MBD, FEA AND DEM SYSTEMS 1832
Jose Ortiz, Neil Macdonald

MS 609: ADVANCED COMPUTATIONAL MODELING OF BATTERIES AND FUEL CELLS
STOCHASTIC 3D MODELING OF AMORPHOUS MICROSTRUCTURES - A POWERFUL TOOL FOR VIRTUAL MATERIALS TESTING 1838
Matthias Neumann, Volker Schmidt

MS 610: NUMERICAL METHODS TO STUDY THE CONTACT MECHANICS OF DRY, ADHESIVE AND LUBRICATED ROUGH SURFACES
PERCOLATION PROPERTIES OF THE FREE VOLUME GENERATED BY TWO ROUGH SURFACES IN CONTACT 1850
Paolo Cinat, Marco Paggi, Claudia Borri

A PARAMETRICALLY TIME-DEPENDENT METHODOLOGY FOR RECIPROCATING CONTACT MECHANICS BETWEEN VISCOELASTIC SOLIDS 1856
Carmine Putignano, Giuseppe Carbone, Daniele Dini

STATISTICAL AND GEOMETRICAL CHARACTERISTICS OF RANDOMLY ROUGH SURFACES USED FOR CONTACT SIMULATIONS. 1864
Rafael Schouwenaars, Carlos G. Figueroa, Víctor Hugo Jacobo, Armando Ortiz

MS 612: NUMERICAL SIMULATIONS FOR SMART-CITY APPLICATIONS
NUMERICAL OPTIMIZATION OF NEAR-ROAD VEGETATION BARRIERS 1876
Ludek Benes, Viktor Sip

MS 615: COMPUTATIONAL MODELS IN MAGNETOHYDRODYNAMICS
LARGE EDDY SIMULATIONS OF MAGNETIC FIELD EFFECT ON TURBULENT FLOW IN A SQUARE DUCT 1886
Jie Mao, Kunlei Zhang, Zhongquan Tan, Ke Liu

700 MATERIALS

MS 701: ADVANCED MATERIALS: COMPUTATIONAL ANALYSIS OF PROPERTIES AND PERFORMANCE
OPTIMAL DESIGN AND ADDITIVE MANUFACTURING OF NOVEL REINFORCING ELEMENTS FOR COMPOSITE MATERIALS 1893
Francesco Fabbrocino, Ilenia Farina, Ada Amendola, Luciano Feo, Fernando Fraternali

A TEMPERATURE SENSITIVE CRYSTAL PLASTICITY MODEL FOR THE PREDICTION OF HIGH TEMPERATURE MECHANICAL BEHAVIOUR OF MULTI-PHASE TIAL ALLOY 1909
M. Umer Ilyas, M. Rizviul Kabir

LAYOUT OPTIMIZATION OF PIEZOELECTRIC ELEMENTS WITH EXTERNAL ELECTRIC CIRCUITS IN SMART CONSTRUCTIONS BASED ON SOLUTION OF THE NATURAL VIBRATIONS PROBLEM 1920
Nataliia Iurlova, Valerii P. Matveenko, Dmitriy A. Oshmarin, Nataliya V. Sevodina, Maksim A. Yurlov

DETERMINATION OF OPTIMAL PARAMETERS FOR A PASSIVE RL-CIRCUIT BY SOLVING THE PROBLEM ON NATURAL VIBRATIONS OF ELECTROELASTIC BODIES 1930
Maksim Yurlov, Valerii P. Matveenko, Dmitriy A. Oshmarin, Nataliya V. Sevodina, Nataliya A. Iurlova

COMPUTATIONAL MODELLING OF MECHANICAL BEHAVIOR OF BIOLOGICAL TISSUES FOR BIOMEDICAL APPLICATIONS 1940
Simin Li, Yang Liu, Juan Du, Begum Zeybek, Lorenzo Zani, Mark P. Lewis, Vadim V. Silberschmidt

ANALYSIS OF FIBRE BRAGG GRATINGS SENSORS OPTIMAL PLACEMENT FOR MONITORING OF DAMAGE PROPAGATION IN LAMINATE COMPOSITES 1950
Mikhail Tashkinov

SEVEN DIFFERENT WAYS TO MODEL VISCOELASTICITY IN A GEOMETRICALLY EXACT SETTING 1959
Alexey Shutov

MS 702: MODELING OF NANOFILLED COMPOSITES
SIMULATING ATOMIC FORCE MICROSCOPY FOR THE DETERMINATION OF THE ELASTIC PROPERTIES OF NANOPARTICLE REINFORCED EPOXY RESIN 1971
Johannes Fankhänel, Andreas Kempe, Raimund Rolfes

MULTI-SCALE FINITE ELEMENT ANALYSIS OF GRAPHENE/POLYMER NANOCOMPOSITES ELECTRICAL PERFORMANCE 1984
Asimina Manta, Matthieu Gresil, Constantinos Soutis

MS 705: IDENTIFICATION OF MATERIAL MODELS
THE BOOTSTRAP APPROACH TO THE STATISTICAL SIGNIFICANCE OF PARAMETERS IN RSM MODEL 2003
Jacek Pietraszek, Leszek Wojnar

MS 706: MODELING OF FIBER-BASED STRUCTURES - TEXTILES AND TEXTILE REINFORCED COMPOSITES
INVERSE IDENTIFICATION OF THE BENDING STIFFNESS OF A BRAIDED POLYETHYLENE TWINE SUBJECT TO LARGE DEFORMATION: APPLICATION TO THE IDENTIFICATION OF THE MESH OPENING RIGIDITY OF FISHING NETS 2010
Barthélémy Morvan, Guilhem Bles, Nicolas Dumergue, Daniel Priour

MULTI-SCALE FRAMEWORK FOR MODELLING CARBON FIBRE WEAVES BASED ON STOCHASTIC REINFORCEMENT GEOMETRY 2029
Andy Vanaerschot, Karen Soete, Stepan Lomov, Dirk Vandepitte

MS 707: MICROMECHANICAL MODELLING: COMPETITION BETWEEN ANALYTICAL AND NUMERICAL METHODS
MULTISCALE MODELING OF THERMO-ELASTIC PROPERTIES OF MICROCRACKED MATERIAL 2039
Antonino Favata, Patrizia Trovalusci, Renato Masiani

EFFECTIVE ELASTICITY COEFFICIENTS IN DRY POROUS MATERIALS. NUMERICAL AND SEMI-ANALYTICAL APPROACHES 2050
Vladimir Sladek, Bruno Musil, Jan Sladek

INFLUENCE OF MICRO CRACKS ON EFFECTIVE MATERIAL PROPERTIES IN FIBER REINFORCED SMART COMPOSITE MATERIALS 2073
Michael Wünsche, Jan Sladek, Vladimir Sladek

DEFECT ACCUMULATION IN NANOPOROUS WEAR-RESISTANT COATINGS UNDER COLLECTIVE RECRYSTALLIZATION. SIMULATION BY HYBRID CELLULAR AUTOMATON METHOD 2080
Dmitry Moiseenko, Pavel Maksimov, Sergey Panin, Viktor Panin

MS 711: FOURIER-BASED METHODS FOR COMPUTING THE BEHAVIOR OF HETEROGENEOUS MATERIALS DEVELOPMENTS, EXTENSIONS AND APPLICATIONS
NONLINEAR COMPOSITE VOXELS AND FFT-BASED HOMOGENIZATION 2099
Matthias Kabel, Andreas Fink, Felix Ospald, Matti Schneider

MS 712: SMART MATERIAL SYSTEMS AND STRUCTURES
A THERMODYNAMICALLY CONSISTENT FINITE STRAIN MICRO-SPHERE FRAMEWORK FOR PHASE-TRANSFORMATION 2110
Richard Ostwald, Thorsten Bartel, Andreas Menzel

OPTIMAL CONTROL TUNNING IN SMART STRUCTURES WITH DELAMINATIONS 2123
Panos Koutsianitis, Amalia Moutsopoulou, Georgios Drosopoulos, Georgios Tairidis, Georgia Foutsitzi, Georgios Stavroulakis

MS 713: MICROSTRUCTURE-BASED MODELLING OF HETEROGENEOUS MATERIALS
DEPENDENCE OF THE MECHANICAL PROPERTIES OF PENTAMODE MATERIALS ON THE LATTICE MICROSTRUCTURE 2134
Ada Amendola, Francesco Fabbrocino, Luciano Feo, Ferdinando Auricchio, Fernando Fraternali

HOMOGENIZATION TECHNIQUES FOR ACCURATE AND APPROXIMATE ESTIMATES FOR OVERALL PROPERTIES OF MICROCRACKED VISCOELASTIC MASONRIES 2151
Amna Rekik, Alain Gasser

APPLICATION OF POLAR ELASTICITY TO THE PROBLEM OF PURE BENDING OF A THICK PLATE 2162
Svitlana Fedorova, Jiří Burša

ARTIFICIAL MICROSTRUCTURE GENERATOR FOR DUAL-PHASE STEELS 2172
Yuliang Hou, Alexandre Dumon, Pièrre Culière, Mohamed Rachik

MS 714: STRENGTH, FATIGUE AND STABILITY OF COMPOSITE STRUCTURES
ON THE STRUCTURAL DESIGN OF IMPERFECTION SENSITIVE LAMINATED COMPOSITE SHELL STRUCTURES SUBJECTED TO AXIAL COMPRESSION 2182
Linus Friedrich, Pawel Lyssakow, Garth Pearce, Martin Ruess, Chiara Bisagni, Kai-Uwe Schröder

STRENGTH AND WEIGHT EQUIVALENT SUBSTITUTION OF LARGE SANDWICH PANELS BY MONOLITHIC CFRP STRUCTURES 2190
Martin Meindlhumer, Martin Schagerl

ROBUST INTEGRAL COMPOSITE AIRCRAFT STRUCTURES 2202
Tamas Havar

MS 715: COMPUTATIONAL ANALYSIS OF COMPOSITE STRUCTURES
MINIMAL MASS DESIGN OF STRENGTHENING TECHNIQUES FOR PLANAR AND CURVED MASONRY STRUCTURES 2210
Gerardo Carpentieri, Francesco Fabbrocino, Mariella De Piano, Valentino Paolo Berardi, Luciano Feo, Fernando Fraternali

MODELING OF FIBER-REINFORCED PLASTICS TAKING INTO ACCOUNT THE MANUFACTURING PROCESS …….. 2220
Cherry Ann Reclusado, Sumito Nagasawa

CONVERGENCE ANALYSIS OF STRESS FIELDS TO HOMOGENIZATION PREDICTIONS IN OPTIMAL PERIODIC COMPOSITE DESIGN …….. 2236
Pedro Coelho, Rui Reis, José Guedes

A NUMERICAL STUDY FOR THE BUCKLING CAPACITY OF WIND TURBINE BLADES: GEOMETRY, LOADING AND MATERIAL INFLUENCE …….. 2247
Efstathios E. Theotokoglou, George A. Balokas, Evgenia K. Savvaki

800 MULTISCALE PROBLEMS

MS 801: MULTISCALE COMPUTATIONAL HOMOGENIZATION FOR BRIDGING SCALES IN THE MECHANICS AND PHYSICS OF COMPLEX MATERIALS

A SEMI-IMPLICIT MICROPLAR DISCRETE-TO-CONTINUUM METHOD FOR GRANULAR MATERIALS …….. 2266
Kun Wang, Waiching Sun

MODELING OF LOW-ALLOYED TRIP-STEELS BASED ON DIRECT MICRO-MACRO SIMULATIONS …….. 2280
Stefan Prüger, Ashutosh Gandhi, Daniel Balzani

HOMOGENIZATION METHOD FOR DESIGNING NOVEL ARCHITECTURED CELLULAR MATERIALS …….. 2292
Zheng-Dong Ma

A HOMOGENIZATION TECHNIQUE FOR ELASTO-PLASTIC COMPOSITES …….. 2316
Federica Covezzi, Stefano De Miranda, Sonia Marfia, Elio Sacco

MS 802: ADVANCES IN THE MODELLING OF MULTI-SCALE, MULTI-PHYSICS AND MULTI-UNCERTAINTY PROBLEMS

THEORETICAL PREDICTION OF PLASTIC INSTABILITY FOR ASYMMETRICAL MATERIALS …….. 2323
Shenghua Wu, Nannan Song, F. M. Andrade Pires

MULTISCALE SIMULATION OF THREE-DIMENSIONAL UNSTEADY GAS FLOWS IN MICROCHANNELS OF TECHNICAL SYSTEMS …….. 2331
Viktoriia Podryga, Yury Karamzin, Tatiana Kudryashova, Sergey Polyakov

HIGHER-ORDER ASYMPTOTIC HOMOGENIZATION OF PERIODIC MATERIALS WITH LOW SCALE SEPARATION …….. 2346
M. Mohammed Ameen, R.H.J. Peerlings, M.G.D. Geers

MS 805: ADVANCED MULTI-PHYSICS AND MULTI-SCALE TECHNIQUES FOR MODELING INELASTIC PROCESSES IN SOLIDS: DAMAGE, FRACTURE AND CONTACT MECHANICS

MESO-SCALE MODELING OF HYBRID INDUSTRIAL/RECYCLED STEEL FIBER-REINFORCED CONCRETE …….. 2353
Antonio Caggiano, Diego Said Schicchi, Guillermo Etse, Enzo Martinelli

MS 806: MULTISCALE MODELLING OF MATERIALS AND STRUCTURES
THE KIRKENDALL SHIFT AND FRENKEL EFFECT DURING MULTI-COMPONENT DIFFUSION PROCESS 2363
Bartek Wierzba

CRACK PATH FIELD AND STRAIN INJECTION TECHNIQUES IN DYNAMIC FRACTURE SIMULATIONS 2372
Oriol Lloberas-Valls, Alfredo E. Huespe, Javier Oliver, Ivo F. Dias

HYBRID FE/XFE FINITE ELEMENT MODEL FOR SIMULATION OF BRITTLE-DUCTILE FRACTURES IN DUAL-PHASE STEEL GRADES. 2390
Konrad Perzynski, Lukasz Madej

EFFECTIVE MESO AND MACRO PROPERTIES FOR FIBRE-REINFORCED-POLYMER CURING COUPLED TO VISCO-ELASTICITY 2398
Rolf Mahnken, Christian Dammann

ON APPLICATION OF THE MESHLESS FINITE DIFFERENCE METHOD TO NUMERICAL HOMOGENIZATION 2417
Irena Jaworska

A MESOSCALE STUDY OF DEFORMATION AND FAILURE OF ANGLE-PLY LAMINATES UNDER TENSILE LOADING BY MEANS OF NUMERICAL HOMOGENIZATION 2426
Marek Romanowicz

MODELING OF PHASE CHANGES IN MICRO-DOMAIN INDUCED BY AN ULTRASHORT LASER PULSE 2437
Ewa Majchrzak, Lukasz Turchan, Jolanta Dziatkiewicz

MS 807: ADVANCED COMPUTATIONAL STRATEGIES FOR MODELLING, SIMULATION AND CHARACTERISATION OF MULTI-SCALE HETEROGENEOUS MATERIALS
VALIDATION OF A DISCRETE CRACK MODEL FOR LIGHTWEIGHT AGGREGATE CONCRETE BEAMS 2449
D. Dias-Da-Costa, R. Graça-E-Costa, R.N.F. Carmo

MS 808: MULTISCALE AND MULTIPHYSICS MODELING OF CEMENTITIOUS MATERIALS
REINFORCEMENT CORROSION IN CONCRETE DUE TO CARBONATION AND CHLORIDE INGRESS UP AND BEYOND INDUCTION PERIOD 2460
Karolina Hájková, Libor Jendele, Vit Šmilauer, T. Sajdlová, Jan Červenka

MS 809: MULTISCALE STOCHASTIC FINITE ELEMENT METHODS
THE STOCHASTIC FINITE ELEMENT METHOD FOR NUCLEAR APPLICATIONS 2471
José David Arregui Mena, Lee Margetts, Llion Evans, D. V. Griffiths, Anton Shterenlikht, Luis Cebamanos, Paul M Mummery

DETERMINATION OF MESOSCALE RANDOM FIELDS FOR THE APPARENT PROPERTIES OF SPATIALLY RANDOM COMPOSITES 2484
Dimitrios Savvas, George Stefanou

MS 811: MULTISCALE MODELING OF CONCRETE AND CONCRETE STRUCTURES
NUMERICAL SIMULATION OF THE AUTOGENOUS SHRINKAGE OF HARDENING PORTLAND CEMENT PASTE ... 2501
Peng Gao, Guang Ye, Jiangxiong Wei, Qijun Yu

900 NUMERICAL SIMULATION METHODS

MS 901: ISOGEOMETRIC METHODS
ANALYSIS OF LAMINATED COMPOSITE PLATES USING ISOGEOMETRIC COLLOCATION METHOD 2514
G. S. Pavan, K. S. Nanjunda Rao

MODELLING STIFFENED LIGHTWEIGHT STRUCTURES WITH ISOGEOMETRIC ANALYSIS VIA MORTAR METHODS 2531
Malte Woidt, Kay Sommerwerk, Matthias C. Haupt, Peter Horst

DOMAIN DECOMPOSITION SOLUTION SCHEMES FOR LARGE-SCALE IGA PROBLEMS 2541
George Stavroulakis, Dimitris Tsapetis, Manolis Papadrakakis

MS 902: INNOVATIVE NUMERICAL APPROACHES FOR MULTI-PHYSICS PROBLEMS
NUMERICAL PROPERTIES OF A DISCONTINUOUS GALERKIN FOMULATION FOR ELECTRO-THERMAL COUPLED PROBLEMS 2558
Lina Homsi, Christophe Geuzaine, Ludovic Noels

MS 903: ADVANCES IN FICTITIOUS DOMAIN METHODS FOR SOLID MECHANICS
TOWARDS LATTICE-BOLTZMANN ON DYNAMICALLY ADAPTIVE GRIDS — MINIMALLY-INVASIVE GRID EXCHANGE IN ESPRESSO 2566
Michael Lahnert, Carsten Burstedde, Christian Holm, Miriam Mehl, Georg Rempfer, Florian Weik

AUTOMATIC CONFORMAL DECOMPOSITION OF ELEMENTS CUT BY NURBS 2591
Jakob W. Steidl, Thomas-Peter Fries

A DESIGN-THROUGH-ANALYSIS APPROACH USING THE FINITE CELL METHOD 2601
Benjamin Wassermann, Tino Bog, Stefan Kollmannsberger, Ernst Rank

MS 905: DISCONTINUOUS GALERKIN METHODS: NEW TRENDS AND APPLICATIONS
DISCONTINUOUS GALERKIN COMPUTATION OF GASEOUS MIXTURE COAXIAL JETS 2614
Nicoletta Franchina, Marco Savini, Francesco Bassi

DISCONTINUOUS GALERKIN SOLUTION OF THE REYNOLDS-AVERAGED NAVIER–STOKES AND KL-KT-LOG(W) TRANSITION MODEL EQUATIONS 2632
A. Bassi, Alessandro Colombo, Antonio Ghidoni, Marco Lorini, Gianmaria Noventa

A STUDY OF MULTIGRID SMOOTHERS USED IN COMPRESSIBLE CFD BASED ON THE CONVECTION DIFFUSION EQUATION 2648
Philipp Birken, Jonathan Bull, Antony Jameson

HIGH-ORDER LINEARLY IMPLICIT TWO-STEP PEER METHODS FOR THE DISCONTINUOUS GALERKIN SOLUTION OF THE INCOMPRESSIBLE RANS EQUATIONS 2664
Francesco Carlo Massa, Gianmaria Noventa, Francesco Bassi, Alessandro Colombo, Antonio Ghidoni, Marco Lorini
OUTPUT-BASED SPACE-TIME ADAPTATION WITH NON-VARIATIONAL TIME INTEGRATION 2684
Krzysztof Fidkowski

MATRIX-FREE MODIFIED EXTENDED BACKWARD DIFFERENTIATION FORMULAE APPLIED TO THE DISCONTINUOUS GALERKIN SOLUTION OF COMPRESSIBLE UNSTEADY VISCOUS FLOWS 2697
Alessandra Nigro, Carmine De Bartolo, Andrea Crivellini, Francesco Bassi

MS 906: MATHEMATICAL ADVANCES IN ISOGEOMETRIC ANALYSIS
PRECONDITIONERS FOR ISOGEOMETRIC ANALYSIS BASED ON SOLVERS FOR SYLVESTER EQUATION 2715
Giancarlo Sangalli, Mattia Tani

MS 907: REGULARIZED ENRICHED APPROXIMATIONS AND QUADRATURE FOR DISCONTINUITIES, SINGULARITIES AND CONTINUOUS-DISCONTINUOUS TRANSITION
EFFECTIVE 3D REGULARIZED XFEM FOR PULL-OUT OF STEEL BARS IN CONCRETE, BENDING AND SHEAR TESTS ON FRP-REINFORCED CONCRETE BEAMS 2726
Elena Benvenuti, Nicola Orlando

MS 908: VERIFICATION AND VALIDATION OF STRUCTURAL MECHANICS SIMULATION MODELS
EXPERIMENTAL VALIDATION OF COMPOSITE STRUCTURES IN EXPLICIT IMPACT ANALYSIS 2734
Konstantinos Fotopoulos, George Lampeas

CODE VERIFICATION IN SHELL ANALYSIS BY THE METHOD OF MANUFACTURED SOLUTIONS 2745
Michael Gfrerer, Martin Schanz

EXPERIMENTAL VALIDATION OF A MODEL UPDATE PROCEDURE FOCUSING ON SMALL GEOMETRIC DEVIATIONS 2757
Thomas Maywald, Arnold Kühhorn, Sven Schrape

OPTIMUM DESIGN, FINITE ELEMENT MODEL UPDATING AND DYNAMIC ANALYSIS OF A FULL LAMINATED GLASS PANORAMIC CAR ELEVATOR 2774
Dimitrios Giagopoulos, Iraklis Chatziparasidis

INVESTIGATION IN THE MECHANICAL BEHAVIOUR OF LOW-VELOCITY IMPACTED CFRP PLATES UNDER DIFFERENT BOUNDARY CONDITIONS 2786
Süleyman Arslan, Zamaan Sadeghi, Athanasios Dafnis, Kai-Uwe Schröder

A NEW SHELL FINITE ELEMENT WITH DRILLING DEGREES OF FREEDOM AND ITS RELATION TO EXISTING FORMULATIONS 2803
Robert Winkler, Dimitrios Plakomytis

MS 910: HIGH ORDER CFD METHODS: CONCLUSIONS AND OUTLOOK
CURVED GRID GENERATION AND DG COMPUTATION FOR THE DLR-F11 HIGH LIFT CONFIGURATION 2843
Ralf Hartmann, Harlan Mcmorris, Tobias Leicht

BS1 - DNS OF THE TAYLOR-GREEN VORTEX AT RE = 1600 2859
Andrea Mastellone, Luigi Cutrone, Francesco Capuano

MS 911: NUMERICAL METHODS IN THE MECHANICS OF GENERALIZED CONTINUA
THE STRUCTURAL SYMMETRY WITHIN THE CONTEXT OF NONLOCAL ELASTICITY 2869
Aurora Angela Pisano, Paolo Fuschi

ISOGEOMETRIC GALERKIN METHODS FOR GRADIENT-ELASTIC BARS, BEAMS, MEMBRANES AND PLATES 2876
Jarkko Niiranen, Sergei Khakalo, Viacheslav Balobanov, Josef Kiendl, Antti H. Niemi, Bahram Hosseini, Alessandro Reali

MS 912: HIGH-ORDER METHODS, SENSITIVITY ANALYSIS AND ADAPTATION FOR THE NAVIER STOKES EQUATIONS
ADAPTIVE MESH REFINEMENT TECHNIQUES FOR HIGH-ORDER FINITE-VOLUME WENO SCHEME 2883
Harshavardhana Srinivasan, Panagiotis Tsoutsanis

IMPLEMENTATION OF A LOW-MACH NUMBER MODIFICATION FOR HIGH-ORDER FINITE-VOLUME SCHEMES FOR ARBITRARY HYBRID UNSTRUCTURED MESHES 2900
Nicholas Simmonds, Panagiotis Tsoutsanis, Adrian Gaylard

APPROPRIATE COMBINATIONS OF CONTROLLER PARAMETERS FOR UNSTEADY FLOW SIMULATIONS WITH ADAPTIVE TIME STEP CONTROL 2916
Kathrin Kozulovic, Graham Ashcroft

MS 913: HIGH-ORDER METHODS FOR POLYGONAL AND POLYHEDRAL MESHES
ADAPTIVE BEM-BASED FEM ON POLYGONAL MESHES FROM VIRTUAL ELEMENT METHODS 2930
Steffen Weißer

VEM AND TOPOLOGY OPTIMIZATION ON POLYGONAL MESHES 2941
Paola F. Antonietti, Matteo Bruggi, Simone Scacchi, Marco Verani

THE VIRTUAL ELEMENT METHOD FOR DISCRETE FRACTURE NETWORK FLOW AND TRANSPORT SIMULATIONS 2953
Matias Fernando Benedetto, Stefano Berrone, Andrea Borio, Sandra Pieraccini, Stefano Scialo

MS 916: DIRECT METHODS FOR LIMIT STATES OF MATERIALS AND STRUCTURES
COMPOSITE FEM MODELS FOR LIMIT AND SHAKEDOWN ANALYSIS 2971
Leonardo Leonetti, Giovanni Garcea, Hung Nguyen-Xuan

ON SHAKEDOWN OF ELASTIC-PLASTIC BODIES WITH TEMPERATURE DEPENDENT PROPERTIES 2983
Michael Peigney

A NUMERICAL STUDY ON THE ENDURANCE LIMIT OF PARTICULATE REINFORCED METAL MATRIX COMPOSITES (PRMMCS) USING THE DIRECT METHOD AND THE STATISTICAL LEARNING 2994
Geng Chen, Bezold Alexander, Christoph Broeckmann, Dieter Weichert

SHAKEDOWN ANALYSIS OF PLATE BENDING UNDER STOCHASTIC UNCERTAINTY BY CHANCE CONSTRAINED PROGRAMMING 3007
Ngoc Trinh Tran, Thanh Ngoc Tran, H.G. Matthies, G.E. Stavroulakis, M. Staat

LIMIT ANALYSIS LOCUS OF HIGH STRENGTH STEEL PLATES UNDER NON-QUADRATIC YIELD CRITERIA 3020
Konstantinos Nikolaou, Christos Bisbos

MS 917: MESOSCOPIC METHODS FOR COMPLEX FLUIDS AND SOFT MATTER
PARTICLE BASED SIMULATION OF FLUID FLOW IN PERIODICALLY GROOVED CHANNELS 3026
Dorothea Kasiteropoulou, Theodoros Karakasidis, Antonios Liakopoulos

MS 918: COMPUTER ALGEBRA SYSTEMS IN MODELLING STATIC AND DYNAMIC PROBLEMS IN MECHANICS OF SOLIDS
HIGH COMPUTATIONAL EFFICIENCY THROUGH GENERIC ANALYTICAL FORMULATION FOR LINEAR SOIL PRESSURE DISTRIBUTION OF RIGID SPREAD RECTANGULAR FOOTINGS 3045
John Bellos, Nikolaos P. Bakas

A REAL-SPACE MODAL ANALYSIS METHOD FOR NON-PROPORTIONAL DAMPED STRUCTURES 3065
Evgueni Stanoev

A FORCE-BASED FORMULATION FOR THE ANALYSIS OF 3-DIMENSIONAL INELASTIC STRUCTURAL FRAMES ... 3091
Theodoros Patsios, Konstantinos Spiliopoulos

THE FEATURES OF ANALYTICAL SOLUTIONS OF BOUNDARY VALUE PROBLEMS OF THE ELASTICITY THEORY FOR FINITE DOMAINS WITH ANGULAR POINTS OF A BOUNDARY 3107
Alexander Kerzhaev, Mikhail Kovalenko, Irina Menshova

MS 919: RECENT ADVANCES IN NUMERICAL SIMULATION AND ANALYSIS OF KINETIC MODELS
DISCRETE VELOCITY MODELS: A STUDY OF THE HYDRODYNAMIC LIMIT 3113
Hans Babovsky

MESO - MACRO MODELS FOR A HARD SPHERE GAS 3121
Sergey Bogomolov, Natalya Esikova, Artem Kuvshinnikov

MS 921: RECENT ADVANCES IN BOUNDARY ELEMENT METHODS
ADVANCES IN THE BOUNDARY ELEMENT METHOD IN GEOMECHANICS 3139
Gernot Beer

ENERGETIC BEM FOR THE NUMERICAL SOLUTION OF DAMPED WAVE PROPAGATION EXTERIOR PROBLEMS 3159
Alessandra Aimi, Mauro Diligenti, Chiara Guardasoni

MODELLING THE PROCESS OF SEQUENTIAL EXCAVATION WITH THE BOUNDARY ELEMENT METHOD 3168
Christian Duenser, Gernot Beer

GREEN'S FUNCTION FOR THE EVALUATION OF ANCHOR LOSSES IN MEMS 3181
Attilio Frangi, Massimiliano Cremonesi

EFFICIENT BOUNDARY ELEMENT FORMULATION OF THERMOELASTICITY 3190
Relindis Rott, Martin Schanz

A CHEBYSHEV INTERPOALTION-BASED FAST MULTIPOLE METHOD FOR POROELASTICITY 3201
Barbara Knöbl, Thomas Traub, Martin Schanz

MS 922: HIGH-ORDER METHODS FOR ELASTIC WAVES AND THEIR APPLICATION
ENERGY STABLE HIGH ORDER FINITE DIFFERENCE METHODS ON STAGGERED GRIDS: AN INITIAL INVESTIGATION 3211
Ossian O'Reilly, Tomas Lundquist, Jan Nordström

MS 923: NOVEL DISCRETIZATION METHODS – MATHEMATICAL AND MECHANICAL ASPECTS
ALTERNATIVE TO RETURN -MAPPING ALGORITH FOR COMPUTING PLASTIC STRAIN. APPLICATION TO DILATANT MATERIALS 3224
Siegfried Maiolino

DISCONTINUOUS GALERKIN METHOD WITH REDUCED INTEGRATION SCHEME FOR THE BOUNDARY TERMS IN ALMOST INCOMPRESSIBLE LINEAR ELASTICITY 3231
Hamid Reza Bayat, Stephan Wulfinghoff, Stefanie Reese

1000 OPTIMIZATION, INVERSE PROBLEMS AND CONTROL

MS 1001: STRUCTURAL AND MULTIDISCIPLINARY OPTIMIZATION
A METHOD FOR BI-DIRECTIONAL COUPLING OF STRUCTURE AND SYSTEM IN THE OPTIMISATION OF MULTI-FUNCTIONAL COMPONENTS 3239
Ajit Panesar, Ian Ashcroft, Ricky Wildman, Richard Hague

SHAPE-TOPOLOGY OPTIMIZATION FOR DESIGNING SHELL STRUCTURES 3249
Hirotaka Nakayama, Masatoshi Shimoda

ON A NASH GAME FOR TOPOLOGY OPTIMIZATION UNDER LOAD-UNCERTAINTY – FINDING THE WORST LOAD 3259
Carl-Johan Thore

ANALYSIS AND DESIGN OF REINFORCED CONCRETE STRUCTURES AS A TOPOLOGY OPTIMIZATION PROBLEM 3271
Matteo Bruggi

ADJOINT OPTIMAL CONTROL PROBLEMS FOR FLUID-STRUCTURE INTERACTION SYSTEMS 3280
Daniele Cerroni, Roberto Da Vià, Sandro Manservisi, Filippo Menghini, Luca Zaniboni

APPLICATION OF WHIRL FLUTTER OPTIMIZATION-BASED SOLUTION TO FULL-SPAN MODEL OF TWIN TURBOPROP AIRCRAFT 3293
Jiri Cecrdle

ON THE CONSIDERATION OF UNCERTAINTY IN DESIGN: OPTIMIZATION – RELIABILITY – ROBUSTNESS 3310
Nicolas Lelièvre, Pierre Beaurepaire, Cécile Mattrand, Nicolas Gayton, Abdelkader Ostmane

A NEW OPTIMIZED ANISOTROPIC MOVING LEAST SQUARES SURROGATE MODEL WITH MAXIMIZED PROGNOSIS 3321
Kevin Cremanns, Dirk Roos

LATIN HYPERCUBE SAMPLING BASED ON ADAPTIVE ORTHOGONAL DECOMPOSITION 3333
Dirk Roos

A BENCHMARK OF CONTEMPORARY METAMODELING ALGORITHMS 3344
Can Bogoclu, Dirk Roos

TOPOLOGY OPTIMIZATION USING A KRIGING-ASSISTED GENETIC ALGORITHM WITH A NOVEL LEVEL SET REPRESENTATION APPROACH 3361
Mitsuo Yoshimura, Koji Shimoyama, Takashi Misaka, Shigeru Obayashi

TOPOLOGY OPTIMIZATION OF CONTACT PROBLEMS BASED ON ALLEN CAHN APPROACH 3377
Andrzej Myslinski

ON THE OPTIMAL DESIGN OF CABLE-STAYED BRIDGES 3386
Gerardo Carpentieri, Mariano Modano, Francesco Fabbrocino, Luciano Feo, Fernando Fraternali

A MIXED-INTEGER LINEAR PROGRAMMING APPROACH FOR GLOBAL DISCRETE SIZE OPTIMIZATION OF FRAME STRUCTURES 3395
Roxane Van Mellaert, Kristo Mela, Teemu Tiainen, Markku Heinisuo, Geert Lombaert, Mattias Schevenels

A SUPER-STRUCTURE BASED OPTIMISATION APPROACH FOR BUILDING SPATIAL DESIGNS 3409
Koen Van Der Blom, Sjonnie Boonstra, Hèrm Hofmeyer, Michael Emmerich

BENDING STIFFNESS OF A MULTILAYERED PLATE 3423
Petr Tovstik, Tatiana M. Tovstik

EFFICIENT SIZING OF STRUCTURES UNDER STRESS CONSTRAINTS 3436
Zhi Hong, Mostafa Abdalla

TOPOLOGY OPTIMIZATION OF WAVE BARRIERS FOR RAILWAY INDUCED VIBRATIONS IN BUILDINGS 3461
Cédric Van Hoorickx, Mattias Schevenels, Geert Lombaert

MS 1002: EVOLUTIONARY ALGORITHMS AND METAHEURISTICS IN CIVIL ENGINEERING AND CONSTRUCTION MANAGEMENT

MULTIOBJECTIVE OPTIMIZATION USING GENETIC ALGORITHMS IN TIME-COST CONSTRUCTION PROJECT SCHEDULING PROBLEM 3473
Jorge Magalhães-Mendes

APPLICATION OF GENETIC ALGORITHMS TO STRUTTED SHEET PILE WALL DESIGN OPTIMIZATION 3485
Mohamed Eid, Remon Isaac

NASH EVOLUTIONARY ALGORITHMS: TESTING PROBLEM SIZE IN RECONSTRUCTION PROBLEMS IN FRAME STRUCTURES 3493
David Greiner, Jacques Periaux, Jose M. Emperador, Blas Galván, Gabriel Winter

OPTIMUM DESIGN OF REINFORCED CONCRETE FRAMES ACCORDING TO EC8 AND MC2010 WITH GENETIC ALGORITHMS 3505
Panagiotis Mergos

CALIBRATION OF THE NUMERICAL MODEL OF A STAND IN DRAGÃO STADIUM BASED ON GENETIC ALGORITHMS 3519
Jorge Leite, Diogo Ribeiro, Hugo Marques, Rui Calçada

MS 1003: ADVANCES IN DESIGN OPTIMIZATION OF STRUCTURES AND MATERIALS
DESIGN OF ISOTROPIC MICROSTRUCTURES VIA A TWO-SCALE APPROACH 3534
Alexis Faure, Georgios Michailidis, Rafael Estevez, Guillaume Parry, Grégoire Allaire

SHAPE OPTIMIZATION OF GRAPHENE SHEETS FOR MAXIMUM FUNDAMENTAL FREQUENCY 3546
Jin-Xing Shi, Masatoshi Shimoda

TOPOLOGY OPTIMIZATION WITH AN IMPLICIT FUNCTION AND PARAMETERIZED CUTTING SURFACE 3553
Peter Dunning

TOPOLOGY OPTIMIZATION OF PERIODIC STRUCTURES FOR COUPLED ACOUSTIC-STRUCTURE SYSTEMS 3565
William Martins Vicente, Renato Picelli, Renato Pavanello, Yi Min Xie

BESO APPROACH TO TOPOLOGY OPTIMIZATION OF GAN PHONONIC CRYSTALS 3583
Luca D'Alessandro, Bichoy Bahr, Luca Daniel, Dana Weinstein, Raffaele Ardito

MS 1004: AERODYNAMIC STRATEGIES FOR THE GLOBAL OPTIMIZATION OF FLYING CONFIGURATIONS IN SUPERSONIC FLOW
LOW BOOM / LOW DRAG SMALL SIZE SUPERSONIC AIRCRAFT DESIGN 3596
Atsushi Ueno, Yasushi Watanabe, Itham Salah El Din, Richard Grenon, Gerald Carrier

DEVELOPMENT OF PHYSICAL METHODS OF THE SUPERSONIC AIRPLANE NEAR-FIELD INVESTIGATION AIMED AT THE SONIC BOOM MINIMIZATION 3612
Sergey Chernyshev, Alexander Ivanov, Andrey Kiselev, Vladimir Mosharov, Leonid Teperin

GLOBAL OPTIMIZATION OF THE SHAPE OF AN AEROSPACE VEHICLE, VIA ITERATIVE OPTIMUM-OPTIMORUM STRATEGY 3621
A. Nastase

MS 1005: MONITORING AND CONTROL OF STRUCTURES
SHAPE IDENTIFICATION ANALYSIS OF CAVITY IN RESIN STRUCTURE BASED ON THERMAL NONDESTRUCTIVE TESTING METHOD 3630
Kotaro Maruoka, Takahiko Kurahashi, Tetsuro Iyama

DESIGN AND INSTALLATION OF A PERMANENT MONITORING SYSTEM FOR PALAZZO LOMBARDIA IN MILANO, ITALY 3640
Marta Berardengo, Alfredo Cigada, Stefano Manzoni, Marcello Vanali

MS 1006: PARAMETER IDENTIFICATION IN SOLID MECHANICS
AN INVERSE PROBLEM STRATEGY BASED ON FORWARD MODEL EVALUATIONS: GRADIENT-BASED OPTIMIZATION WITHOUT ADJOINT SOLVES 3652
Miguel Aguilo

COHESIVE ZONE MODEL IDENTIFICATION WITH DCB TEST PARAMETER ESTIMATION SENSITIVITY 3676
Racine Ly, Julien Jumel, Martin Shanahan, Florian Lavelle

MS 1007: ADDITIVE MANUFACTURING AND OPTIMIZATION
TOPOLOGY OPTIMIZATION FOR ADDITIVE MANUFACTURING WITH CONTROLLABLE SUPPORT STRUCTURE COSTS 3689
Matthijs Langelaar

STUDY OF TOPOLOGY OPTIMIZATION PARAMETERS AND SCAFFOLD STRUCTURES IN ADDITIVE MANUFACTURING 3700
Alain Garaigordobil, Ruben Ansola, Estrella Veguería

FINITE ELEMENT APPROXIMATION OF A TIME-DEPENDENT TOPOLOGY OPTIMIZATION PROBLEM 3711
Matteo Bruggi, Nicola Parolini, Francesco Regazzoni, Marco Verani

DESIGNING MANUFACTURABLE VISCOELASTIC DEVICES USING A TOPOLOGY OPTIMIZATION APPROACH WITHIN A TRULY-MIXED FEM FRAMEWORK 3724
Paolo Venini, Marco Pingaro, Carlo Cinquini

ISO-XFEM FOR TOPOLOGY OPTIMIZATION OF STRUCTURES UNDER MULTIPLE LOAD CASES AND ACCELERATION LOADING 3739
Meisam Abdi, Ian Ashcroft, Ricky Wildman

MS 1008: ULTRASONIC GUIDED WAVES TESTING AND MONITORING
LAMINATE ELEMENT METHOD AND ITS APPLICATION TO THE STUDY OF GUIDED WAVE RESONANCE PHENOMENA IN LAYERED ELASTIC STRUCTURES WITH DEFECTS 3753
Evgeny Glushkov, Natalia Glushkova, Artem Eremin, Rolf Lammering

INVESTIGATIONS OF MECHANICAL GUIDED WAVES PROPAGATION IN PIPES REPAIRED LOCALLY BY COMPOSITE PATCHES 3761
Slah Yaacoubi, Weina Ke, F. Dahmene, M. El Mountassir

MS 1009: ADJOINT METHODS FOR STEADY & UNSTEADY OPTIMIZATION

TWO-STEPS SHAPE OPTIMIZATION ALGORITHM IMPROVING HYDRODYNAMICS STABILITY 3770
Takashi Nakazawa

ON THE CORRECT APPLICATION OF AD CHECKPOINTING TO ADJOINT MPI-PARALLEL PROGRAMS 3778
Ala Taftaf, Laurent Hascoët

SPHERICITY: MESH OPTIMIZATION FOR ARBITRARY ELEMENT TOPOLOGY 3797
Pavlos Alexias, Eugene De Villiers

LINKING PARAMETRIC CAD WITH ADJOINT SURFACE SENSITIVITIES 3812
Ilias Vasilopoulos, Dheeraj Agarwal, Marcus Meyer, Trevor T. Robinson, Cecil G. Armstrong

A CODE-COUPLING APPROACH TO THE IMPLEMENTATION OF DISCRETE ADJOINT SOLVERS BASED ON AUTOMATIC DIFFERENTIATION 3828
Jan Backhaus, Anna Engels-Putzka, Christian Frey

CAD KERNEL AND GRID GENERATION ALGORITHMIC DIFFERENTIATION FOR TURBOMACHINERY ADJOINT OPTIMIZATION 3843
Ismael Sanchez Torreguitart, Tom Verstraete, Lasse Mueller

ALTERNATIVE SOLUTION ALGORITHMS FOR PRIMAL AND ADJOINT INCOMPRESSIBLE NAVIER-STOKES 3858
Mattia Oriani, Guillaume Pierrot

DEFROSTER NOZZLE SHAPE OPTIMIZATION USING THE CONTINUOUS ADJOINT METHOD 3883
Lefki Germanou, Evangelos Papoutsis-Kiachagias, Antoine Delacroix, Kyriakos Giannakoglou

SENSITIVITY ANALYSIS FOR FORCED RESPONSE IN TURBOMACHINERY USING AN ADJOINT HARMONIC BALANCE METHOD 3893
Anna Engels-Putzka, Christian Frey

A TWO-LEVEL HYBRID APPROACH FOR OPTIMAL ACTIVE FLOW CONTROL ON A THREE-ELEMENT AIRFOIL 3912
Anil Nemili, Emre Özkaya, Nicolas R. Gauger, Felix Kramer, Frank Thiele

TRANSITION FROM 2D CONTINUOUS ADJOINT LEVEL SET TOPOLOGY TO SHAPE OPTIMIZATION 3928
J.R.L. Koch, E.M. Papoutsis-Kiachagias, K.C. Giannakoglou

AERODYNAMIC SHAPE OPTIMIZATION USING THE TRUNCATED NEWTON METHOD AND CONTINUOUS ADJOINT 3942
Mehdi Ghavami Nejad, Evangelos M. Papoutsis-Kiachagias, Kyriakos C. Giannakoglou

UNSTEADY ADJOINT TO THE CUT-CELL METHOD USING MESH ADAPTATION ON GPU'S 3957
Konstantinos Samouchos, Stergios Katsanoulis, Kyriakos Giannakoglou

SHAPE OPTIMIZATION OF TURBOMACHINERY ROWS USING A PARAMETRIC BLADE MODELLER AND THE CONTINUOUS ADJOINT METHOD RUNNING ON GPUS 3972
K.T. Tsiakas, F. Gagliardi, X.S. Trompoukis, K.C. Giannakoglou

GEOMETRIC CONTINUITY CONSTRAINTS FOR ADJACENT NURBS PATCHES IN SHAPE OPTIMISATION 3985
Xingchen Zhang, Yang Wang, Mateusz Gugala, Jens-Dominik Mueller

TOWARDS MULTIDISCIPLINARY ADJOINT OPTIMIZATION OF TURBOMACHINERY COMPONENTS 3999
Marc Schwalbach, Tom Verstraete

TOWARDS AN OUTPUT-BASED RE-MESHING FOR TURBOMACHINERY APPLICATIONS 4011
Mateusz Gugala, Marcus Meyer, Jens-Dominik Mueller

OPTIMISATION OF A U-BEND USING A CAD-BASED ADJOINT METHOD WITH DIFFERENTIATED CAD KERNEL 4023
Salvatore Auriemma, Mladen Banovic, Orest Mykhaskiv, Herve Legrand, Jens-Dominik Mueller, Tom Verstraete, Andrea Walther

A CONTINUOUS ADJOINT APPROACH FOR VEHICLE INTERIOR NOISE REDUCTION 4035
Christos Kapellos, Michael Hartmann

A MIXED OPERATOR OVERLOADING AND SOURCE TRANSFORMATION APPROACH FOR ADJOINT CFD COMPUTATION 4047
Zahrasadat Dastouri, Sinan Gezgin, Uwe Naumann

MS 1010: INVERSE PROBLEMS, DESIGN AND OPTIMIZATION

EFFICIENT CALIBRATION OF DISCRETE ELEMENT MATERIAL MODEL PARAMETERS USING LATIN HYPERCUBE SAMPLING AND KRIGING 4061
Michael Rackl, Carolin D. Görnig, Kevin J. Hanley, Willibald A. Günthner

UPDATING PRIOR PARAMETERS BASED ON LIKELIHOOD FUNCTION-BAYESIAN METHOD FOR PARAMETER ESTIMATION AT HIGH MEASUREMENT UNCERTAINTY 4073
Sankaran Somasundharam, K Srinivas Reddy

MS 1011: SURROGATE-ASSISTED EVOLUTIONARY ALGORITHMS IN AERODYNAMIC DESIGN/OPTIMIZATION

CONSTRAINED MULTI-POINT AERODYNAMIC SHAPE OPTIMIZATION OF THE VISCOUS DPW WING THROUGH EVOLUTIONARY PROGRAMMING AND SUPPORT VECTOR MACHINES 4082
Daniel González Juárez, Esther Andrés Pérez, Mario Jaime Martin Burgos

SURROGATE-BASED GLOBAL OPTIMIZATION OF A CYLINDER BY THE USE OF EVOLUTIONARY ALGORITHMS, SUPPORT VECTOR MACHINES AND NON UNIFORM B-SPLINES 4092
Eesther Andrés-Pérez, Daniel González-Juárez, Mario J. Martin-Burgos, Leopoldo Carro-Calvo, Sancho Salcedo-Sanz

PERFORMANCE COMPARISON OF KRIGING AND SVR SURROGATE MODELS APPLIED TO THE OBJECTIVE FUNCTION PREDICTION WITHIN AERODYNAMIC SHAPE OPTIMIZATION 4104
Daniel Viúdez-Moreiras, Esther Andrés-Pérez, Daniel González-Juárez, Mario J. Martin Burgos

HEAT TRANSFER OPTIMIZATION OF A RIBBED SURFACE USING SURROGATE-ASSISTED GENETIC ALGORITHMS 4117
Panagiotis Tsirikoglou, Ghader Ghorbaniasl, Simon Abraham, Chris Lacor

MULTI-FIDELITY EXTENSION TO NON-INTRUSIVE PROPER ORTHOGONAL DECOMPOSITION BASED SURROGATES 4129
Tariq Benamara, Piotr Breitkopf, I. Lepot, Caroline Sainvitu

EVOLUTIONARY AERODYNAMIC SHAPE OPTIMIZATION THROUGH THE RBF4AERO PLATFORM 4146
Dimitrios Kapsoulis, Varvara Asouti, Kyriakos Giannakoglou, Stefano Porziani, Emiliano Costa, Corrado Groth, Ubaldo Cella, Marco Evangelos Biancolini

THE RBF4AERO BENCHMARK TECHNOLOGY PLATFORM 4156
Massimo Bernaschi, Alessandro Sabellico, Giorgio Urso, Emiliano Costa, Stefano Porziani, Fabrizio Lagasco, Corrado Groth, Ubaldo Cella, Marco Evangelos Biancolini, Dimitrios H.Kapsoulis, Varvara G. Asouti, Kyriakos C. Giannakoglou

MS 1013: SOLUTION OF LARGE-SCALE INVERSE PROBLEMS
DATA ASSIMILATION USING MRI DATA 4164
Geir Nævdal, Ove Sævareid, Rolf J. Lorentzen

MS 1014: DESIGN OPTIMIZATION AND INVERSE PROBLEMS FOR WAVE PROPAGATION PROBLEMS
FULL-WAVEFORM INVERSION FOR THE RECONSTRUCTION OF ELECTROMAGNETIC PROPERTY PROFILES USING PLANE MICROWAVES 4178
Namho Joh, Jun Won Kang

1100 REDUCTION METHODS

MS 1101: REDUCED BASIS, POD AND PGD MODEL ORDER REDUCTION TECHNIQUES
REDUCED BASIS METHOD FOR POISSON-BOLTZMANN EQUATION 4187
Cleophas Kweyu, Martin Hess, Lihong Feng, Matthias Stein, Peter Benner

AN ITERATIVE MODEL ORDER REDUCTION SCHEME FOR A SPECIAL CLASS OF BILINEAR DESCRIPTOR SYSTEMS APPEARING IN CONSTRAINT CIRCUIT SIMULATION 4196
Pawan Goyal, Peter Benner

ELECTRICAL FIELDS SIMULATION IN HETEROGENOUS DOMAINS USING THE PROPER GENERALIZED DECOMPOSITION 4213
Chady Ghnatios, Francisco Chinesta, Anais Barasinski

MODEL ORDER REDUCTION IN STRUCTURAL DYNAMICS 4221
Raul Rodriguez Sanchez, Martin Buchschmid, Gerhard Müller

TOWARDS A NON-INTRUSIVE PROPER GENERALIZED DECOMPOSITION SCHEME FOR MODEL ORDER REDUCTION 4248
Xi Zou, Pedro Díez, Michele Conti, Ferdinando Auricchio

A PGD-BASED TIME SPACE DECOMPOSITION FOR THE UNSTEADY NAVIER-STOKES EQUATIONS APPLIED TO INCOMPRESSIBLE FLOWS 4263
Michel Visonneau

MS 1103: MATHEMATICAL SURROGATE MODELLING IN ELECTROMAGNETICS

RATIONAL MODELING OF MULTIVARIATE MULTI-FIDELITY DATA 4284
Elizabeth Rita Samuel, Dirk Deschrijver, Luc Knockaert, Tom Dhaene, Annie Cuyt

ON THE USE OF S-PARAMETER TRANSFORMATIONS TO IMPROVE SURROGATE MODEL BEHAVIOUR OF MULTIPORT NETWORKS 4295
Petrie Meyer

RECENT ADVANCES IN SURROGATE MODELLING OF REFLECTOR ANTENNA SYSTEMS 4302
Dirk De Villiers

A SHORT-TIME PRONY METHOD FOR THE DETECTION OF TRANSIENTS 4313
Annie Cuyt, Wen-Shin Lee, Min-Nan Tsai

SURROGATE MODELING OF ANTENNA RADIATION CHARACTERISTICS BY GAUSSIAN PROCESSES 4322
Jan Pieter Jacobs, Dirk De Villiers

ACCELERATION OF MESH-BASED PHYSICAL OPTICS FOR ELECTROMAGNETIC SCATTERING ANALYSIS 4330
Dao P. Xiang, Matthys M. Botha

DESIGN OF AXIALLY SYMMETRIC POWER COMBINERS USING SURROGATE BASED OPTIMIZATION 4339
Ryno Beyers, Dirk De Villiers

1200 STRUCTURAL STABILITY AND DYNAMICS

MS 1201: COMPUTATIONAL STRUCTURAL DYNAMICS

FLEXURAL VIBRATION ANALYSIS OF GRAPHENE NANOPLATELETS REINFORCED NANOCOMPOSITE BEAMS 4351
Jie Yang, Chuang Feng

DYNAMIC RESPONSE OF REAL OFFSHORE WIND TURBINES ON MONOPILES IN STRATIFIED SEABED 4361
Guillermo M. Álamo, Juan J. Aznárez, Luis A. Padrón, Alejandro E. Martínez-Castro, Rafael Gallego, Orlando Maeso

DYNAMIC ANALYSIS OF SPECIAL CARS ON UNEVEN ROADS 4377
Szymon Tengler

STRUCTURED SPATIAL DISCRETIZATION OF DYNAMICAL SYSTEMS 4398
Marko Jokic, Andrej Jokic, Bruno Dogancic

WAVE PROPAGATION IN THIN PRETWISTED ANISOTROPIC STRIPS 4406
Maloth Thirupathi, Mira Mitra, P J Guruprasad

NUMERICAL ROUTINE FOR DYNAMIC ANALYSIS OF TRANSMISSION LINES GUYED TOWERS SUBMITTED TO BROKEN CABLE 4416
Thiago Brazeiro Carlos, João Kaminski Jr.

INVARIANCE OF EIGENFREQUENCIES AND EIGENMODES UNDER GEOMETRIC TRANSFORMATION IN ELONGATED ELASTIC STRUCTURES 4436
Maryam Morvaridi, Michele Brun

MS 1202: ADVANCED BEAM MODELS

A COMPUTATIONALLY EFFECTIVE FORMULATION TO FINITE ROTATIONS - SMALL STRAINS DESCRIPTION OF BEAM ELEMENTS 4445
Salvatore Lopez

A COROTATIONAL FINITE ELEMENT TO MODEL BENDING VIBRATIONS OF METALLIC STRANDS 4455
Francesco Foti

3D BEAM-COLUMN FINITE ELEMENT UNDER NON-UNIFORM SHEAR STRESS DISTRIBUTION DUE TO SHEAR AND TORSION 4467
Paolo Di Re, Daniela Addessi, Filip C. Filippou

SINGULARLY PERTURBED PROBLEMS IN MECHANICS (SOME FUNDAMENTAL ASPECTS OF DESIGNING AND COMPUTING) 4481
Lyudmila K. Kuzmina

A DEGENERATE-CONTINUUM BASED TIMOSHENKO BEAM APPROACH FOR THE AEROELASTIC ANALYSIS OF THE WIND TURBINE BLADES 4488
Anthoula Panteli, Dimitris Manolas, Konstantinos Spiliopoulos

DYNAMICS OF WIND TURBINE BLADES USING A GEOMETRICALLY-EXACT BEAM FORMULATION 4503
Celso Faccio Junior, Alfredo Gay Neto

EXTENDED FORMULA FOR THE CRITICAL VELOCITY OF A LOAD MOVING ON A BEAM SUPPORTED BY A FINITE DEPTH FOUNDATION 4520
Zuzana Dimitrovová

FIRST-ORDER DISPLACEMENT-BASED ZIGZAG THEORIES FOR COMPOSITE LAMINATES AND SANDWICH STRUCTURES: A REVIEW 4528
Marco Di Sciuva

MS 1203: THE MODELS AND INVESTIGATIONS METHODS OF DYNAMICS OF THE SOLIDS SYSTEMS WITH DRY FRICTION

VIRTUAL TESTBENCH FOR THE OMNI WHEEL DYNAMICS SIMULATION: NEW CONTACT TRACKING ALGORITHM 4553
Ivan Kosenko, Sergey Stepanov, Kirill Gerasimov, Mikhail Stavrovskiy

MODEL AND INVESTIGATION OF DYNAMICS OF SOLID SYSTEM WITH TWO MASSIVE ECCENTRICS ON A ROUGH PLANE 4572
Sergey Semendyaev, Alexey Tsyganov

NON-REGULAR VEHICLE DYNAMICS. APPLICATION TO COLLISION 4584
Ahmed Bouzar Essaidi, Bachir Menkouz, Moussa Haddad, Taha Chettibi

TESTING, SIMULATING AND UNDERSTANDING UNDER-PLATFORM DAMPER DYNAMICS 4599
Chiara Gastaldi, Muzio M. Gola

MS 1204: NONLINEAR DYNAMICS OF ROTATING STRUCTURES
A NEW DYNAMIC SUBSTRUCTURING METHOD FOR NONLINEAR AND DISSIPATIVE SYSTEMS 4611
Colas Joannin, Benjamin Chouvion, Fabrice Thouverez

SOME REMARKS ON TIME INTEGRATION OF 3D ROTOR-STATOR ASSEMBLY 4624
Benoit Prabel

DELAYED FEEDBACK CONTROL METHOD FOR CALCULATING SPACE-TIME PERIODIC SOLUTIONS OF VISCOELASTIC PROBLEMS 4638
Ustim Khristenko, Patrick Le Tallec

COMBINING FINITE ELEMENT ANALYSIS AND ANALYTICAL MODELLING FOR EFFICIENT SIMULATIONS OF NON-LINEAR GEAR DYNAMICS 4655
Shadi Sweiki, Jakub Korta, Antonio Palermo, Rocco Adduci, Domenico Mundo

DAMPING INDUCED BY DRY FRICTION: ANALYSES AND EXPERIMENTS FOR MODELING IMPROVEMENT 4664
Marc-André Douville, Béatrice Faverjon, Eric Chatelet, Georges Jacquet-Richardet

MS 1206: ADVANCES IN NUMERICAL METHODS FOR LINEAR AND NON-LINEAR DYNAMICS AND WAVE PROPAGATION
RAYLEIGH METHOD APPLIED TO A 46-M-HIGH CONCRETE MAST 4674
Alexandre De M. Wahrhaftig, Reyolando M. L. R. F. Brasil

A DISPERSION MINIMIZED MIMETIC METHOD FOR A COLD PLASMA MODEL 4682
Vrushali A. Bokil, Vitaliy Gyrya, Duncan A. Mcgregor

A VOLUME-AGGLOMERATION MULTIRATE TIME ADVANCING APPROACH 4703
Emmanuelle Itam, Stephen Wornom, Bruno Koobus, Alain Dervieux

THERMOMECHANICAL NUMERICAL SIMULATION OF IMPACTS ON ELASTIC-PLASTIC SOLIDS WITH THE FINITE VOLUME METHOD 4717
Thomas Heuzé

MS 1207: COMPUTATIONAL SIMULATION OF SMART STRUCTURES AND MATERIALS
TRANSVERSE WAVE PROPAGATION IN A ONE-DIMENSIONAL STRUCTURE COUPLED TO ITS ELECTRICAL ANALOGUE: COMPARISON OF TRANSFER MATRIX MODELS 4728
Boris Lossouarn, Mathieu Aucejo, Jean-François Deü

MS 1208: BIFURCATIONS AND STABILITY
BUCKLING OF ANNULAR PLATE JOINT WITH CIRCULAR BEAM 4741
Sergei Filippov, Maria Boyarskaya

THE CURL AND FLUTING OF PAPER: THE EFFECT OF ELASTO-PLASTICITY 4752
Anna-Leena Erkkilä, Teemu Leppänen, Tero Tuovinen

STABILITY OF STEEL STRUCTURES WITH CLEARANCES AND IMPERFECTION 4770
Katarzyna Rzeszut, Andrzej Garstecki

MS 1209: DYNAMIC ANALYSIS OF BEAMS UNDER MOVING VEHICLES: APPLICATION TO RAILWAY TRACK MODELLING, DESIGN AND REHABILITATION
NUMERICAL ANALYSIS FOR THE DYNAMIC RESPONSE CHARACTERISTICS OF THE PRESTRESSED CONCRETE SLEEPER 4780
Tsutomu Watanabe, Kodai Matsuoka, Shintaro Minoura

NUMERICAL DYNAMIC ANALYSIS OF BEAMS ON NONLINEAR ELASTIC FOUNDATIONS UNDER HARMONIC MOVING LOAD 4794
Diego Froio, Roberto Moioli, Egidio Rizzi

MS 1210: ADVANCES IN MODELING AND ANALYSIS OF FGM STRUCTURES
ELASTOSTATIC AND MODAL AND BUCKLING ANALYSIS OF SPATIAL FGM BEAM STRUCTURES 4810
Justin Murin, Mehdi Aminbaghai, Juraj Hrabovsky, Vladimir Kutis, Juraj Paulech, Stephan Kugler

THERMO-ELASTICITY IN SHELL STRUCTURES MADE OF FUNCTIONALLY GRADED MATERIALS 4838
Stephan Kugler, Peter Fotiu, Justin Murin

FINITE BEAM ELEMENT WITH PIEZOELECTRIC LAYERS AND FUNCTIONALLY GRADED MATERIAL OF CORE 4863
Vladimír Kutiš, Justín Murín, Juraj Paulech, Juraj Hrabovský, Roman Gogola, Jakub Jakubec

NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS WITH NON-CONSTANT COEFFICIENTS 4878
Juraj Hrabovský, Justín Murín, Mehdi Aminbaghai, Vladimír Kutiš, Juraj Paulech

TWO-WAY COUPLED ELECTRO-THERMAL ANALYSIS OF FGM SYSTEM CALCULATED BY THE NEW LINK FINITE ELEMENT 4888
Juraj Paulech, Vladimír Kutiš, Juraj Hrabovský, Justín Murín, Roman Gogola, Tibor Sedlár, Gabriel Gálik

MS 1211: COMPUTATIONAL STRATEGIES FOR STRUCTURAL ROBUSTNESS ASSESSMENT
PROGRESSIVE COLLAPSE FRAGILITY MODELS OF RC FRAMED BUILDINGS BASED ON PUSHDOWN ANALYSIS 4900
Emanuele Brunesi, Roberto Nascimbene, Fulvio Parisi

MECHANICAL PROPERTIES OF S355 UNDER EXTREME COUPLED EFFECT OF HIGH TEMPERATURES AND HIGH STRAIN RATES 4920
Daniele Forni, Bernardino Chiaia, Ezio Cadoni

IMPLEMENTATION OF PROGRESSIVE DAMAGE IN FINITE ELEMENT CODES FOR THE ASSESSMENT OF ROBUSNTESS 4928
Domenico Asprone, Bernardino Chiaia, Valerio De Biagi, Gaetano Manfredi, Fulvio Parisi

SEISMIC RESIDUAL CAPACITY ASSESSMENT OF FRAMED STRUCTURES DAMAGED BY EXCEPTIONAL ACTIONS 4942
Antonio Formisano, Giuseppe Iazzetta, Giuseppe Marino, Francesco Fabbrocino, Raffaele Landolfo

MS 1212: DYNAMICS AND SEISMIC RESPONSE OF ROCKING AND SELF-CENTERING STRUCTURES
ROCKING RESPONSE OF MASONRY BLOCK STRUCTURES USING MATHEMATICAL PROGRAMMING 4959
Francesco Portioli, Lucrezia Cascini, Raffaele Landolfo

CONTROLLED ROCKING, DISSIPATIVE CONTROLLED ROCKING AND MULTI-HIERARCHICAL ACTIVATION: NUMERICAL ANALYSIS AND EXPERIMENTAL TESTING 4969
Royce Liu, Alessandro Palermo

THE ROLE OF THE PRESTRESSED TENDONS ON THE SEISMIC PERFORMANCE OF HYBRID ROCKING BRIDGE BENTS 4983
Anastasios I. Giouvanidis, Elias G. Dimitrakopoulos

MODELLING CONTACT IN ROCKING STRUCTURES WITH A NONSMOOTH DYNAMICS APPROACH 4997
Anastasios I. Giouvanidis, Elias G. Dimitrakopoulos

A SIMPLE ANALYTICAL MODEL FOR THE ROCKING PREWEC SYSTEM 5010
Dimitrios Kalliontzis, Sri Sritharan

AN ANALYTICAL MODEL FOR DYNAMIC RESPONSE OF AN ELASTIC SDOF SYSTEM FIXED ON TOP OF A ROCKING SINGLE-STORY FRAME STRUCTURE: EXPERIMENTAL VALIDATION 5023
Jonas A. Bachmann, Christoph Jost, Quentin Studemann, Michalis F. Vassiliou, Bozidar Stojadinovic

SMOOTH-ROCKING OSCILLATOR UNDER NATURAL ACCELEROGRAMS 5063
Blerta Lipo, Gianmarco De Felice

EXPERIMENTAL INVESTIGATION OF THE SEISMIC RESPONSE OF A COLUMN ROCKING AND ROLLING ON A CONCAVE BASE 5074
Jonas A. Bachmann, Patrick Blöchlinger, Matthias Wellauer, Michalis F. Vassiliou, Bozidar Stojadinovic

SEISMIC ASSESSMENT OF MASONRY ROCKING COLUMNS AND FRAMES UNDER GROUND MOTION EXCITATIONS 5098
Ioannis Kavvadias, Lazaros Vasiliadis

PARAMETRIC INVESTIGATION OF THE DYNAMIC RESPONSE OF RIGID BLOCKS SUBJECTED TO SYNTHETIC NEAR-SOURCE GROUND MOTION RECORDS 5113
Michalis Fragiadakis, Ioannis Psycharis, Yenan Cao, George P. Mavroeidis

A MACRO-ELEMENT FORMULATION FOR ROCKING FLEXIBLE BODIES WITH A DEFORMABLE BASE 5129
Evangelos Avgenakis, Ioannis N. Psycharis

MS 1213: INNOVATIVE STRUCTURAL SYSTEMS FOR SEISMIC RESISTANT BUILDINGS

NONLINEAR ANALYSIS AND EXPERIMENTAL BEHAVIOUR OF AN INNOVATIVE STEEL FRAME WITH REINFORCED CONCRETE INFILL WALLS 5142
Andrea Dall'Asta, Graziano Leoni, Francesco Morelli, Walter Salvatore, Alessandro Zona

MODAL PROPERTIES AND SEISMIC RESPONSE OF EXISTING BUILDING RETROFITTED BY EXTERNAL BRACINGS WITH VISCOUS DAMPERS 5156
Laura Gioiella, Enrico Tubaldi, Fabrizio Gara, Andrea Dall'Asta

MS 1214: HISTORIC MASONRY STRUCTURES: MODELLING, ASSESSMENT & RETROFIT

SIMPLE CLOSED FORM HOMOGENIZATION MODEL FOR THE NON LINEAR STATIC AND DYNAMIC ANALYSIS OF RUNNING BOND MASONRY WALLS IN AND OUT OF PLANE LOADED 5172
Elisa Bertolesi, Gabriele Milani

CONSTRUCTION PHASES ANALYSIS OF UNREINFORCED MASONRY BUILDINGS THROUGH EQUIVALENT FRAME MODEL 5190
Francesco Pugi, Alessio Francioso, Giacomo Sevieri

A COMBINED EXPERIMENTAL AND NUMERICAL STUDY OF THE PULL-OUT MECHANISM OF THREADED TITANIUM BARS EMBEDDED IN MARBLE BLOCKS 5207
Stavros K. Kourkoulis, Aikaterini Marinelli, Ioanna Dakanali

FINITE-DISCRETE ELEMENT MODELLING OF MASONRY INFILL WALLS SUBJECTED TO OUT-OF-PLANE LOADS 5219
Laura Liberatore, Marta Bruno, Omar Al Shawa, Monica Pasca, Luigi Sorrentino

PALAZZO LA SAPIENZA IN PISA: STRUCTURAL ASSESSMENT AND RETROFIT OF AN HISTORICAL MASONRY BUILDING IN ITALY 5230
Silvia Caprili, Federico Mangini, Nicola Mussini, Walter Salvatore

THREE-DIMENSIONAL NONLINEAR BEHAVIOUR OF MASONRY WALLS MODELLED WITH DISCRETE ELEMENTS 5248
Daniele Baraldi, Antonella Cecchi

NUMERICAL VALIDATION OF EQUIVALENT-FRAME MODELS FOR URM WALLS 5262
Rossella Siano, Guido Camata, Vincenzo Sepe, Enrico Spacone, Pere Roca, Luca Pelà

SEISMIC ASSESSMENT OF HISTORICAL MASONRY TOWERS IN THE NORTH-EAST REGION OF ITALY 5274
Marco Valente, Gabriele Milani

NATIVITY CHURCH IN BETHLEHEM: FULL 3D NON-LINEAR FE APPROACH FOR STRUCTURAL DAMAGE PREDICTION 5287
Gabriele Milani, Marco Valente, Claudio Alessandri

DYNAMIC BEHAVIOUR AND EARTHQUAKE PERFORMANCE OF GREEK BASILICA CHURCHES WITH FOUNDATION DEFORMABILITY 5301
George Manos, Evaggelos Kozikopoulos, Lambros Kotoulas

PROBLEMS RELATED TO THE USE OF FIBER REINFORCED CEMENTITIOUS MATERIALS AS STRENGTHENING OF MASONRY MEMBERS 5314
Daniela Sinicropi, Antonio Borri, Marco Corradi, Michele Paradiso

FAST KINEMATIC LIMIT ANALYSIS OF FRP REINFORCED MASONRY VAULTS THROUGH A NEW GENETIC ALGORITHM NURBS-BASED APPROACH 5326
Andrea Chiozzi, Gabriele Milani, Antonio Tralli

THE ROLE OF RESTORATION MORTARS IN THE EARTHQUAKE PROTECTION OF KAISARIANI MONASTERY 5340
Antonia Moropoulou, Maria Apostolopoulou, Petros Moundoulas, Eleni Aggelakopoulou, Louiza Siouta, Asterios Bakolas, Maria Douvika, Panagiotis Karakitsios, Panagiotis G. Asteris

FIRST RESULTS OF THE VIBRATION-BASED STRUCTURAL HEALTH MONITORING OF A MASONRY DOME 5359
Nicola Cavalagli, Gabriele Comanducci, Massimiliano Gioffrè, Vittorio Gusella, Filippo Ubertini

BEHAVIOR OF SINGLE STORY BEARING WALL MASONRY STRUCTURES IN VARIOUS CONDITIONS 5373
Abdelraouf Kassem

SEISMIC PROTECTION OF MONUMENTS AND HISTORIC STRUCTURES – THE SEISMO RESEARCH PROJECT 5382
Constantine Spyrakos, Charilaos Maniatakis

MS 1215: NONLINEAR VIBRATIONS OF CONSERVATIVE AND NONCONSERVATIVE SYSTEMS: PHENOMENA AND ADVANCED NUMERICAL METHODS

FORCED RESPONSE OF SHROUDED BLADES WITH VARIABLE OPERATING POINTS 5396
Ferhat Kaptan, Lars Panning-Von Scheidt, Jörg Wallaschek, Victor Salit

GEOMETRY EFFECTS ON THE NONLINEAR OSCILLATIONS OF VISCOELASTIC CYLINDRICAL SHELLS 5406
Zenon Del Prado, Marco Amabili, Paulo Gonçalves, Frederico Da Silva

REDUCED ORDER MODELS FOR DYNAMIC BEHAVIOR OF PRESTRESSED ELASTOMER DAMPING DEVICES 5417
Antoine Legay, Jean-François Deü, Benjamin Morin

LOCALIZATION OF VIBRATORY ENERGY OF A LINEAR SYSTEM IN A CHAIN OF FOUR NONLINEAR OSCILLATORS 5427
Simon Charlemagne, Claude-Henri Lamarque, Alireza Ture Savadkoohi

CONTINUATION OF QUASI-PERIODIC SOLUTIONS WITH TWO-FREQUENCY HARMONIC BALANCE METHOD .. 5438
Louis Guillot, Pierre Vigué, Christophe Vergez, Bruno Cochelin

MS 1216: STRUCTURAL ANALYSIS AND VIBRATIONS

ASYMPTOTIC ANALYSIS OF DEFORMATIONS OF THE SLIGHTLY ORTHOTROPIC SPHERICAL LAYER UNDER NORMAL PRESSURE 5444
Andrei L. Smirnov, Svetlana M. Bauer, Liudmila A. Venatovskaia, Eva B. Voronkova

EXPERIMENTAL VALIDATION OF FINITE ELEMENT MODELS FOR REINFORCED CONCRETE BEAMS WITH AND WITHOUT DISCONTINUITIES 5454
Marios Filippoupolitis, Carl Hopkins, Siu-Kui Au

MS 1217: COMPUTATIONAL METHODS IN EARTHQUAKE ENGINEERING AND STRUCTURAL DYNAMICS

SEISMIC BEHAVIOR INDICES OF OLD TYPE REINFORCED CONCRETE MEMBERS 5464
Anthos Ioannou, Stavroula J. Pantazopoulou

EFFECT OF SOIL SPATIAL VARIABILITY ON THE DYNAMIC BEHAVIOR OF A SLOPE 5483
Michael Michael, Tamara Al-Bittar, Abdul-Hamid Soubra

ENERGY-MOMENTUM METHOD FOR NONLINEAR DYNAMIC OF 2D COROTATIONAL BEAMS 5496
Sophy Chhang, Mohammed Hjiaj, Jean-Marc Battini, Carlo Sansour

THE INFLUENCE OF UNCERTAINTY OF DESIGN PARAMETERS ON DYNAMIC CHARACTERISTICS OF STRUCTURE WITH DAMPERS 5507
Magdalena Lasecka-Plura, Roman Lewandowski

FINITE-DISCRETE NUMERICAL MODELLING OF REINFORCED CONCRETE STRUCTURES 5523
Nikolina Zivaljic, Zeljana Nikolic, Hrvoje Smoljanovic, Ivan Balic

BEAM DYNAMIC STRESSES INCREMENTS AFTER PARTIAL DECONSTRUCTION OF FOUNDATION 5533
Vladimir Gordon, Olga Pilipenko, Timur Gasimov

RELIABILITY OF SYSTEMS EQUIPPED WITH VISCOUS DAMPERS WITH VARIABLE PROPERTIES 5550
Andrea Dall'Asta, Laura Ragni, Fabrizio Scozzese, Enrico Tubaldi

AN INVESTIGATION ON EFFECT OF SUBGRADE REACTION MODULUS OF SOIL ON NATURAL PERIOD OF LOW RISE BUILDINGS 5567
Rafet Sisman, Abdurrahman Sahin

A PROBABILISTIC APPROACH TOWARDS AN EVALUATION OF EXISTING CODE PROVISIONS FOR SEISMICALLY ISOLATED STRUCTURES 5574
Anastasios Tsiavos, Bozidar Stojadinovic

EFFECT OF THE SEISMIC EXCITATION'S INCIDENCE ANGLE ON THE NONLINEAR BEHAVIOR OF BASE ISOLATED BUILDINGS CONSIDERING POUNDING TO ADJACENT MOAT WALLS 5585
Eftychia A. Mavronicola, Panayiotis C. Polycarpou, Petros Komodromos

NUMERICAL NON-LINEAR SIMULATION OF THE IN-PLANE BEHAVIOUR OF R/C FRAMES WITH MASONRY INFILLS UNDER SEISMIC TYPE LOADING 5598
George Manos, Vasilios Soulis

MODELING THE MECHANICAL RESPONSE OF A LEAD-CORE BEARING DEVICE: DAMAGE MECHANICS APPROACH 5613
Todor Zhelyazov, Rajesh Rupakhety, Simon Olafsson

INVESTIGATION OF THE PERFORMANCE OF VARIOUS MODAL CORRELATION CRITERIA IN STRUCTURAL DAMAGE IDENTIFICATION 5626
Manolis Georgioudakis, Vagelis Plevris

MS 1218: STABILITY AND CONTROL OF FLEXIBLE STRUCTURES

EXPLOITING MULTIPLE REFERENCE MODELS FOR ADAPTIVE CONTROL OF FLEXIBLE STRUCTURES 5646
Ilaria Venanzi, Laura Ierimonti

PASSIVE CONTROL OF WAVE PROPAGATION IN PERIODIC ANTI-TETRACHIRAL META-MATERIALS 5658
Marco Lepidi, Andrea Bacigalupo

FLEXIBLE COMPLEX SYSTEM OF A DOUBLE-STRING UNDER EXTREME MOVING LOADS 5670
Jarosław Rusin

ON THE MODELING OF SELF-DAMPING IN STRANDED CABLES 5680
Francesco Foti, Luca Martinelli, Federico Perotti

COMPUTATIONAL EFFECTIVENESS OF LMI DESIGN STRATEGIES FOR VIBRATION CONTROL OF LARGE STRUCTURES 5689
Francisco Palacios-Quiñonero, Josep Rubió-Massegú, Josep M. Rossell, Hamid Reza Karimi

A SMART BASE RESTRAINT FOR WIND TURBINES TO MITIGATE UNDESIRED EFFECTS DUE TO STRUCTURAL VIBRATIONS 5702
Nicola Caterino, Christos T. Georgakis, Mariacristina Spizzuoco, Antonio Occhiuzzi

MS 1220: DYNAMICS OF STRUCTURES SUBJECT TO SEISMIC EXCITATIONS

NONLINEAR DYNAMIC BEHAVIOUR OF BASE-ISOLATED BUILDINGS WITH THE FRICTION PENDULUM SYSTEM SUBJECTED TO NEAR-FAULT EARTHQUAKES 5712
Fabio Mazza, Sandro Sisinno

ABSOLUTE FINITE ELEMENT COORDINATES IN THE DYNAMICS OF LARGE FLEXIBLE STRUCTURES 5726
Evtim Zahariev

MODAL ANALYSIS OF A FRAME MODEL UNDER UNMEASURED SEISMIC INPUT 5742
Claudio Valente, Vincenzo Sepe, Marco Di Pilla, Fabrizio Iezzi, Rossella Siano, Luigia Zuccarino

EQUIVALENT UNIAXIAL ACCELEROGRAM FOR CSS-BASED ISOLATION SYSTEMS ASSESSMENT UNDER TWO-COMPONENTS SEISMIC EVENTS 5753
Marco Furinghetti, Alberto Pavese

MS 1221: COMPUTATIONAL STRUCTURAL STABILITY

MIXED SOLID MODELS IN NUMERICAL ANALYSIS OF SLENDER STRUCTURES 5768
Domenico Magisano, Leonardo Leonetti, Giovanni Garcea

COMPUTATIONAL ANALYSIS OF THE COLLAPSE BEHAVIOUR OF THIN-WALLED POLYGONAL STEEL BEAMS 5789
Raffaele Ardito

INFLUENCE OF STRAIN DEFINITIONS ON TRUSSES CRITICAL LOADS 5799
Reyolando Brasil, Jose Balthazar

ULTIMATE STRENGTH ANALYSIS OF STIFFENED PANELS OF SHIP STRUCTURES UNDER COMBINED LOAD 5810
Li Hong, Meng Linghua, Qin Zhongwen, Zhang Enguo, Li Li

MS 1222: INFLUENCE OF LIQUEFIABLE SOIL ON SINGLE AND CLOSELY CLUSTERED STRUCTURES
NUMERICAL ANALYSES OF THE INFLUENCE OF STRUCTURAL SLENDERNESS ON THE SEISMIC RESPONSE OF SINGLE AND CLUSTERED STRUCTURES 5817
Gonzalo Barrios, Tam Larkin, Nawawi Chouw

EFFECT OF HIGHER MODES ON STRUCTURAL RESPONSE WITH NONLINEAR SOIL-FOUNDATION-STRUCTURE INTERACTION 5826
Xiaoyang Qin, Nawawi Chouw, Tam Larkin

MS 1224: INNOVATIVE SOLUTIONS FOR THE SEISMIC PROTECTION OF INDUSTRIAL BUILDINGS
SEISMIC RETROFIT OF AN INDUSTRIAL STRUCTURE THROUGH AN INNOVATIVE SELF-CENTERING HYSTERETIC DAMPER: MODELLING, ANALYSIS AND OPTIMIZATION. 5834
Francesco Morelli, Andrea Piscini, Walter Salvatore

PERFORMANCE ASSESSMENT OF SEISMIC RETROFITTING MEASURES ON SILO STRUCTURES USING INNOVATIVE SEISMIC PROTECTION SYSTEMS 5851
Marius Pinkawa, Benno Hoffmeister, Markus Feldmann

SEISMIC RETROFIT OF INDUSTRIAL SILOS BY MEANS OF BASE ISOLATION DEVICES 5868
Alper Kanyilmaz, Carlo Andrea Castiglioni, Julia Georgi

EFFICIENCY OF SEISMIC ISOLATION ON INDUSTRIAL PLANTS - CASE STUDY OF A GAS TANK 5896
José Henriques, Francesco Morelli, Bram Vandoren, Walter Salvatore, Hervé Degée

PERFORMANCE-BASED NONLINEAR RESPONSE HISTORY ANALYSIS FRAMEWORK FOR THE "PROINDUSTRY" PROJECT CASE STUDIES 5912
Marco Faggella, Raffaele Laguardia, Rosario Gigliotti, Francesco Morelli, Franco Braga, Walter Salvatore

PERFORMANCE BASED EARTHQUAKE ASSESSMENT OF AN INDUSTRIAL SILOS STRUCTURE AND RETROFIT WITH SLIDING ISOLATORS 5926
Edoardo Rossi, Michelangelo Ventrella, Marco Faggella, Rosario Gigliotti, Franco Braga

MS 1225: SEISMIC PERFORMANCE ASSESSMENT OF STRUCTURES AND SEISMIC RISK MITIGATION STRATEGIES
A CRITICAL REVIEW OF FRAGILITY CURVES FOR EXISTING RC BUILDINGS 5934
Monica Mastroberti, Marco Vona

DYNAMIC IDENTIFICATION AND MODELLING OF A RC HOSPITAL BUILDING 5942
Giorgio Lacanna, Pauline Deguy, Maurizio Ripepe, Massimo Baglione, Marco Tanganelli, Stefania Viti, Mario De Stefano

SEISMIC HAZARD OF URBAN AREAS: A CASE-STUDY 5952
Giorgio Lacanna, Pauline Deguy, Maurizio Ripepe, Massimo Coli, Barbara Paoletti, Sara Barducci, Marco Tanganelli, Stefania Viti, Mario De Stefano

SPO2FRAG V1.0: SOFTWARE FOR PUSHOVER-BASED DERIVATION OF SEISMIC FRAGILITY CURVES 5962
Iunio Iervolino, Georgios Baltzopoulos, Dimitrios Vamvatsikos, Roberto Baraschino

COMPUTER-AIDED SEISMIC RISK ASSESSMENT AT URBAN SCALE. MODEL DEFINITION AND VALIDATION ON A CASE STUDY 5977
Alberto Basaglia, Alessandra Aprile, Francesco Pilla, Enrico Spacone

TOWARDS INTEGRATED SEISMIC RISK ASSESSMENT IN PALESTINE - APPLICATION TO THE CITY OF NABLUS ... 5987
Ricardo Monteiro, Paola Ceresa, Vania Cerchiello, Jamal Dabeek, Antonella Di Meo, Barbara Borzi

REASSESS V1.0: A COMPUTATIONALLY-EFFICIENT SOFTWARE FOR PROBABILISTIC SEISMIC HAZARD ANALYSIS 5999
Iunio Iervolino, Eugenio Chioccarelli, Pasquale Cito

SIMPLIFIED PERIOD ESTIMATION OF ITALIAN RC BRIDGES FOR LARGE-SCALE SEISMIC ASSESSMENT 6013
C. Zelaschi, R. Monteiro, R. Pinho

1300 UNCERTAINTY QUANTIFICATION AND ERROR ESTIMATION

MS 1301: THE STOCHASTIC COMPUTER METHODS IN MECHANICS
FLOW FIELD ESTIMATION IN OPEN CHANNEL BASED ON KALMAN FILTER FINITE ELEMENT METHOD 6029
Taichi Yoshiara, Takahiko Kurahashi, Yasuhide Kobayashi, Toshihiko Eto

STOCHASTIC TECHNIQUES FOR THE NUMERICAL SOLUTION OF ENGINEERING BOUNDARY VALUE PROBLEMS 6041
Victor Maceiras, Manuel Casteleiro

THE BOOTSTRAP APPROACH TO THE STATISTICAL SIGNIFICANCE OF PARAMETERS IN THE FIXED EFFECTS MODEL 6061
Jacek Pietraszek, Renata Dwornicka, Agnieszka Szczotok

ON FULLY COUPLED THERMO-ELASTO-PLASTIC STOCHASTIC FINITE ELEMENT ANALYSIS OF STEEL STRUCTURES 6069
Marcin Kaminski, Michał Strakowski

AN INTRODUCTION TO STOCHASTIC FINITE ELEMENT METHOD ANALYSIS OF HYPERELASTIC STRUCTURES ... 6078
Marcin Kamiński, Damian Sokołowski

MS 1303: ANALYSIS AND DESIGN OF SAFETY CRITICAL SYSTEMS UNDER UNCERTAINTY
IDENTIFICATION OF INTERVAL FIELDS FOR SPATIAL UNCERTAINTY REPRESENTATION IN FINITE ELEMENT MODELS 6091
Matthias Faes, Jasper Cerneels, Dirk Vandepitte, David Moens

INTRODUCTION OF BEAD INGS INTO A CRASH TUBE USING A ROBUST OPTIMISATION APPROACH 6099
Stefan Scheiblhofer

DYNAMICAL ANALYSIS OF PLATE MODELS WITH UNCERTAIN STRUCTURAL PROPERTIES USING THE INTERVAL FIELD METHOD 6107
Maurice Imholz, Dirk Vandepitte, David Moens

POWER GRID ROBUSTENSS TO SEVERE FAILURES: TOPOLGICAL AND FLOW BASED METRICS COMPARISION 6121
Roberto Rocchetta, Edoardo Patelli

MS 1304: STOCHASTIC MODELING AND IDENTIFICATION OF UNCERTAINTIES IN COMPUTATIONAL MECHANICS
MULTILEVEL STOCHASTIC REDUCED-ORDER MODEL IN LINEAR STRUCTURAL DYNAMICS FOR COMPLEX STRUCTURES 6136
Olivier Ezvan, Anas Batou, Christian Soize

TO ESTIMATE NON-STATIONARY STOCHASTIC DISTRIBUTED LOAD ON A BEAM STRUCTURE FROM RESPONSE SAMPLES 6148
S.Q. Wu, J. Zhu

A ROBUST POLYNOMIAL CHAOS KALMAN FILTER FRAMEWORK FOR CORROSION DETECTION IN REINFORCED CONCRETE STRUCTURES 6160
Wael Slika, George Saad

ANALYSIS OF STOCHASTIC DYNAMIC SOIL-STRUCTURE INTERACTION PROBLEMS BY MEANS OF COUPLED FINITE ELEMENTS-PERFECTLY MATCHED LAYERS 6172
Manthos Papadopoulos, Stijn François, Geert Degrande, Geert Lombaert

A MULTISCALE FRAMEWORK FOR THE STOCHASTIC ASSIMILATION AND MODELING OF UNCERTAINTY ASSOCIATED NCF COMPOSITE MATERIALS 6186
Loujaine Mehrez, Roger Ghanem, Colin Mcauliffe, William R. Rodgers, Venkat Aitharaju

MS 1306: ERCOFTAC SIG-45: UNCERTAINTY QUANTIFICATION IN CFD AND FLUID STRUCTURE INTERACTION
A MULTI-FIDELITY ADAPTIVE SAMPLING METHOD FOR METAMODEL-BASED UNCERTAINTY QUANTIFICATION OF COMPUTER SIMULATIONS 6196
Riccardo Pellegrini, Cecilia Leotardi, Umberto Iemma, Emilio Campana, Matteo Diez

A LEAST-SQUARES, ADAPTIVE UNCERTAINTY PROPAGATION APPROACH FOR A PLASMA-COUPLED COMBUSTION SYSTEM 6213
Kunkun Tang, Luca Massa, Jonathan Wang, Jonathan Freund

UNCERTAINTY QUANTIFICATION IN NUMERICAL SIMULATIONS OF THE FLOW IN THORACIC AORTIC ANEURYSMS 6226
Alessandro Boccadifuoco, Alessandro Mariotti, Simona Celi, Nicola Martini, Maria Vittoria Salvetti

MS 1307: NON-INTRUSIVE SURROGATE MODELS FOR UNCERTAINTY QUANTIFICATION IN HIGH DIMENSIONS
OPTIMIZATION OF A PHOTOACOUSTIC GAS SENSOR USING MULTIFIDELITY RBF METAMODELING 6250
Cedric Durantin, Justin Rouxel, Jean-Antoine Desideri, Alain Gliere

BEARING CAPACITY OF STRIP FOOTINGS ON SPATIALLY RANDOM SOILS USING KRIGING AND MONTE CARLO SIMULATION 6260
Jawad Thajeel, Tamara Al-Bittar, Nour Issa, Abdul Hamid Soubra

AN OPTIMAL SENSOR PLACEMENT METHOD FOR SHM BASED ON BAYESIAN EXPERIMENTAL DESIGN AND POLYNOMIAL CHAOS EXPANSION 6272
Giovanni Capellari, Eleni Chatzi, Stefano Mariani

CALIBRATION OF NESTED COMPUTER MODELS 6283
Sophie Marque-Pucheu, Guillaume Perrin, Josselin Garnier

A STATISTICAL APPROACH FOR BUILDING SPARSE POLYNOMIAL CHAOS EXPANSIONS 6307
Simon Abraham, Ghader Ghorbaniasl, Chris Lacor

MS 1308: MODELLING AND INVERSE METHODS IN NONLINEAR DYNAMICAL SYSTEMS
FINITE ELEMENT METAMODELING OF UNCERTAIN STRUCTURES 6316
Vasilis Dertimanis, Dimitris Giagopoulos, Eleni Chatzi

DYNAMICS OF MULTIBODY SYSTEMS SUBJECT TO UNILATERAL MOTION CONSTRAINTS USING A NEW NUMERICAL ALGORITHM FOR CONTACT DETECTION 6328
Antonios Pournaras, Fotios Karaoulanis, Sotirios Natsiavas

EQUATIONS OF MOTION FOR MECHANICAL SYSTEMS SUBJECT TO ACATASTATIC CONSTRAINTS 6342
Elias Paraskevopoulos, Sotirios Natsiavas

PARAMETER ESTIMATION OF NONLINEAR LARGE SCALE SYSTEMS THROUGH STOCHASTIC METHODS AND MEASUREMENT OF ITS DYNAMIC RESPONSE 6369
Dimitrios Giagopoulos, Alexandros Arailopoulos

MS 1309: SCALABLE MULTI-FIDELITY MODELING FOR DESIGN, UNCERTAINTY QUANTIFICATION, AND INVERSE PROBLEMS
VARIABLE FIDELITY MODELLING IN MODERN AIRCRAFT DESIGN 6383
Marian Zastawny

MS 1310: COMPUTATIONAL METHODS FOR THE SOLUTION OF STOCHASTIC DIFFERENTIAL EQUATIONS
FOUR-DIMENSIONAL PATH INTEGRATION METHOD FOR ESTIMATING THE STOCHASTIC ROLL RESPONSE 6398
Wei Chai, Arvid Naess, Bernt Leira

SEISMIC RELIABILITY ASSESSMENT OF FRAME STRUCTURES WITH FLEXIBILITY-BASED STOCHASTIC BEAM-COLUMN ELEMENTS 6407
Georgios Balokas, Michalis Fragiadakis

1400 SPECIAL TECHNOLOGICAL SESSIONS

MS 1401: TOICA: THERMAL OVERALL INTEGRATED CONCEPT AIRCRAFT
THERMAL TRADE OFF SUSTAINED BY MULTI DISCIPLINARY AND MULTI LEVEL OPTIMIZATION 6420
Olivier Tabaste, Cedric Campguilhem

COLLABORATIVE ENGINEERING TECHNOLOGIES ENABLING MULTI-PARTNER THERMAL ANALYSIS IN EARLY DESIGN STAGES OF AIRCRAFT 6436
Erik Baalbergen, Wim Lammen, Bert De Wit, Robert Maas, Marie Moghadasi, Johan Kos, Fabio Chiacchio

VISUAL ANALYTICS FOR EVALUATION OF VALUE IMPACT IN ENGINEERING DESIGN 6461
Timoleon Kipouros, Ola Isaksson

STS 3: INNOVATIVE DESIGN OPTIMIZATION TOOLS LINKED TO INDUSTRIAL AERONAUTICAL APPLICATIONS: TARGETING GREENER PERFORMANCES
COMBINING AN RBF-BASED MORPHER WITH CONTINUOUS ADJOINT FOR LOW-SPEED AERONAUTICAL OPTIMIZATION APPLICATIONS 6471
Evangelos Papoutsis-Kiachagias, Matej Andrejasic, Stefano Porziani, Corrado Groth, David Erzen, Marco Evangelos Biancolini, Emiliano Costa, Kyriakos Giannakoglou

STS 5: TRANSITION LOCATION EFFECT ON SHOCK WAVE BOUNDARY LAYER INTERACTION
LOAD CONTROL OF NATURAL-LAMINAR-FLOW WING VIA BOUNDARY LAYER CONTROL 6485
Wieńczysław Stalewski, Janusz Sznajder

STS 7: MORPHING TECHNOLOGIES FOR AIRCRAFT WINGS
AN ADAPTIVE TRAILING EDGE FOR LARGE COMMERCIAL AIRCRAFT 6496
Antonio Concilio, Ignazio Dimino, Rosario Pecora

STS 9: ADVANCED WING HIGH-LIFT SYSTEMS
APPLICATION OF SYNTHETIC JETS ACTUATORS IN WING-PYLON JUNCTION AREA TO IMPROVE THE HIGH LIFT PERFORMANCES 6505
Petr Vrchota

YOUNG INVESTIGATORS MINISYMPOSIUM
A NANO-MACRO BOTTOM-UP APPROACH TOWARDS BRITTLE FRACTURE 6515
Sandeep Patil, Yousef Heider, Carlos Hernandez Padilla, Eduardo Cruz Chu, Bernd Markert

ECCOMAS OLYMPIAD MINISYMPOSIUM
INTEGRATION OF DESIGN AND ANALYSIS THROUGH BOUNDARY INTEGRAL EQUATIONS 6526
Benjamin Marussig, Jürgen Zechner, Gernot Beer, Thomas-Peter Fries

Contributed Sessions

CS 110: NUMERICAL MODELS IN BIOMECHANICS

EVALUATION OF ANGULAR KINEMATICS OF LOWER LIMB AMPUTEES USING QUANTITATIVE FLUOROSCOPIC IMAGING 6539
Alexander Breen, Mihai Dupac

PRELIMINARY STUDY TO INVESTIGATE THE EFFECT OF PISTON-LIKE AND ROCKING MOTIONS OF THE STAPES FOOTPLATE ON THE BASILAR MEMBRANE VIBRATION 6547
Philipp Wahl, Sebastian Ihrle, Pascal Ziegler, Peter Eberhard

IN-HOUSE MULTIBODY HUMAN MODEL BASED ON EULER PARAMETERS FOR THE FAST IMPACT SCENARIO CALCULATION 6556
Jan Spicka, Michal Hajzman, Tomasz Bonkowski

EXPERIMENTAL STUDY ON THE FUZZY-PID HYBRID CONTROL ALGORITHM FOR UNLOADING SYSTEM IN MECHATRONIC DEVICE FOR GAIT RE-EDUCATION 6567
Slawomir Duda, Grzegorz Gembalczyk

FE BONE STRUCTURAL ANALYSIS WITH CT MAPPING OF INHOMOGENEOUS MATERIAL PROPERTIES 6574
Miguel Tobias Bahia, Emílio Graciliano Ferreira Mercuri, Mildred Ballin Hecke

SOLVING THE INVERSE PROBLEM OF ESTIMATING FUZZY VISCOELASTIC CONSTITUTIVE PARAMET ERS 6588
Ruifei Peng, Haitian Yang

DEVELOPMENT OF A CENTRIFUGAL BLOOD PUMP FOR ECMO AND VAD OPERATIONS 6606
Shinhwa Choi, Nahmkeon Hur, Mohammad Moshfeghi, Seongwon Kang, Wonjung Kim

EFFECT OF PERIVASCULAR TISSUE ON INFLATION-EXTENSION BEHAVIOR OF ABDOMINAL AORTA 6616
Tereza Voňavková, Lukáš Horný, Jan Veselý, Tomáš Adámek, Rudolf Žitný

EFFECT OF STERILIZATION ON MECHANICAL PROPERTIES OF BIOLOGICAL COMPOSITE 6625
Jan Vesely, Hynek Chlup, Rudolf Zitny, Tomas Grus

FINITE-ELEMENT MODEL OF INTRAOCULAR PRESSURE MEASUREMENT BY MAKLAKOV APPLANATION TONOMETER 6631
Dmitry V. Franus

CS 210: NUMERICAL SIMULATION OF COMPOSITE MATERIALS

MOMENT-CURVATURE DIAGRAMS FOR EVALUATION OF SECOND ORDER EFFECTS IN RC ELEMENTS 6637
Helena Barros, Carla Ferreira, Tatiana Marques

INVESTIGATIONS ON AN INNOVATIVE METAL-THERMOPLASTIC COMPOSITE ASSEMBLY 6652
Simon Paroissien, Patrick Rozycki, Thierry Renault

A FAILURE CRITERIA FOR UNIDIRECTIONAL FIBER REINFORCED COMPOSITES BASED ON MICROMECHANICS BY ASYMPTOTIC HOMOGENIZATION 6664
Rafael Quelho De Macedo, José Miranda Guedes, Rafael Thiago Luiz Ferreira, Maurício Vicente Donadon

EVOLUTION EQUATIONS BASED APPROACH FOR MODELING FATIGUE IN AMORPHOUS GLASSY POLYMERS. ON THE INVESTIGATION OF FATIGUE DAMAGE DEVELOPMENT IN POLYCARBONATE 6675
Sami Holopainen

CS 211: CRACK PROPAGATION
ULTRASONIC GUIDED WAVES INSPECTION OF PIPES FROM ONE END TO THE FIRST BENT 6688
Mihai Valentin Predoi, Marian Soare, Ovidiu Vasile, Mihail Boiangiu

CS 212: NUMERICAL MODELING OF DAMAGE, FAILURE AND FRACTURE
DUCTILE FRACTURE CRITERIA IN PREDICTION OF SLANT FRACTURE 6699
Petr Kubik, Frantisek Sebek, Jindrich Petruska

DAMAGE MODELING USING STRAIN GRADIENT BASED FINITE ELEMENT FORMULATION 6711
Filip Putar, Jurica Soric, Tomislav Lesicar, Zdenko Tonkovic

RELIABILITY ANALYSIS IN FRACTURE MECHANICS ACCORDING TO COMBINED FAILURE CRITERIA 6721
Rudy Chocat, Paul Beaucaire, Loïc Debeugny, Jean-Pierre Lefebvre, Caroline Sainvitu, Piotr Breitkopf, Eric Wyart

NEW BENCHMARK FOR THE LIFE ASSESSMENT OF A THIN-WALLED PIPE SUBJECTED TO STRESS ASSISTED CORROSION 6734
Yulia Pronina, Elena Sedova

A HIGHER ORDER PHASE-FIELD APPROACH TO FRACTURE FOR FINITE-DEFORMATION CONTACT PROBLEMS 6741
Marlon Franke, Christian Hesch, Maik Dittmann

MODELLING THERMAL BARRIER COATINGS AND THEIR INFLUENCE ON THE LIFETIME OF ROCKET ENGINE NOZZLE STRUCTURES 6764
Marek Fassin, Stephan Wulfinghoff, Stefanie Reese

ANALYTICAL SOLUTION FOR DYNAMIC FRACTURE OF TWO COPLANAR LIMITED-PERMEABLE CRACKS IN MAGNETO-ELECTRO-ELASTIC MATERIAL 6774
Peiwei Zhang

DAMAGE DETECTION OF A BRIDGE BY PARAMETRIC STATISTICAL MOMENT METHOD 6788
Isabella Failla, Nicola Impollonia, Giuseppe Ricciardi

CS 230: MODELLING OF CONCRETE AND MEASURING
A COHESIVE ZONE MODEL FOR THE CHARACTERISATION OF THE INTERFACIAL TRANSITION ZONE (ITZ) BETWEEN CEMENT PASTE AND AGGREGATE 6799
Etienne Malachanne, Marie Salgues, Mouad Jebli, Frederic Jamin

AN EXACT SHEAR STRAIN APPROACH FOR RC FRAME ELEMENTS WITH AXIAL-SHEAR INTERACTION 6810
Alexander Kagermanov, Paola Ceresa

CROSS-SECTIONAL FAILURE CRITERION COMBINED WITH STRAIN-HARDENING DAMAGE MODEL FOR SIMULATION OF THIN-WALLED TEXTILE-REINFORCED CONCRETE SHELLS 6823
Ehsan Sharei, Rostislav Chudoba, Alexander Scholzen

SEISMIC ASSESSMENT OF A 5-STOREY RETROFITTED RC BUILDING 6832
Hussein Bark, George Markou, Christos Mourlas, Manolis Papadrakakis

CS 310: CAD, CAM AND CAE
VECTOR FIELD GUIDED TOOL PATHS FOR FIVE-AXIS MACHINING 6847
Stanislav Makhanov

AN 'A PRIORI' MODEL REDUCTION FOR ISOGEOMETRIC BOUNDARY ELEMENT METHOD 6882
Shengze Li, Jon Trevelyan, Weihua Zhang, Zhuxuan Meng

RESEARCH OF DEFORMATION QUALITIES OF POROUS MATERIALS ENHANCING COMPUTER DESIGN OF WETSUIT 6892
Irina Cherunova, Elena Sirota, Nikolai Kornev, Mathias Paschen, Sebastian Schreier, Tatyana Lesnikova, Pavel Cherunov

MATHEMATICAL MODELLING OF CITY AERODYNAMICS 6900
Svetlana Valger, Natalya Fedorova, Alexander Fedorov

CS 320: GRID GENERATION AND ADAPTIVE TECHNIQUES
A STABLE AND CONSERVATIVE TIME-DEPENDENT INTERFACE FORMULATION ON SUMMATION-BY-PARTS FORM: AN INITIAL INVESTIGATION 6911
Samira Nikkar, Jan Nordström

A LAPLACIAN MESH DEFORMATION TECHNIQUE FOR SIMULATION-DRIVEN DESIGN OPTIMIZATION 6928
Mario J. Martin-Burgos, Daniel González-Juárez, Esther Andrés-Pérez

AN INTEGRATED FRAMEWORK FOR WRAPPING AND MESH GENERATION OF COMPLEX GEOMETRIES 6938
David Martineau, Jeremy Gould, Jacques Papper

CS 410: COMPUTATIONAL FLUID MECHANICS
THE VORTICITY CREATION PROCESS AT PHYSICAL SURFACES 6955
Horia Dumitrescu, Vladimir Cardos

HEAT RECOVERY BY CROSS FLOW 6966
Karel Adamek, Jan Kolar, Pavel Peukert

THE INLET AND OUTLET BOUNDARY PROBLEM WITH THE PREFERENCE OF MASS FLOW 6979
Martin Kyncl, Jaroslav Pelant

BUDGET ANALYSIS OF TURBULENT KINETIC ENERGY IN CORNER SEPARATION : RANS VS LES 6999
Jean-François Monier, Feng Gao, Jérôme Boudet, Liang Shao, Lipeng Lu

A SHALLOW WATER EQUATION BASED ON DISPLACEMENT AND PRESSURE AND THE ZU-CLASS SYMPLECTIC METHOD 7014
Feng Wu, Wan-Xie Zhong

IMMERSED BOUNDARY METHODS FOR COMPRESSIBLE LAMINAR FLOWS 7029
Rakesh Ramakrishnan, Anant Girdhar, Santanu Ghosh

IMPACT OF EXTERNAL SURROUNDINGS ON NATURAL CONVECTION IN A VERTICAL CHANNEL ASYMMETRICALLY HEATED 7045
Delphine Ramalingom, Alain Bastide

INFLUENCE OF ADJACENT ROOMS ON THE DEVELOPMENT OF GAS EXPLOSION 7056
Iurii Polandov, Sergei Dobrikov

DOMAIN DECOMPOSITION APPROACH FOR NEAR-WALL TURBULENCE MODELING 7066
Sergey Utyuzhnikov

LARGE EDDY SIMULATION OF SALTATION OVER GAUSSIAN HILLS 7074
Gang Huang, Catherine Le Ribault, Serge Simoëns, Ivana Vinkovic, J.M.Vignon

AN EXPERIMENTAL AND NUMERICAL STUDY ON THE PERFORMANCE OF AN INNOVATIVE VERTICAL-AXIS WIND TURBINE 7087
Katarzyna Kludzinska, Krzysztof Tesch, Piotr Doerffer

NUMERICAL SIMULATION OF NONLINEAR FREE SURFACE WATER WAVES USING A COUPLED POTENTIAL FLOW-URANS/VOF APPROACH 7095
Bulent Duz, Tim Bunnik, Geert Kapsenberg

THE OPTIMAL CONTROL OF A MULTI-MASS VIBRATION PROPULSION SYSTEM IN A VISCOUS INCOMPRESSIBLE FLUID 7121
Artem Nuriev, Olga Zakharova

HIGH-RESOLUTION SIMULATION OF INTERNAL WAVES ATTRACTORS AND IMPACT OF INTERACTION OF HIGH AMPLITUDE INTERNAL WAVES WITH WALLS ON DYNAMICS OF WAVES ATTRACTORS 7130
Ilias Sibgatullin, Michael Kalugin

NUMERICAL SIMULATION OF HEAT AND MASS TRANSFER PROCESSES IN LARGE-SCALE FLUIDIZED BED COMPLEX STRUCTURE APPARATUS AS AN EXAMPLE OF THE REACTOR OF ISOPARAFFINS DEHYDROGENATION 7138
Sergei Solovev, Svetlana Egorova, Alexander Lamberov, Olga Soloveva

CS 420: MULTI-PHASE AND CHEMICALLY REACTING FLOWS

SHOCK CAPTURING COMPUTATIONS WITH STABILIZED POWELL-SABIN ELEMENTS 7149
Giorgio Giorgiani, Herve Guillard, Boniface Nkonga

NUMERICAL INVESTIGATION OF METASTABLE CONDENSING FLOWS WITH AN IMPLICIT UPWIND METHOD ... 7165
Lucia Azzini, Teus Van Der Stelt, Matteo Pini

LEVEL SET METHOD FOR SIMULATING THE INTERFACE KINEMATICS: APPLICATION OF A DISCONTINUOUS GALERKIN METHOD 7182
Roozbeh Mousavi, Florian Kummer, Martin Oberlack, Peter F. Pelz

LARGE EDDY SIMULATION OF TURBULENT COMPRESSIBLE FLOWS USING THE CHARACTERISTIC BASED SPLIT SCHEME AND MESH ADAPTATION 7222
Renato Linn, Armando Awruch

NUMERICAL SIMULATION OF HYDROGEN JET INJECTION AND IGNITION IN SUPERSONIC FLOW 7237
Yulia Zakharova, Natalya Fedorova, Svetlana Valger, Marat Goldfeld, Olga Vankova

A 3D UNSTEADY NUMERICAL SIMULATION OF THE REFRIGERANT FLOW EVAPORATION IN A PLATE HEAT EXCHANGER 7247
Mirza Popovac, Gerwin Schmid, Michael Lauermann

RESEARCH OF UNSTEADY FLOW REGIMES IN CHANNEL OF HYPERSONIC INLET 7255
Natalya Fedorova, Marat Goldfeld

EXPLICIT DARCY'S LAW BOUNDARY CONDITION WITH COMBINED CONTINUUM AND DISCRETE MODEL FOR PRESSURE DRIVEN MEMBRANE APPLICATIONS 7266
Tomi Naukkarinen, Teemu Turunen-Saaresti

CS 450: NUMERICAL METHODS AND CONVERGENCE ACCELERATION IN CFD

A MULTIGRID FORMULATION FOR FINITE DIFFERENCE METHODS ON SUMMATION-BY-PARTS FORM: AN INITIAL INVESTIGATION 7274
Andrea Alessandro Ruggiu, Per Weinerfelt, Tomas Lundquist, Jan Nordström

IMPROVED DUAL TIME–STEPPING USING SECOND DERIVATIVES 7285
Jan Nordström, Andrea Ruggiu

A DIRECT SOLVER FOR THE ADVECTION-DIFFUSION EQUATION USING GREEN'S FUNCTIONS AND LOW-RANK APPROXIMATION 7302
Jonathan Bull

AN IMPROVED DISCRETIZATION METHOD FOR BOUNDED CONVECTIVE SCHEMES ON UNSTRUCTURED CO-LOCATED GRIDS 7317
Vlado Przulj

A LOW-MACH, LOW-REYNOLDS PRECONDITIONING SCHEME WITH PARTICULAR ATTENTION TO VISCOUS TIME-STEPPING 7334
Jens Fiedler, Graham Ashcroft

PERFORMANCE IMPROVEMENT OF FLOW COMPUTATIONS WITH AN OVERSET-GRID METHOD INCLUDING BODY MOTIONS USING A FULL MULTIGRID METHOD 7347
Kunihide Ohashi, Hiroshi Kobayashi

SOLVING LINEAR SYSTEMS WITH MULTIPLE RIGHT-HAND SIDES WITH GMRES : AN APPLICATION TO AIRCRAFT DESIGN 7358
Aloïs Bissuel, Grégoire Allaire, Laurent Daumas, Frédéric Chalot, Michel Mallet

COMPARISON OF DIFFERENT SPATIAL/ANGULAR AGGLOMERATION MULTIGRID SCHEMES FOR RADIATIVE HEAT TRANSFER COMPUTATIONS 7372
Georgios Lygidakis, Ioannis Nikolos

CS 460: UNSTEADY FLOW COMPUTATION
LINEARISED FREQUENCY DOMAIN GUST ANALYSIS OF LARGE CIVIL AIRCRAFT 7390
Philipp Bekemeyer, Reik Thormann, Sebastian Timme

CONSISTENT NON-REFLECTING BOUNDARY CONDITIONS FOR BOTH STEADY AND UNSTEADY FLOW SIMULATIONS IN TURBOMACHINERY APPLICATIONS 7403
Daniel Schlüß, Christian Frey, Graham Ashcroft

REDUCED ORDER MODELLING OF GUST ANALYSIS USING COMPUTATIONAL FLUID DYNAMICS 7423
Reik Thormann, Philipp Bekemeyer, Sebastian Timme

ON THE REGULARISATION OF NON-REFLECTING BOUNDARY CONDITIONS NEAR ACOUSTIC RESONANCE 7435
Christian Frey, Hans-Peter Kersken

ON GRID RESOLUTION REQUIREMENTS FOR LES OF WALL-BOUNDED FLOWS 7454
Saleh Rezaeiravesh, Mattias Liefvendahl, Christer Fureby

A HYBRID MESH HARMONIC BALANCE SOLVER FOR THE AEROELASTIC ANALYSIS OF TURBOMACHINERY 7466
Hans-Peter Kersken, Graham Ashcroft, Christian Frey

NUMERICAL AND EXPERIMENTAL INVESTIGATIONS OF DYNAMIC STALL PROBLEM ON A 2D HELICOPTER ROTOR BLADE SECTION 7477
Antonello Marino, Serena Russo, Giovanni Paolo Reina, Gennaro Esposito, Francesco Capizzano, Carlo De Nicola

INVESTIGATION OF A MODERN & HYBRID TURBULENCE MODELLING APPROACH FOR TRANSIENT AUTOMOTIVE AERODYNAMICS SIMULATIONS 7488
Rene Devaradja, Petr Simanek, Jacques Papper, Pavla Policka

FREE-SURFACE FLOW SIMULATIONS FOR MOORED AND FLOATING OFFSHORE PLATFORMS 7515
Arthur Veldman, Roel Luppes, Peter Van Der Plas, Henri Van Der Heiden, Bulent Duz, Henk Seubers, Joop Helder, Tim Bunnik, Rene Huijsmans

MODELLING OF UNSTEADY SECONDARY VORTICES GENERATED BEHIND THE RADIAL GAP OF THE AXIAL TURBINE BLADE WHEEL 7532
Petr Straka

PATHWAYS TO IMPROVED AERODYNAMIC DESIGN 7544
James Page, Paul Hield, Callum Mantell, Paul Tucker

SIMULATIONS OF BLAST WAVE PROPAGATION IN OPEN AND CLOSED SPACE 7557
Natalya Fedorova, Svetlana Valger, Yulia Zakharova

CS 500: HIGH PERFORMANCE COMPUTING
ON A DYNAMIC SCHEDULING ALGORITHM FOR MASSIVELY PARALLEL COMPUTATIONS OF ATOMIC ISOTOPE 7572
Elizaveta Dorofeeva, Jatin Arora, Stefan Typel, Gevorg Poghosyan, Peter Sanders, Achim Streit

APPLICATION THE SPARSE MATRIX METHOD TO CALCULATE THE METAL ELASTIC STRESS-STRAIN STATE, USING THE FINITE ELEMENT METHOD 7581
Inna Barankova, Uliana Mikhailova, Victor Demidovich

CS 610: AERO-ACOUSTICS
FLUID ANALYSIS USING FICTITIOUS DOMAIN FINITE ELEMENT METHOD 7589
Yukihiro Terakado, Tahahiko Kurahashi

ON THE SIMULATION OF AERODYNAMIC NOISE WITH DIFFERENT TURBULENCE MODELS 7599
Xin Huang, Michael Schäfer

NUMERICAL INVESTIGATION OF THERMOACOUSTIC ENGINE USING IMPLICIT LARGE EDDY SIMULATION 7609
Mustapha Mahdaoui, Rachid Bennacer, Smaïne Kouidri, N. Martaj

SEMI-DISCRETE COINCIDENCE IN THE MID-FREQUENCY SOUND TRANSMISSION THROUGH RIB-STIFFENED PANELS 7620
Edwin Reynders, Cédric Van Hoorickx, Arne Dijckmans

CS 620: COMPUTATIONAL CONTACT MECHANICS
SOLVING UNILATERAL CONTACT PROBLEMS IN MULTIBODY DYNAMICS USING A PRIMAL-DUAL INTERIOR POINT METHOD 7628
Dario Mangoni, Alessandro Tasora

CS 630: SIMULATION OF FLUID-STRUCTURE INTERACTION
3D DRBEM MODELING FOR ROTATING INITIALLY STRESSED ANISOTROPIC FUNCTIONALLY GRADED PIEZOELECTRIC PLATES 7640
Mohamed Abdelsabour Fahmy

DIRECT EDDY CURRENT METHOD FOR VOLUMETRIC FLAWS OF CYLINDRICAL SHAPE 7659
Valentina Koliskina, Andrei Kolyshkin, Rauno Gordon, Olev Märtens

NUMERICAL SIMULATION OF BUOYANT PLUMES USING A FIXED POINT ITERATIVE METHOD 7666
Blanca Bermúdez, Alejandro Rangel-Huerta, W. Fermín Guerrero S., José David Alanis Urquieta

INVESTIGATION OF THE FLOW OVER AN OSCILLATING CYLINDER WITH A VERY LARGE EDDY SIMULATION MODEL 7674
Anastasia Kondratyuk, Michael Schäfer, Awais Ali

NUMERICAL STUDY OF TWO OPTIMIZED COUPLING INTERFACE TREATMENTS FOR STEADY CONJUGATE HEAT TRANSFER PROBLEMS 7682
Marc Paul Errera, Roch Roukoz El Khoury

EFFECT OF DISC GEOMETRY ON THE DYNAMIC STABILITY OF DIRECT SPRING OPERATED PRESSURE RELIEF VALVES 7695
István Erdődi, Csaba Hős

CFD SIMULATION OF THERMO- AERODYNAMIC INTERACTION IN A SYSTEM HUMAN-CLOTH-ENVIRONMENT UNDER VERY LOW TEMPERATURE AND WIND CONDITIONS 7703
Irina Cherunova, Sina Samarbaksch, Nikolai Kornev

CS 720: COMPUTATIONAL MATERIALS SCIENCE
ON THE INFLUENCE OF THE INCLUSIONS' MORPHOLOGY ON THE ACCURACY OF THE PREDICTION OF REINFORCED COMPOSITES MECHANICAL BEHAVIOUR 7711
Viwanou Hounkpati, Vladimir Salnikov, Philippe Karamian-Surville, Alexandre Vivet

NUMERICAL SIMULATION OF ADVANCED PULTRUSION PROCESSES WITH MICROWAVE HEATING 7720
Evgeny Barkanov, Pavel Akishin, Rudolf Emmerich, Matthias Graf

EFFICIENT NUMERICAL SIMULATION OF INDUSTRIAL SHEET METAL BENDING PROCESSES 7739
Christian Zehetner, Paula Reimer, Franz Hammelmüller, Hans Irschik, Wolfgang Kunze

NUMERICAL MODELLING AND SIMULATION OF SHEET METAL CUTTING PROCESSES 7749
Paula Reimer, Christian Zehetner, Franz Hammelmüller, Wolfgang Kunze

SIMULATION OF PRECIPITATION IN V-CONTAINING HSLA STEEL FOR THE STRENGTHENING ENHANCEMENT 7757
Piyada Suwanpinij, Paolo Massaro, Annalisa Pola, Prasonk Srichareonchai

CS 750: COMPUTATIONAL MODELING OF COMPOSITES
NUMERICAL ANALYSIS OF SANDWICH PANELS SUBJECTED TO POINT LOADS 7765
Zbigniew Pozorski, Jolanta Pozorska

BOND OF REINFORCEMENT IN REACTIVE POWDER CONCRETE: EXPERIMENTAL STUDY 7771
Mingde Sun, Ri Gao, Aili Li, Yongjing Wang

MODELLING SOIL-FIBRE COMPOSITE BEHAVIOUR USING A MICROMECHANICAL APPROACH 7779
Thomas Bower, Anthony Jefferson, Peter Cleall, Paul Lyons

NUMERICAL SIMULATION OF MECHANICAL BEHAVIOR OF WOVEN COMPOSITE AT DIFFERENT STRAIN RATE BY A COLLABORATIVE ELASTO-PLASTO-DAMAGE MODEL WITH FRACTIONAL DERIVATIVES 7786
Alina Krasnobrizha, Patrick Rozycki, Laurent Gornet, Pascal Cosson

NUMERICAL SIMULATION OF MECHANICAL BEHAVIOUR OF COMPOSITE SANDWICH PANELS WITH DEFECTS 7800
Aleksandr Anoshkin, Valerii Zuiko, Mikhail Alikin, Anna Tchugaynova

TIMOSHENKO BEAM ELEMENT WITH ANISOTROPIC CROSS-SECTIONAL PROPERTIES ……. 7810
Alexander R. Stäblein, Morten H. Hansen

STABILIZATION OF THE NUMERICAL PRIMING TO STABILIZATION PHENOMENON IN THE CRUSHING OF A COMPOSITE PANEL USING LOCAL AND DIFFUSE DAMAGE LAWS …….. 7820
Christine Espinosa, Floran Tostain, Samuel Rivallant

STRUCTURAL ANALYSIS OF ADHESIVE BONDING FOR THICK-WALLED TUBULAR COMPOSITE PROFILES …….. 7837
Geminiano Mancusi, Agostina Orefice, Luciano Feo, Fernando Fraternali

CS 751: SMART MATERIALS AND STRUCTURES
TOWARDS REAL-TIME STRUCTURAL HEALTH MONITORING DAMAGE DETECTION WITHOUT USER INPUT ….. 7853
Mohmmad Salmanpour, Zahra Sharif Khodaei, Ferri Aliabadi

FIBROUS TECTONICS:A RETHINKING OF COMPOSITE PRODUCTION THROUGH INNOVATION AND EXPLORATION OF MOLDING TECHNIQUES AND METHODOLOGIES …….. 7863
David Costanza

CS 830: COMPUTATIONAL NANOTECHNOLOGY
EFFECT OF NANOSIZED ASPERITIES AT THE SURFACE OF A NANOHOLE …….. 7875
Mikhail Grekov, Aleksandra Vakaeva

CS 840: MULTI-SCALE COMPUTATIONAL METHODS FOR SOLIDS AND FLUIDS
THE GEMA FRAMEWORK – AN INNOVATIVE FRAMEWORK FOR THE DEVELOPMENT OF MULTIPHYSICS AND MULTISCALE SIMULATIONS …….. 7886
Carlos Augusto Teixeira Mendes, Marcelo Gattass, Deane Roehl

NONLINEAR PATCH METHOD AND APPLICATION …….. 7895
Antoine Brunet, Pierre Sarrailh, François Rogier, Jean-François Roussel, Denis Payan

ASSESSMENT OF A CONTINUUM MICROMECHANICS-BASED MULTISCALE MODEL FOR CONCRETE BY MEANS OF SENSITIVITY ANALYSIS AND UNCERTAINTY PROPAGATION …….. 7905
Luise Göbel, Andrea Osburg, Tom Lahmer

SPATIAL CLUSTERING STRATEGIES FOR HIERARCHICAL MULTI-SCALE MODELLING OF METAL PLASTICITY …… 7922
Md Khairullah, Jerzy Gawad, Dirk Roose, Albert Van Bael

BLACK-BOX SOLVER FOR ONE-DIMENSIONAL MULTISCALE MODELLING USING THE QTT FORMAT …….. 7938
Ivan Oseledets, Maxim Rakhuba, Andrei Chertkov

SCALE EFFECT IN MICROFLOWS MODELLING WITH THE MICROPOLAR FLUID THEORY …….. 7948
Anna Kucaba-Pietal

CS 930: HIGH-ORDER DISCRETIZATION METHODS

A COMPARISON OF VARIOUS NODAL DISCONTINUOUS GALERKIN METHODS FOR THE 3D EULER EQUATIONS 7956
Michael Bergmann, Svetlana Drapkina, Graham Ashcroft, Christian Frey

A 3D ISOGEOMETRICAL BOUNDARY ELEMENT ANALYSIS FOR NON-LINEAR GRAVITY WAVE PROPAGATION ... 7967
Jorge Maestre, Jordi Pallarés, Ildefonso Cuesta

HYBRID OPENMP/MPI PARALLELIZATION OF A HIGH–ORDER DISCONTINUOUS GALERKIN CFD SOLVER 7992
Francesco Bassi, Alessandro Colombo, Andrea Crivellini, Matteo Franciolini

ON THE CHOICE OF SHOCK CAPTURING SCHEMES FOR THE SOLUTION OF THE LWR TRAFFIC FLOW EQUATION USING A HIGH ORDER MODAL DISCONTINUOUS GALERKIN DISCRETIZATION 8013
Alberto Costa Nogueira Junior, João Lucas De Sousa Almeida, Cláudio Alessandro De Carvalho Silva

COMPARISON OF FINITE VOLUME HIGH-ORDER SCHEMES FOR THE WO-DIMENSIONAL EULER EQUATIONS 8033
Jens Wellner

NURBS-BASED ISOGEOMETRIC ANALYSIS FOR BALLISTIC EVALUATION OF TITANIUM PLATES 8047
Mattia Montanari, Petros Siegkas, Antonio Pellegrino, Nik Petrinic

HYBRID RIEMANN SOLVERS FOR LARGE SYSTEMS OF CONSERVATION LAWS 8057
Birte Schmidtmann, Mariia Astrakhantceva, Manuel Torrilhon

CS 940: EXTENDED DISCRETIZATION METHODS

THE USE OF ENRICHED BASE FUNCTIONS IN THE THREE-DIMENSIONAL SCALED BOUNDARY FINITE ELEMENT METHOD 8066
Sascha Hell, Wilfried Becker

METHODOLOGY OF COMBINED APPLICATION OF DIRECTIONAL DERIVATIVES AND THE EXTENDED FINITE ELEMENT METHOD (X-FEM) FOR SOLVING VIBRATION EIGENVALUE PROBLEMS 8085
Daria Serbichenko, Pascal Cosson, Patrick Rozycki

XFEM ANALYSIS – INCLUDING BUCKLING – OF COMPOSITE SHELLS CONTAINING DELAMINATION 8103
Wilhelm J.H. Rust, Saleh Yazdani, Peter Wriggers

CS 960: MESHLESS METHODS

THREE-DIMENSIONAL CRACK PROPAGATION ANALYSIS USING MESHLESS POINT COLLOCATION METHOD 8115
Eiji Tanaka

A CELL-BASED SMOOTHED POINT INTERPOLATION METHOD FOR AXISYMMETRIC PROBLEMS 8133
Arash Tootoonchi, Arman Khoshghalb

EFFICIENT SHAPE FUNCTION COSTRUCTION OF EFG MESHLESS METHOD 8142
P. Metsis, M. Papadrakakis

CS 980: NUMERICAL AND SYMBOLIC COMPUTATION
MODERNIZING SOFTWARE IN SCIENCE AND ENGINEERING: FROM C/C++ APPLICATIONS TO MOBILE PLATFORMS 8162
Liliana Favre, Claudia Pereira, Liliana Martinez

NEW EDUCATIONAL MATLAB TOOL TO EXPLAIN TWO DIMENSIONAL FINITE ELEMENT METHOD TO GRADUATE STUDENTS 8177
Rafet Sisman, Abdurrahman Sahin

CONTRIBUTION OF NUMERICAL THERMAL SIMULATION OF MASONRY TO THE GLOBAL OBJECTIVE OF BUILDINGS ENERGY EFFICIENCY 8187
Hipólito Sousa, Rui Sousa, Luís Silva, Gonçalo Sousa

CS 990: PARTICLE-BASED METHODS
TOWARDS MULTIPHYSICS SIMULATION OF DEEP PENETRATION LASER WELDING USING SMOOTHED PARTICLE HYDRODYNAMICS 8196
Haoyue Hu, Peter Eberhard, Florian Fetzer, Peter Berger

MULTIPLE TIMESCALE MODELLING OF PARTICLE SUSPENSIONS IN METAL MELTS SUBJECTED TO EXTERNAL FORCES 8207
Anton Manoylov, Georgi Djambazov, Valdis Bojarevics, Koulis Pericleous

STATISTICAL DISTRIBUTION FUNCTION OF CHARGED PARTICLES IN MAGNETIC FIELD 8224
B.Chetverushkin, Nicola D'Ascenzo, Valeri Saveliev

A STABILISED TOTAL LAGRANGIAN CORRECTED SMOOTH PARTICLE HYDRODYNAMICS TECHNIQUE IN LARGE STRAIN EXPLICIT FAST SOLID DYNAMICS 8231
Giorgio Greto, Sivakumar Kulasegaram, Chun H. Lee, Antonio J. Gil, Javier Bonet

BLOOD FLOW SIMULATION USING SMOOTHED PARTICLE HYDRODYNAMICS 8241
Mohammed Al-Saad, Sivakumar Kulasegaram, Stephane P.A. Bordas

CS 1010: COMPUTATIONAL INVERSE PROBLEMS AND OPTIMIZATION
ON THE VALIDITY OF TIDAL TURBINE ARRAY CONFIGURATIONS OBTAINED FROM STEADY-STATE ADJOINT OPTIMISATION 8247
Christian T. Jacobs, Matthew D. Piggott, Stephan C. Kramer, Simon W. Funke

DEVELOPMENT OF ACCURATE PNEUMATIC TYRE FINITE ELEMENT MODELS BASED ON AN OPTIMISATION PROCEDURE 8262
Chrysostomos-Alexandros Bekakos, George Papazafeiropoulos, Dan J. O'Boy, Jan Prins

METEOROLOGICAL DATA ASSIMILATION USING AN ADJOINT PROGRAM GENERATED BY AUTOMATIC DIFFERENTIATION 8275
Yasuyoshi Horibata

EVOLUTIONARY TOPOLOGY OPIMIZATION USING PARAMETERIZED B-SPLINE SURFACE 8283
Igor Pehnec, Damir Vucina, Frane Vlak

GEARBOX DESIGN VIA MIXED-INTEGER PROGRAMMING 8294
Bastian Dörig, Thorsten Ederer, Peter F. Pelz, Marc E. Pfetsch, Jan Wolf

A FRAMEWORK FOR THE DESIGN BY OPTIMIZATION OF HYDROFOILS UNDER CAVITATING CONDITIONS 8305
Paolo Olivucci, Stefano Gaggero

A COMPUTATIONAL STRATEGY FOR TRAJECTORY OPTIMIZATION OF UNDERACTUATED MULTIBODY SYSTEMS WITH CONTACTS 8324
Silvia Manara, Alessio Artoni, Marco Gabiccini

ANALYSIS OF NON-LINEAR ENVIRONMENTAL LOAD COMBINATIONS BY EXTENDED CONTOUR-LINE ALGORITHMS 8337
Bernt Leira

ADJOINT-BASED ROBUST OPTIMIZATION USING POLYNOMIAL CHAOS EXPANSIONS 8351
Joao Miranda, Dinesh Kumar, Chris Lacor

APPLICATION OF THE DECONVOLUTION METHODS FOR PROCESSING OF MEASUREMENT SIGNALS IN THE FAST PROCESSES 8365
Marat Goldfeld, Valery Pickalov

INVERSION METHODS BASED ON KALMAN FILTERING FOR IDENTIFICATION IN TUNNELING PROBLEMS 8378
Tamara Nestorović, Luan T. Nguyen

CS 1020: EVOLUTIONARY AND DETERMINISTIC METHODS FOR DESIGN, OPTIMIZATION AND CONTROL
EVALUATION OF DIFFERENT APPROACHES FOR THE OPTIMIZATION OF LAYOUT AND CONTROL OF BOOSTER STATIONS 8391
Philipp Pöttgen, Peter F. Pelz

MANIFESTATION OF ASCENDANCY OF EXTINCTION-REIGNITION ON SOUNDING HYBRID ROCKET USING DESIGN INFORMATICS 8403
Kazuhisa Chiba, Masahiro Kanazaki, Toru Shimada

COMPUTING AN OPERATING STRATEGY FOR AN ACTIVE BODY CONTROL WITH DYNAMIC PROGRAMMING IN THE AUTOMOTIVE AREA 8419
Marlene Utz, Phillipp Hedrich, Peter F. Pelz

SEISMIC RELIABILITY-BASED DESIGN OF STRUCTUTRES ISOLATED BY FPS 8431
Paolo Castaldo, Guglielmo Amendola, Bruno Palazzo

SEISMIC RETROFIT OF EXISTING BUILDINGS THROUGH THE DISSIPATIVE COLUMNS 8445
Paolo Castaldo, Bruno Palazzo, Francesco Perri, Ivana Marino, Marco Maria Faraco

AN EXAMPLE OF ENERGY DISSIPATION OPTIMIZATION FOR STEEL MRFS WITH PIN-JOINTED COLUMN BASES 8460
Elide Nastri, Annabella Paciello

ECCENTRICALLY BRACED FRAMES DESIGNED FOR THE ENERGY DISSIPATION OPTIMIZATION 8476
Elide Nastri

AN EXAMPLE OF COLUMN DESIGN OPTIMIZATION FOR FAILURE MODE CONTROL OF REINFORCED CONCRETE FRAMES 8492
Roberta Muscati

A COMPARISON OF MILP AND MINLP SOLVER PERFOMANCE ON THE EXAMPLE OF A DRINKING WATER SUPPLY SYSTEM DESIGN PROBLEM 8509
Lea Rausch, Philipp Leise, Thorsten Ederer, Lena C. Altherr, Peter F. Pelz

RELIABILITY TEST DEMONSTRATION METHOD FOR EXPONENTIAL LIFE SYSTEM WITH RELIABILITY GROWTH UNDER THE CONDITION OF IN-TIME CORRECTIVE STRATEGY 8528
Yunyan Xing, Ping Jiang, Feng Yao, Zhiwei Yang, Michael Emmerich, Thomas Bäck

CS 1100: REDUCTION ORDER METHODS
EVALUATION OF LINE-FITTING METHOD OF MODEL ORDER REDUCTION 8539
Valery Makhavikou, Roland Kasper, Dmitry Vlasenko

AN EFFICIENT ORDER REDUCTION STRATEGY IN EARTHQUAKE NONLINEAR RESPONSE ANALYSIS OF STRUCTURES 8552
Franz Bamer, Abbas Kazemi Amiri, Christian Bucher

ACTIVE SUBSPACES FOR THE PRELIMINARY FLUID DYNAMIC DESIGN OF UNCONVENTIONAL TURBOMACHINERY 8572
Juan S. Bahamonde, Matteo Pini, Piero Colonna

EXTENSION AND APPLICATION OF A NONLINEAR REDUCED ORDER MODEL TO GUST LOAD PREDICTION IN TIME DOMAIN 8587
Christoph Strobach, Klemens Lindhorst, Matthias Haupt, Peter Horst

CS 1200: STRUCTURAL DYNAMICS
RESEARCH ON HYSTERETIC BEHAVIORS OF A SAPARATED SHOCK ABSORBER APPLIED IN RAILWAY BRIDGE 8597
Aili Li, Ri Gao, Mingde Sun, Jilei Zhang

FINITE ELEMENT ANALYSIS OF TIMBER BEAMS WITH FLAWS 8606
Janka Kovacikova, Mats Ekevad, Olga Ivankova, Sven Berg

NONLINEAR FORCED RESPONSE OF A STATOR VANE WITH MULTIPLE FRICTION CONTACTS USING A COUPLED STATIC/DYNAMIC APPROACH 8612
Marco Lassalle, Christian Maria Firrone

COMPUTATIONAL METHOD OF DETERMINATION OF INTERNAL EFFORTS IN LINKS OF MECHANISMS AND ROBOT MANIPULATORS WITH STATICALLY DEFINABLE STRUCTURES CONSIDERING THE DISTRIBUTED DYNAMICALLY LOADINGS 8627
Zhumadil Baigunchekov, Muratulla Utenov, Nurzhan Utenov, Saltanat Zhilkibayeva

FE MODELLING OF SFRC BEAMS UNDER IMPACT LOADS 8640
Pegah Behinaein, Ali Abbas, Demetris Cotsovos

AN APROXIMATE ANALYTICAL SOLUTION FOR NONLINEAR FGM SHELL STRUCTURE WITH VARIABLE IN TIME PARAMETERS 8654
Victor Gristchak, Yuliia Fatieieva

UNCERTAINTY OF MODELS IN INTELLIGENT SYSTEMS UNDER STOCHASTIC LOADING 8665
Amalia Moutsopoulou, Georgios Stavroulakis, Tasos Pouliezos

COMPUTATIONAL MODELLING OF COLD-FORMED STEEL SCREWED CONNECTIONS AT AMBIENT AND ELEVATED TEMPERATURES 8678
Luis Mesquita, Rui Dias, Armandino Parente, Paulo Piloto

MONITORING AND ANALYSIS OF STRESS FIELD FOR ORTHOTROPIC STEEL DECK OF DASHENGGUAN YANGZTE RIVER BRIDGE 8691
Ying Wang, Y.S. Song

CS 1201: COMPUTATIONAL SOIL MECHANICS
A MORE COMPREHENSIVE MODELING OF CONTACT FORCE DURING SHEAR TESTING USING DEM 8708
Varvara Roubtsova, Mohamed Chekired

THREE-DIMENSIONAL BE-FE MODEL OF BUCKET FOUNDATIONS IN POROELASTIC SOILS 8725
Jacob D. R. Bordón, Juan J. Aznárez, Orlando F. Maeso

CS 1202: STRUCTURAL ANALYSIS AND MULTI BODY DYNAMICS
ON THE NUMERICAL INFLUENCES OF INERTIA REPRESENTATION FOR RIGID BODY DYNAMICS 8739
Xiaoming Xu, Wanxie Zhong

APPROACHES TO THE CREATION OF MULTIBODY MODELS OF THE VVER 1000 NUCLEAR REACTOR CONTROL ASSEMBLY 8756
Pavel Polach, Michal Hajžman

CS 1300: UNCERTAINTY QUANTIFICATION AND ERROR ESTIMATION
THE UPLIFT CAPACITY OF HORIZONTAL PLATE ANCHORS IN SPATIALLY VARIABLE CLAY USING SPARSE POLYNOMIAL CHAOS EXPANSIONS 8769
Tom Charlton, Mohamed Rouainia

FATIGUE RELIABILITY OF AGEING RAILWAY BRIDGES: FEASIBILITY OF PROBABILISTIC APPROACH 8778
Nirosha D. Adasooriya

AN EFFICIENT AERODYNAMIC SHAPE OPTIMIZATION FRAMEWORK FOR ROBUST DESIGN OF AIRFOILS USING SURROGATE MODELS 8787
Daigo Maruyama, Dishi Liu, Stefan Görtz

QUANTIFICATION OF SPATIAL VARIABILITY FOR TRANSVERSE ELASTIC MODULUS OF SPRUCE WOOD …….. 8801
Alireza Farajzadeh Moshtaghin, Steffen Franke, Thomas Keller, Anastasios Vassilopoulos

COMPARISON BETWEEN A POLYNOMIAL CHAOS SURROGATE MODEL AND MARKOV CHAIN MONTE CARLO FOR INVERSE UNCERTAINTY QUANTIFICATION BASED ON AN ELECTRIC DRIVE TEST BENCH …….. 8809
Philipp Glaser, Michael Schick, Kosmas Petridis, Vincent Heuveline

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MINIMAL MASS DESIGN OF STRENGTHENING TECHNIQUES FOR PLANAR AND CURVED MASONRY STRUCTURES

G. Carpentieri[1], F. Fabbrocino[2], M. De Piano[3], V. P. Berardi[3], L. Feo[3] and F. Fraternali[3]

[1]University of Naples Federico II, Department of Engineering
80132 Naples, Italy
e-mail: gcarpentieri@unisa.it

[2] Pegaso University, Department of Engineering
80132 Naples, Italy
e-mail: francesco.fabbrocino@unipegaso.it

[3] University of Salerno, Department of Civil Engineering
84084, Fisciano, SA, Italy
e-mail: depianomariella@yahoo.it, l.feo@unisa.it, f.fraternali@unisa.it

Keywords: Masonry structures, Strengthening, Tensegrity, Topology Optimization, Composite materials

Abstract. *We present a discrete element model of a masonry structure strengthened through the application of reinforcing elements designed to work in tension. We describe the reinforced masonry structure as a tensegrity network of masonry rods, mainly working in compression, and tension elements corresponding to fiber-reinforced composite reinforcements, which are assumed to behave as elastic-perfectly-plastic members. We optimize a background structure connecting each node of the discrete model of the structure with all the neighbors lying inside a sphere of prescribed radius, in order to determine a minimal mass resisting structure under the given loading conditions and prescribed yielding constraints. Fiber-reinforced composite reinforcements can be naturally replaced by any other reinforcements that are strong in tension (e.g., timber or steel beams/ties). Some numerical examples illustrate the potential of the proposed strategy in designing tensile reinforcements of a three-dimensional structure composed of a masonry vault and supporting walls.*

1 Introduction

It is known that old masonry often features nearly zero tensile strength [1, 2, 3]. Nowadays, it is a common practice to reinforce such structures by applying tensile reinforcements made of traditional materials, such as steel, or innovative high-strength materials [4, 5, 6]. Strips and/or meshes of materials like Fiber Reinforced Polymers (FRP) or Fabric Reinforced Cementitious Matrix (FRCM) composites are often bonded to masonry structures to improve their mechanical properties [7, 8, 9]. However, it is worth remarking that such strengthening techniques, when improperly used, may lead to an excessive over-strength of the reinforced structure, and reduced 'cracking-adaptation' capacity [10].

The modern *Discrete Element Modeling* (DEM) of masonry structures includes computer-assisted, funicular-network procedures [11], *Lumped Stress Models* [12, 13, 14, 15], and *Thrust Network Approaches* (TNA) [16, 17, 18]. A recent study [19] has presented a tensegrity approach to the 'minimal-mass' FRP-/FRCM reinforcement of masonry vaults and domes. Such a procedure employs tensegrity concepts to find an optimal resisting mechanism of the reinforced structure, under given loading conditions, in line with the 'Italian Guide for the Design and Construction of Externally Bonded FRP Systems for Strengthening Existing Structures' [21]. The latter indeed allows the designer to describe the response of the reinforced structure through simplified schemes, on assuming that tensile stresses are directly taken by the FRP reinforcements, and the stress level may be determined by adopting a distribution of stresses that satisfies the equilibrium conditions but not necessarily the strain compatibility (cf. Sect. 5.2.1 of [21]). The approach proposed in [19] describes the reinforced structure as a tensegrity network of masonry rods, working in compression, and tension elements corresponding to the FRP-/FRCM- reinforcements, which are assumed to behave as elastic-perfectly-plastic members. It optimizes a *background structure* connecting each node of a discrete model of the structure with all the neighbors lying inside a sphere of prescribed radius, in order to determine a minimal mass resisting structure under the given loading conditions and prescribed yielding constraints [22]. The FRP/FRCM reinforcements can be naturally replaced by any other reinforcements that are strong in tension (e.g., timber or steel beams/ties).

The present study generalizes the approach presented in [19, 20] to the case of 2D and 3D discrete models of masonry structures with arbitrary shape. Such an extension allows us to explore the potential of the tensegrity modeling of reinforced masonry structures in the design of non-invasive reinforcement patterns of systems formed by masonry walls, vaults and domes. We formulate a design procedure that seeks for an optimal and lightweight pattern of reinforcing elements giving rise to a minimal mass resisting mechanisms of the examined structure, under given loads and yielding constraints. Due to the safe theorem of the limit analysis of elastic-plastic bodies [23], the existence of such a mechanism ensures that the reinforced structure is safe under the examined loading conditions, on assuming elastic-perfectly-plastic response of all members. The input variables of the proposed procedure consist of a 3D point cloud defining the geometry of the structure to be reinforced, obtainable, e.g., through in-situ laser-scanning, together with the material densities and yielding strengths of masonry and reinforcing elements.

The remainder of the paper is structured as follows. Section 2 illustrates the adopted minimal mass modeling of a reinforced masonry structure under given yielding constraints and loading conditions. Next, Sect. 3 presents case studies dealing with the FRP-/FRCM-reinforcement of an independent cloister vault (Sect. 3.1) three-dimensional structural system formed by a cloister vault and supporting walls (Sect. 3.2). We conclude with final remarks and directions of future research in Sect. 4.

2 Minimal-mass reinforcement of a masonry structure

Let us generalize the optimization strategy presented in Sections 2 and 3 of Ref. [19] to the general case of an arbitrary masonry structure, whose geometry is described by three-dimensional set of n_n nodes with position vectors $\mathbf{n}_k$ ($k = 1, ..., n_n$). Such nodes may be condensed over one or multiple structural surfaces, e.g, the intrados and the extrados surfaces of a planar wall or a vaulted structure.

We introduce a background structure (refer, e.g., to the example of Fig. 1) by connecting each node $\mathbf{n}_k$ with all the nodes $\mathbf{n}_j$ such that it results $|\mathbf{n}_k - \mathbf{n}_j| \leq r_k$ (*interacting neighbors*), through two elements working in parallel: a compression element (or *bar*) $\mathbf{b}_i = \mathbf{n}_k - \mathbf{n}_j$; and a tension element (or *string*) $\mathbf{s}_i = \mathbf{n}_k - \mathbf{n}_j$. Assuming that such a background structure is subject to a number m of static loading conditions, we write its equilibrium equations as follows

$$\mathbf{A}\mathbf{x}^{(j)} = \mathbf{w}^{(j)} \tag{1}$$

where j is the loading condition index ($j = 1, ..., m$); $\mathbf{A}$ is the static (or equilibrium) matrix; $\mathbf{w}^{(j)}$ is the external load vector; and $\mathbf{x}^{(j)}$ is the vector collecting all the force densities in bars and strings (refer to [22] for the analytic expression of $\mathbf{A}$).

We now assume that bars and strings behave as as elastic-perfectly-plastic members, with yield strength σ_{b_i} in the generic bar (compressive yield strength), and yield strength σ_{s_i} in the generic string (tensile yield strength). We let A_{b_i} denote the cross-section area of $\mathbf{b}_i$, and let A_{s_i} denote the cross-section area of $\mathbf{s}_i$. The masses of such members are respectively given by $m_{b_i} = \rho_{b_i} A_{b_i} b_i$, and $m_{s_i} = \rho_{s_i} A_{s_i} s_i$, where ρ_{b_i} and ρ_{s_i} respectively denote the mass densities of $\mathbf{b}_i$ and $\mathbf{s}_i$; b_i denotes the length of $\mathbf{b}_i$ and s_i denotes the length of $\mathbf{s}_i$. Moreover, in correspondence with the j-th loading condition, we let $\lambda_{b_i}^{(j)}$ denote the force density carried by $\mathbf{b}_i$, and let $\gamma_{s_i}^{(j)}$ denote the force density carried by $\mathbf{s}_i$, such that $\lambda_{b_i}^{(j)} > 0$ when $\mathbf{b}_i$ is compressed, and $\gamma_{s_i}^{(j)} > 0$ when $\mathbf{s}_i$ is stretched. Yielding constraints impose that it results

$$\lambda_i^{(j)}\, b_i \leq \sigma_{b_i} A_{b_i}, \quad \gamma_i^{(j)}\, s_i \leq \sigma_{s_i} A_{s_i} \tag{2}$$

in correspondence with all the bars and strings, and all the loading conditions.

We seek for an optimized resisting mechanism of the examined structure through the following linear program [22, 19]

$$\begin{aligned} &\underset{\mathbf{x}^{(j)},\mathbf{y}}{\text{minimize}} \quad m = \mathbf{d}^T \mathbf{y} \\ &\text{subject to} \quad \begin{cases} \mathbf{A}\mathbf{x}^{(j)} = \mathbf{w}^{(j)} \\ \mathbf{C}\mathbf{x}^{(j)} \leq \mathbf{D}\mathbf{y} \\ \mathbf{x}^{(j)} \geq \mathbf{0}, \mathbf{y} \geq \mathbf{0} \end{cases}, \end{aligned} \tag{3}$$

where

$$\mathbf{y} = [A_{b_1} \ \cdots \ A_{b_{n_b}} \mid A_{s_1} \ \cdots \ A_{s_{n_s}}]^T \tag{4}$$

$$\mathbf{d}^T = [\rho_{b_i} b_i \ \cdots \rho_{b_{n_b}} b_{n_b} \mid \rho_{s_i} s_i \ \cdots \rho_{s_{n_s}} s_{n_s}] \tag{5}$$

$$\mathbf{C} = \begin{bmatrix} \mathrm{diag}(b_1, \cdots, b_{n_b}) & \underline{0} \\ \mathbf{0} & \mathrm{diag}(s_1, \cdots, s_{n_s}) \end{bmatrix} \tag{6}$$

$$\mathbf{D} = \begin{bmatrix} \mathrm{diag}(\sigma_{b_1}, \cdots, \sigma_{b_{n_b}}) & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\sigma_{s_1}, \cdots, \sigma_{s_{n_s}}) \end{bmatrix} \tag{7}$$

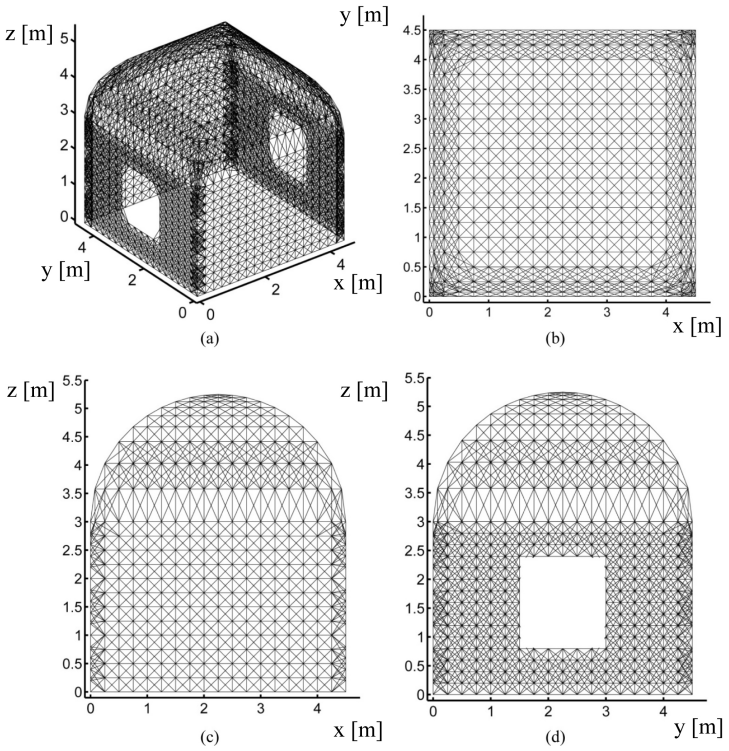


Figure 1: Background structure associated with a 3D point cloud describing the geometry of a cloister vault supported by perimeter walls (dimensions in meters): (a) 3d view; (b) top view; (c)-(d) side views. The connection distance r_k is equal to 0.75 m for nodes at the connection between walls and cloister, and 0.5 m for the reimaning nodes.

The solution to problem (3) provides minimal-mass configuration of the background structure; chooses whether a bar or a string connects each couple of interacting nodes; and returns bars and strings with zero cross-section areas in correspondence with the interacting nodes that do not need to be connected in the minimal mass configuration, under the given equilibrium (1) and yielding (2) constraints.

3 Numerical Results

The present section provides a collection of numerical applications of the minimal mass optimization procedure described in Sect. 2, which are aimed at designing optimal reinforcements of 3D structural complex formed by a cloister vault and supporting walls (Sects. 3.1, 3.2).

On assuming that each analyzed structure is formed by masonry struts with uniform compressive yield strength σ_b, and tension reinforcements with uniform yield strength σ_s, we employ the in-house software 'tensopt' [25] to numerically solve problem (3). We convene to mark the reinforcing elements by red lines and the masonry struts by solid black lines.

3.1 Cloister vault

We first examine a cloister vault made of 'Neapolitan' tufe brick masonry, which is largely diffused in the area of Naples, with 15.0 kN/m^3 self-weight, and 13 MPa compressive strength σ_b. We assume a tensile strength σ_b equal to 376.13 MPa, which corresponds to an average value of the bond strengths of the FRP and FRCM reinforcements of masonry structures analyzed in [7, 8], respectively (we employed formula (5.6) of [21] to estimate such a strength).

Fig. 2 shows the minimal mass FRP/FRCM reinforcements that we obtained for the present example ($t_f = 0.17$ mm). The geometry of the examined vault are illustrated in above figure, together with the corresponding background structure, which features 441 nodes and 4508 connections (see Figs. 2a-c). The optimal reinforcement of such a vault under vertical loading is mainly formed by parallel FRP/FRCM strips with 0.17 mm thickness and 82 mm maximum width near the crown (Figs. 2d-f). The above reinforcements are integrated with diagonal FRP/FRCM strips with about 140 mm maximum width near the intersections of the four vault segments, under combined vertical and seismic loading (Figs. 2g-l). The analyzed seismic loading consists of horizontal forces with magnitude equal to 0.35 of the magnitude of vertical forces in all nodes, which mimic the effects of a seismic excitation of the examined structure, through a conventional, static approach [26]. The compressed network include couples of diagonal arches near the corners, parallel-line arches, and diagonal struts over the vault segments (Figs. 2d-l).

3.2 3D system formed by planar and curved masonry structures

This second example is concerned with a 3D system composed of 4 orthogonal walls featuring 4.5 m horizontal length, 3.0 m height and 50 cm thickness, which support a cloister vault with 2.25 m central rise and 25 cm thickness (cf. Fig. 1). The two walls parallels to the y axis of a Cartesian frame with the z-axis placed along the vertical show 1.5 m $\times$ 1.6 m central openings. The background structure illustrated in Fig. 1 features 1385 nodes and 15378 potential connections. It is worth noting that in the present case we model both the perimeter walls and the vault as 2D membranes lying in the 3D Cartesian space.

We here assume σ_b = 1.21 MPa, σ_s = 112.5 MPa, $t_f = 0.17$ mm, and masonry selfweight equal to 15.0 kN/m^3 (tufe masonry). The optimal design reinforcement for the current example is illustrated in Figs. 3 and 4, under the action of pure vertical loading (structure selfweight), and the combined action of selfweight and seismic loading in the $+y$ direction, respectively. The optimal reinforcements under the action of pure vertical loading are mainly placed along the perimeter at the base of the cloister ($z \approx 3$ m); along horizontal lines over the two piers of the y-walls with openings; and along diagonal lines at the intersections of the vault segments (Fig. 3).

For what concerns the seismic loading condition (cf. Fig. 4), we observe that the optimal reinforcement strategy combines that corresponding to vertical loading with additional diagonal reinforcements over the two piers of the walls featuring central openings, and reinforcements aligned-with- or orthogonal-to-the junctions between the vaults segments, when moving towards the crown of the vault (cf. Figs. 4 and 3). Due to the adopted membrane modeling of

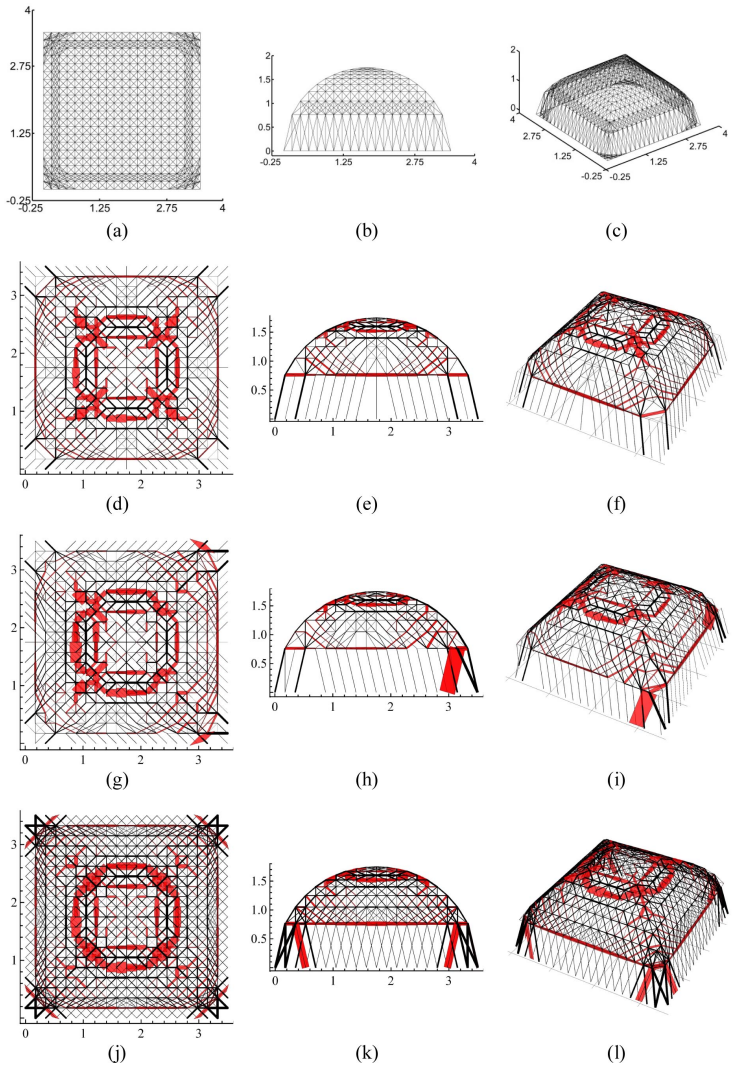


Figure 2: Top, side and 3D views of the optimal reinforcement patterns of a cloister vault with FRP/FRCM strips of thickness 0.17 mm (marked in red), under different loading conditions. The widths of the FRP/FRCM reinforcements are magnified by a factor 2 for visual clarity. (a)-(c): Background structure. (d)-(f): Vertical loading. (g)-(i): Seismic loading in the $+x$-direction. (j)-(l): Combined vertical loading and seismic loading in two perpendicular directions.

all the elements forming the current structure, only the two walls parallel to the direction of the seismic forces ($+y$-axis) are actually interested by the effects of such forces, among all the vault supports.

Comparing the results shown in Figs. 2 and 3-4, we realize that the presence of perimeter walls in the current model leads us to design different topologies of the reinforcing elements, as compared to those predicted by the modeling of the vault as an independent structure constrained by fixed spherical hinges at the base. This is mainly due to the fact that the perimeter walls do not carry forces orthogonal to their planes in the current model, and therefore cannot be exactly replaced by spherical hinges. It is worth noting that the current model predicts major reinforcements over the perimeter walls, and lighter reinforcements over the surface of the vault, as compared to that employed in the previous section.

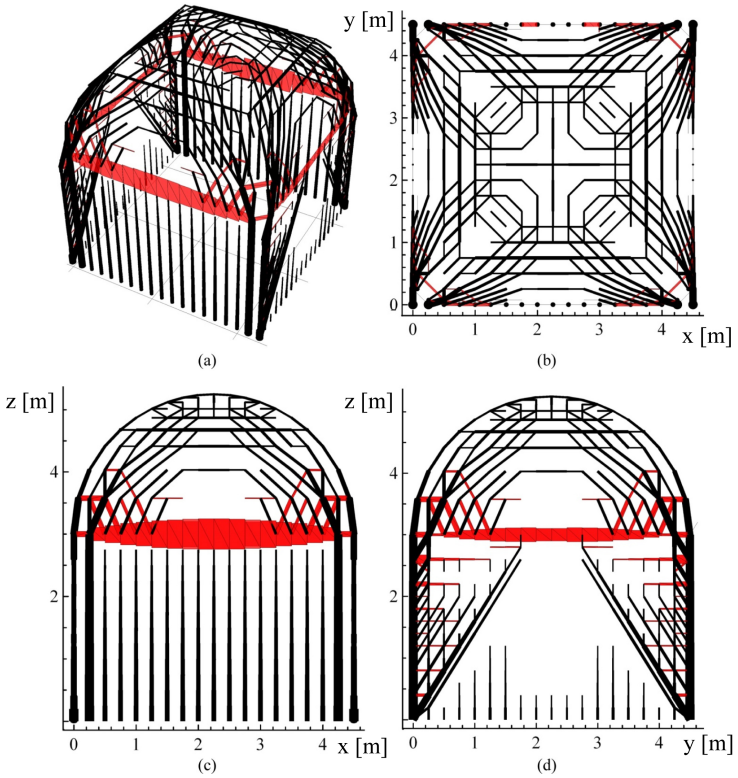


Figure 3: Optimal reinforcement patterns of the cloister vault supported by walls under vertical loading (reinforcements marked in red): (a) 3d view; (b) top view; (c) xz view; (d) yz view (α = 3.343, $V_f = 0.647 \times 10^{-3}$ m^3, $\mu_f = 0.567 \times 10^{-3}$). The widths of the reinforcements are magnified by a factor 2 for visual clarity.

4 Concluding remarks

We have presented an extension of the tensegrity approach formulated in Ref. [19] for the minimal mass reinforcement of masonry vaults and domes that do not react in tension. Such an extension allows us to analyze masonry structures of general shape and dimensions, including 2D walls, 3D walls, and structural complexes formed by an arbitrary combination of walls, vaults and domes.

The reinforcements analyzed in the present study consist of linear elements, such as, e.g., FRP-/FRCM-reinforcements, steel ties, timber beams, and any other reinforcements that are strong in tension. The adopted optimization approach allows us to design non-invasive reinforcement patterns, which can be able to preserve a sufficient crack-adaption capacity of the structure [10, 19, 27], under the respect of the equilibrium equations and material yield limits.

The given numerical results have highlighted that the proposed reinforcement design approach is able to handle both in-plane and out-of-plane loadings, walls with openings, and arbitrary support conditions of vaulted structures. It is worth remarking that the proposed strengthening approach matches the safe theorem of the limit analysis of elastic-plastic bodies [23, 10], and is in line with the recommendations of modern standards for the the design and construction of strengthening techniques for existing structures [21].

Future directions of the present study will be aimed at analyzing the minimal mass reinforcement of a variety of case-studies dealing with masonry structures of arbitrary geometry and complexity. Additional future research lines include the generalization of the proposed design approach to tensegrity materials and structures [28]-[34], and a wide campaign of exper-

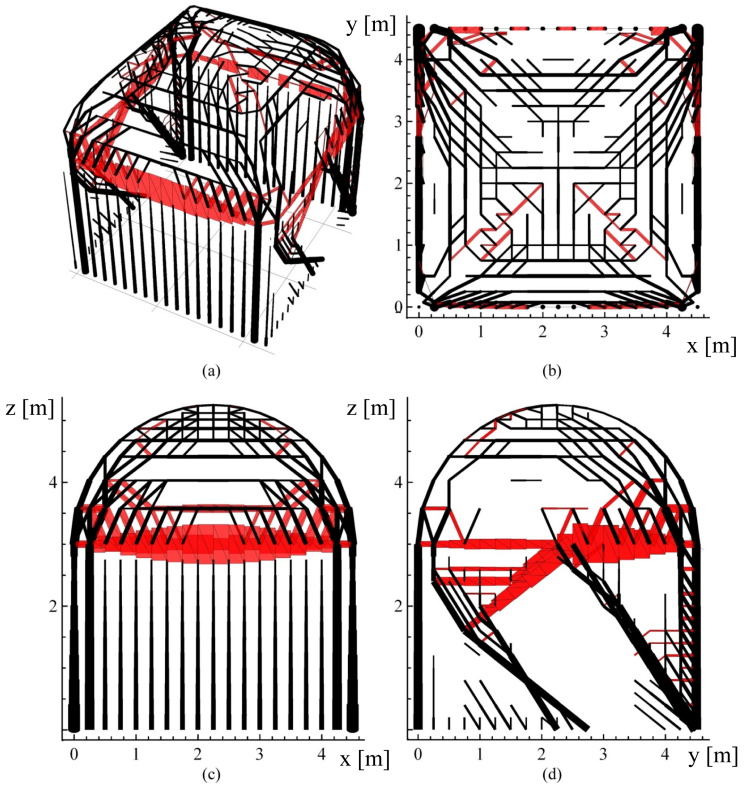


Figure 4: Optimal reinforcement patterns of the cloister vault supported by walls under combined vertical and seismic loadings in the +y-direction (reinforcements marked in red): (a) 3d view; (b) top view; (c) xz view; (d) yz view ($\alpha = 3.343$, $V_f = 1.010 \times 10^{-3}\ m^3$, $\mu_f = 0.807 \times 10^{-3}$). The widths of the reinforcements are magnified by a factor 2 for visual clarity.

imental validations of the design procedure presented in Sect. 2, through laboratory testing of real-scale and reduced-scale models under static and dynamic loading [35].

REFERENCES

[1] R. Hassanli, M.A. ElGawady, J.E. Mills, *Strength and Seismic Performance Factors of Posttensioned Masonry Walls*, J. Struct. Eng., 141(11): 04015038, 2015.

[2] B. Ghiassi, M. Soltani, A.A. Tasnimi, *In-plane Lateral Response of Brick Masonry Walls Retrofitted with Reinforced Concrete Layer* The 14th World Conference on Earthquake Engineering, October 12-17, 2008, Beijing, China, 2008.

[3] A. Fortunato, E. De Chiara, F. Fraternali, M. Angelillo, *Advanced models for the limit analysis of masonry structures*. *CompDyn 2015*, 5th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering, 3716-3725, 2015.

[4] S. Petry, K. Beyer, *Forcedisplacement response of in-plane-loaded URM walls with a dominating flexural mode* Earthquake Engng Struct. Dyn. DOI: 10.1002/eqe, 2015.

[5] G. Magenes, G.M. Calvi, *In-plane seismic response of brick masonry walls* Earthq Eng Struct D, 1997.

[6] M. ElSayed, W. El-Dakhakhni, M. Tait, *Response Evaluation of Reinforced Concrete Block Structural Walls Subjected to Blast Loading* J. Struct. Eng., 141(11): 04015043, 2015.

[7] C. Mazzotti, B. Ferracuti, A. Bellini, *Experimental bond tests on masonry panels strengthened by FRP*. Compos Part B-Eng, 80:223–237, 2015.

[8] F.G. Carozzi, C. Poggi, *Mechanical properties and debonding strength of Fabric Reinforced Cementitious Matrix (FRCM) systems for masonry strengthening*. Compos Part B-Eng, 70:215–230, 2015.

[9] F. Fabbrocino, I. Farina I, V.P. Berardi, A.J.M. Ferreira, F. Fraternali, *On the thrust surface of unreinforced and FRP-/FRCM-reinforced masonry domes*. Compos Part B-Eng, 83, 297-305, 2015.

[10] J. Heyman, *The Stone Skeleton*. Cambridge University Press, Cambridge, 1995.

[11] D. O'Dwyer, *Funicular analysis of masonry vaults*. Comput Struct, 73:187–197, 1999.

[12] F. Fraternali, M. Angelillo, A. Fortunato, *A lumped stress method for plane elastic problems and the discrete-continuum approximation*. Int J Solids Struct, 39:6211–6240, 2002.

[13] F. Fraternali, *A thrust network approach to the equilibrium problem of unreinforced masonry vaults via polyhedral stress functions*. Mech Res Commun, 37:198–204, 2010.

[14] F. Fraternali, *A mixed lumped stress–displacement approach to the elastic problem of masonry walls*. Mech Res Commun, 2011;38:176–180,

[15] F. Fraternali, G.Carpentieri, *On the Correspondence Between 2D Force Networks and Polyhedral Stress Functions*. International Journal of Space Structures, 29(3):145–159, 2014.

[16] P. Block, J. Ochsendorf, *Thrust Network Analysis: A new methodology for three-dimensional equilibrium*. IASS J, 48(3):167–173, 2007.

[17] P. Block, *Thrust Network Analysis: Exploring Three-dimensional equilibrium*. PhD dissertation, Massachusetts Institute of Technology, Cambridge, USA, 2009.

[18] F. De Goes, P. Alliez, H. Owhadi, M. Desbrun, *On the equilibrium of simplicial masonry structures*. ACM Transactions on Graphics, 32(4):93, 2013.

[19] F. Fraternali, G. Carpentieri, M. Modano, F. Fabbrocino, R.E. Skelton, *A tensegrity approach to the optimal reinforcement of masonry domes and vaults through fiber-reinforced composite materials*. Compos. Struct., 134:247–254, 2015.

[20] F. Fraternali, G. Carpentieri, M. Modano, F. Fabbrocino, R.E. Skelton, *A tensegrity approach to the optimal reinforcement of masonry domes and vaults through fiber-reinforced composite materials*. Compos. Struct., 134, 247-254, 2015.

[21] Italian National Research Council (CNR), *Guide for the Design and Construction of Externally Bonded FRP Systems for Strengthening Existing Structures - Materials, RC and PC structures, masonry structures*. CNR-DT 200/2013 - R1, Rome, Italy, 2013.

[22] K. Nagase, R.E. Skelton, *Minimal mass tensegrity structures*, Journal of the International Association for Shell and Spatial Structures, 55(1):37–48, 2014.

[23] W.T. Koiter, *General Theorems for Elastic-Plastic Solids*, in: Progress in Solid Mechanics. Sneddon J.N., Hill R. (eds.), North-Holland, Amsterdam, 1:165-221, 1960.

[24] P. Foraboschi, A. Vanin, *New methods for bonding FRP strips onto masonry structures: experimental results and analytical evaluations*. Composites 4(1):1–23, 2013.

[25] F. Fraternali, G. Carpentieri, *Tensopt, Computer software*. Retrieved from http://www.fernandofraternaliresearch.com/downloads.asp?pg=9, 2014.

[26] European Committee for Standardization, *Eurocode 8: Design of structures for earthquake resistance. Part 1: General rules, seismic actions and rules for buildings*. EN 1998-1:2004, Brussels, Belgium, 2014.

[27] F. Fraternali, *Free discontinuity finite element models in two-dimensions for in-plane crack problems*. *Theoretical and Applied Fracture Mechanics*, 47 (3), pp. 274-282, 2007.

[28] G. Carpentieri, M. Modano, F. Fabbrocino, F. Fraternali, *Optimal design and dynamics of truss bridges*. *CompDyn 2015*, 25-27 May, 2015.

[29] R.E. Skelton, F. Fraternali, G. Carpentieri, A. Micheletti, *Minimum mass design of tensegrity bridges with parametric architecture and multiscale complexity*. *Mech Res Commun*, 58:124–132. ISSN: 0093-6413, DOI: 10.1016/j.mechrescom.2013.10.017, 2014.

[30] G. Carpentieri, R.E. Skelton, F. Fraternali, *Minimum mass and optimal complexity of planar tensegrity bridges*. *Int J Space Struct*, 30(3-4):221–244. ISSN: 0266-3511, DOI: http://dx.doi.org/10.1260/0266-3511.30.3-4.221, 2015.

[31] F. Fraternali, L. Senatore, C. Daraio, Solitary waves on tensegrity lattices. *Journal of the Mechanics and Physics of Solids*, **60**, 1137-1144, 2012.

[32] F. Fraternali, G. Carpentieri, A. Amendola, R.E. Skelton, V.F. Nesterenko, *Multiscale tunability of solitary wave dynamics in tensegrity metamaterials*. *Applied Physics Letters*, 105 (20), art. no. 201903, 2014.

[33] A. Amendola, C.J. Smith, R. Goodall, F. Auricchio, L. Feo, G. Benzoni, F. Fraternali, *Experimental response of additively manufactured metallic pentamode materials con-fined between stiffening plates*. Compos. Struct., 142:254–262, 2016.

[34] R.E. Skelton, R. Montuori, V. Pecoraro, *Globally stable minimal mass compressive tensegrity structure*. Compos. Struct., 141:346–354, 2016.

[35] M. Modano, F. Fabbrocino, A. Gesualdo, G. Matrone, I. Farina, F. Fraternali, *On the forced vibration test by vibrodyne*. *CompDyn 2015*, 25-27 May, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MODELING OF FIBER-REINFORCED PLASTICS TAKING INTO ACCOUNT THE MANUFACTURING PROCESS

Cherry Ann T. Reclusado[1], **Sumito Nagasawa**[2]

[1] Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI
Eckerstraße 4, 79104 Freiburg, Germany
cherryann.reclusado@emi.fraunhofer.de

[2] Fuji Heavy Industries Ltd.
1-1 Subaru-cho, Ota-shi, 373-8555 Gunma, Japan
nagasawa.sumito@subaru-fhi.co.jp

Keywords: Integrative simulation, process and structural simulation, crash simulation, fiber-reinforced plastics, orientation tensor, degree of anisotropy

Abstract: *In a joint project of the EMI and Fuji Heavy Industries Ltd. a method for the modeling of fiber-reinforced plastics was developed, taking into account the manufacturing process. Within this study, an approach for a detailed characterization and modeling of a fiber reinforced plastic is described and presented by the example of a PPGF30 material. The characterization of the orientation dependent material behavior includes tensile tests at different strain rates as well as tensile-unloading, compression and shear tests in 0°-, 45°- and 90°-direction. Also, quasi-static and dynamic three-point bending tests are performed and act as validation tests for the simulation model. For further validation of the method and to evaluate the simulation model's approximation to reality, dynamic three-point bending tests are performed on a component with a ribbed structure. Regarding the modeling of the mechanical behavior, the fiber orientation distribution is taken into account by means of injection molding simulations, both in the sample plate and the component. These simulations provide information about the orientation state at discrete material points in terms of an orientation tensor. By means of the eigenvalues and the respective eigenvectors of the orientation tensors, the degree of anisotropy and the principle fiber direction are defined. However, the degree of anisotropy is considered in a gradual manner by defining several material classes, each covering different ranges of the greatest eigenvalue. This is a very time consuming approach, because for each material class, one set of parameters has to be calibrated iteratively. The impact of considering the degree of anisotropy on the simulation results is therefore investigated as well. Another crucial aspect within this work is the development of a program to automatically translate and map the injection molding simulation results to appropriate variables in the structural simulation model. Furthermore, high-resolution CT-scans of the sample plate and the component are created in order to perform a fiber analysis of the real material and hence to verify the injection molding simulation results.*

1 INTRODUCTION

Injection molded glass fiber-reinforced plastics (GFRP) play an important role as modern lightweight construction materials, used for example for interior or exterior parts in the automotive industry. Those materials combine high-strength glass fibers with a shaping polymer matrix. A suitable distribution of the fiber orientation and the degree of anisotropy within a component can contribute positively to its resistance against deformations. Furthermore, the injection molding process allows for a functional integration and the production of ribbed structures, which lead to an additional stiffness of thin-walled components. Although continuous fiber-reinforced plastics with a selective orientation of the fiber strands can achieve even higher stiffnesses, the use of injection molded plastics is preferred due to lower manufacturing costs and cycle times. On the other hand, high development costs and times have to be considered. Especially the complex distribution of the fibers in injection molded components complicates the prediction of material properties and may require the cost-intensive manufacturing of prototypes. Often numerical simulations are used to predict the material behavior. Injection molding simulations (e.g. using Moldflow) allow for the calculation of fiber distributions, orientations and joint lines whereas structural simulations (e.g. using LS-DYNA) are used to calculate the mechanical behavior of a component. For such a structural simulation material parameters are needed which have to be determined in characterization tests. The consideration of all those results from injection molding and structural simulations as well as from the characterization tests is called »integrative simulation«. As in the conventional approach, the characterization tests are simulated and the material parameters as well as the structural model are validated by comparing stress-strain- and force-displacement-curves. For an integrative simulation, in addition, process-related parameters like fiber orientation and degree of anisotropy are included in the structural model.

Within this study the thorough investigation and modeling of a GFRP, subjected to dynamic loading, are presented by the example of a PPGF30 material, a long glass fiber reinforced thermoplastic (LGFRP). The characterization tests for obtaining the material parameters for modeling the orientation dependent material behavior include tensile tests at different strain rates, as well as tensile-unloading, compression and shear tests on specimens extracted parallel (0°-direction), perpendicular (90°-direction) and in 45°-direction to the main flow direction in the sample plate. When determining the material parameters, the fiber distribution is taken into account by means of an injection molding simulation. This simulation provides information about the orientation state at discrete material points in terms of an orientation tensor. By means of the eigenvalues and the respective eigenvectors of the orientation tensors the degree of anisotropy and the principle fiber direction are defined. Similar to the described approach in [1] several material classes are declared to cover different degrees of anisotropy and for each material class one MAT_108 material card [3], available in the finite element code LS-DYNA, is generated. Furthermore, the material cards are assigned by means of the keyword ELEMENT_SHELL_COMPOSITE [2]. The calibration of the MAT_108 material cards is based on the results from the dynamic tensile tests at nominal strain rate of 100/s. To model the failure behavior one MAT_ADD_EROSION card is assigned to each MAT_108 material card.

The validation of the parameters is done by comparing the stress-strain- and force-displacement-curves of the structural simulation and the experiments. In addition, to evaluate the predictive power of the prepared models and the introduced approach to incorporate process simulation results in the modeling of the mechanical behavior of the PPGF30 material, dynamic three-point bending tests and a structural simulation of these tests are performed using a component with a ribbed structure.

Another crucial aspect within this work is the development of a program to automatically map the injection molding simulation results to appropriate variables in the structural simulation model. This mapping tool not only translates information about the orientation state, but it also enables to overcome the mesh inconsistency between the tetrahedron mesh of the injection molding simulation and the shell mesh of the structural simulation. Furthermore, high-resolution CT-scans of selected regions of the sample plate and the component are created to validate the injection molding simulation results.

2 INTEGRATIVE SIMULATION

As already mentioned the use of injection molded GFRP enables the development of high-strength and very stiff components, but it comes with a complex material behavior. Different fiber orientation distributions, which make it difficult to predict the principle direction and the degree of anisotropy, can arise in various positions in the component. In addition, the outer and inner layers of a thin walled component can show a varying orientation distribution. While the fibers in the outer layers are likely oriented in the flow direction, the fibers in the inner layers are more deviated – in the most extreme case perpendicular to the flow direction. The varying fiber orientation distribution, which results due to the flow of the viscous fiber-reinforced plastic melt while injected in the cavity, represents a challenge for the structural simulation, because as one can conclude it is no longer sufficient to simply assume the principal fiber alignment and degree of anisotropy. To fully exploit the advantages of a GFRP the fiber orientation distribution needs to be considered both in the finished component and the sample plate, from which the specimens for the characterization tests are extracted. One way to detect the fiber orientation within a component is to make CT-scans, but the disadvantage of this technique is that the component has to be manufactured first. Alternatively, a process simulation can be performed. By simulating the injection molding process of the GFRP component the filling of the cavity and the orientation state of the final part can be predicted at discrete points. Incorporating quantities from a process simulation in the material modeling and structural simulation is called »integrative simulation«. The methodology of how the integrative simulation is realized in the present work is illustrated in Figure 1.

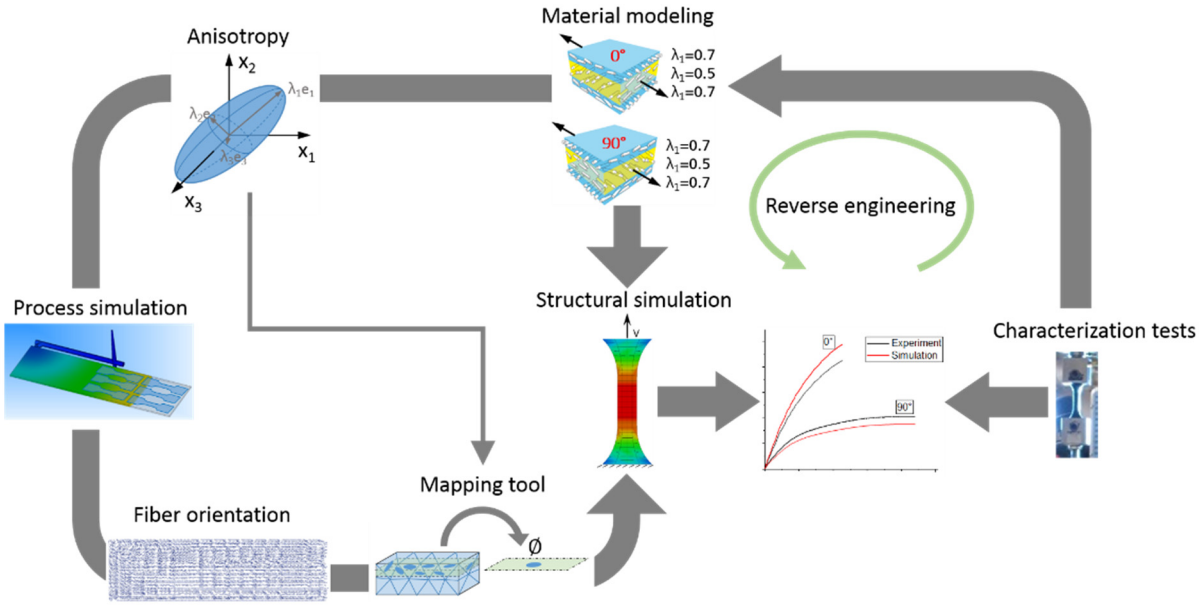


Figure 1: Methodology of the realized integrative simulation.

Centerpiece of this methodology is a mapping-tool which automatically translates the information about the fiber orientation and distribution and which also resolves the mesh inconsistencies caused by different discretizations of the geometry used in the injection molding and in the structural simulation. In order to model different degrees of anisotropy, several material classes are introduced. For each material class, a material card is calibrated, for which the elastic constants are determined using a method described by Advani and Tucker [4]. The plastic parameters, which are the same for all material classes/cards, are identified by an optimization software.

In case of injection molded GFRP the orientation tensor is calculated at discrete material points in the process simulation. Strictly speaking the second order orientation tensor a_{ij} is calculated, which can be described as a symmetric 3x3 matrix. In order to visualize a_{ij}, its eigenvalues and eigenvectors are generated by performing a principal axis transformation. The eigenvectors display the principal directions of the fiber alignment, while the eigenvalues indicate the fiber orientation distribution, ranging from 0 to 1, in the corresponding direction. The sum of all three eigenvalues is equal to 1. By displaying the eigenvectors and eigenvalues in a Cartesian coordinate system an orientation ellipsoid can be defined (see Figure 2).

$$a_{ij} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ & a_{22} & a_{23} \\ sym. & & a_{33} \end{bmatrix} \rightarrow \begin{matrix} \text{Eigenvectors: } e_i \\ \text{Eigenvalues: } \lambda_i \end{matrix}$$

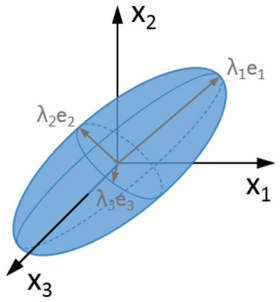


Figure 2: Orientation ellipsoid stretched from the eigenvectors and eigenvalues of the second order orientation tensor.

It is very unlikely that the fibers within one finite volume in the component are aligned in one direction. Instead the fibers are randomly deviated to a greater or lesser extent and lead to varying degrees of anisotropy and hence to varying shapes of the orientation ellipsoid.

3 BASIC PRINCIPLE OF THE DEVELOPED MAPPING TOOL

The flow-chart describing the mapping tool is shown in Figure 3. After importing the shell element model, the tetrahedral element model, and the orientation tensors a calculation algorithm is started. Those orientation tensors, introduced by Advani and Tucker, have been determined by an injection molding simulation for the tetrahedral elements. The mapping tool now divides the shell elements into several layers and calculates the space that would be occupied by each layer based on its thickness. Afterwards, the tetrahedral elements corresponding to this space are assigned and averaged orientation tensors are calculated for each layer of the shell elements, taking into account its volume fraction. The eigenvalues (λ_1, λ_2, λ_3) and eigenvectors (e_1, e_2, e_3) of the orientation tensors are determined by a principal axis transformation. It is assumed that only the largest eigenvalue λ_1 and its eigenvector e_1 are relevant for the calculation of the necessary parameters for the structural simulation composite shell element model. The vector $\lambda_1 e_1$ of each layer is now projected onto the shell element in order to eliminate its component normal to the element surface. The rotation angle, which rotates the fiber direction of the local shell element coordinate system in the direction of the projected

eigenvector, is directly entered into the composite shell element model. Depending on the length of the projected eigenvector, a material class is assigned to each layer. The number of the corresponding material card is also entered into the composite shell element model.

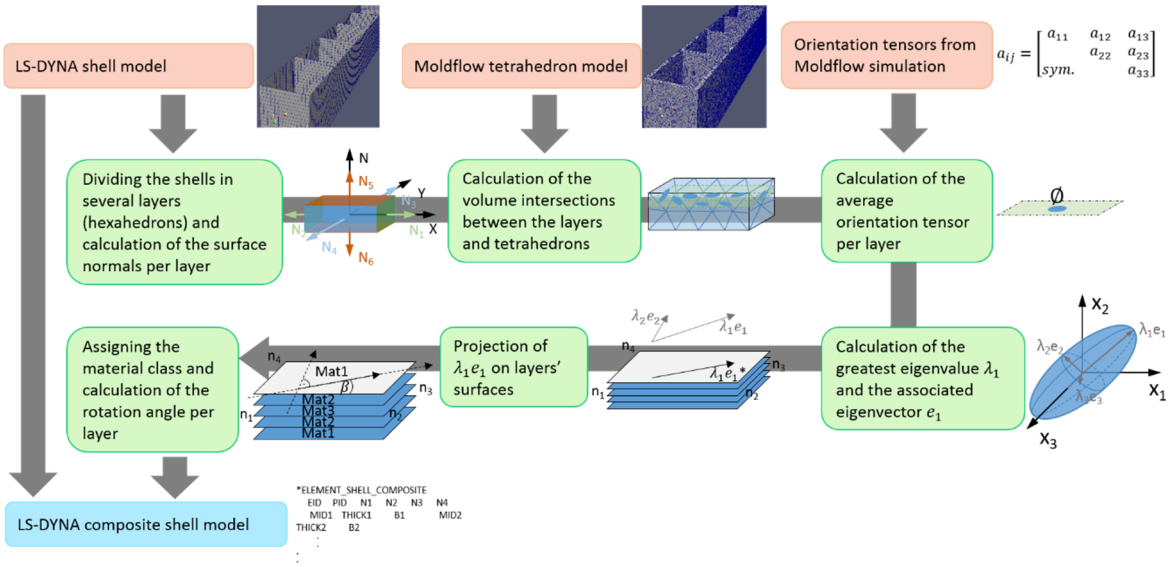


Figure 3: Flow-chart of the developed mapping-tool.

4 DEGREE OF ANISOTROPY

For the definition of the degree of anisotropy, several material classes have to be declared at first. As mentioned before the sum of all three eigenvalues equals 1. In the present study a long fiber reinforced thermoplastic is investigated and the probability of fibers oriented in thickness direction is very low. Therefore, it is assumed that the eigenvectors corresponding to the largest and second largest eigenvalue are mainly aligned parallel to the shell plane. The classification of the degree of anisotropy can simply be done by defining ranges of the largest eigenvalue and each range is represented by a material class. Because the sum of all eigenvalues is 1 and the eigenvalue in the thickness direction is assumed to be 0, the minimum value of the largest eigenvalue is 0.5. In the present study three material classes are defined with the background that a composite made of a GFRP material is parted in two outer layers, two transition layers and one central layer (see Figure 4). An example for a range specification of the largest eigenvalue and the assignment of the corresponding material classes is presented in Figure 5.

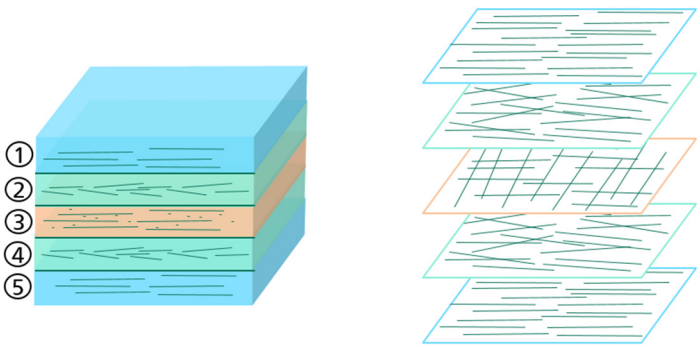


Figure 4: Significant layers: two outer layers (1 and 5), two transition layers (2 and 4), and one central layer (3) within a composite.

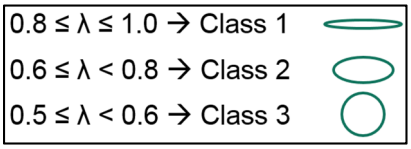
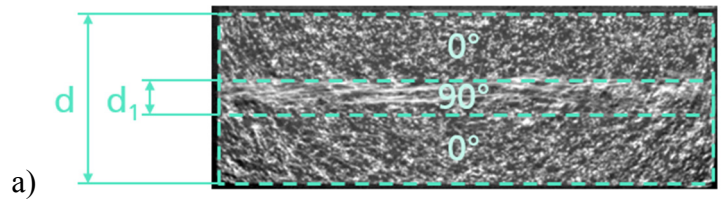


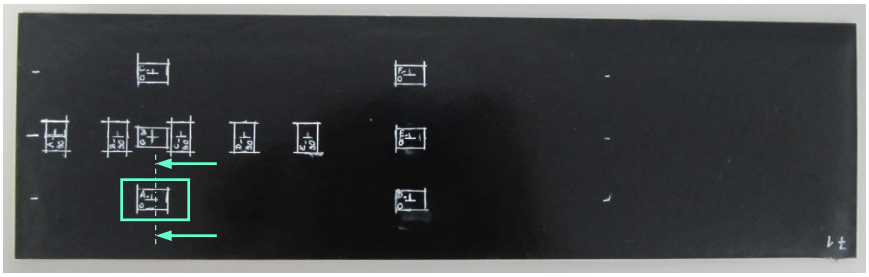
Figure 5: Material classification (degree of anisotropy) used for the material modeling.

5 MATERIAL MODELING

Advani and Tucker [4] proposed an approach to calculate the engineering constants of a material with arbitrary fiber distribution by means of the engineering constants in the uni-directional state. However, from the tensile tests one obtains a combined Young‘s and shear modulus because the fiber distribution in the tested specimens are arbitrary. The approach proposed by Advani and Tucker is therefore reversed. Judging from the injection molding simulation results of the sample plate the central layer is of material class 3 and the outer and transition layers of material class 2. Also from the CT-scans no significant difference between the outer and transition layers can be observed. As an example the CT-scan of a specimen located in the central part of a tensile specimen aligned in 0°-direction is shown in Figure 6 a. The extraction position from the sample plate and the location of the presented CT-scan can be seen in Figure 6 b. It can be observed, that the thickness of the central layer is approximately one fifth of the total thickness. Hence, specifying five layers over the thickness, with each layer having a thickness of one fifth of the total thickness, can be justified.



a)



b)

Figure 6: CT-scan of a specimen extracted from position A on the sample plate.

To calculate the Young’s and shear moduli for the three specified material classes the Young’s moduli and the in-plane shear modulus determined from tensile and shear tests in 0°- and 90°-direction are used.

It is assumed, that the Young's moduli and the in-plane shear modulus are composed as follows:

$$E_{0^\circ} = \eta E_2^{(3)} + (1-\eta)\, E_1^{(2)}\,, \tag{1}$$

$$E_{90^\circ} = \eta E_1^{(3)} + (1-\eta)\, E_2^{(2)}\,, \tag{2}$$

$$G_{0^\circ/90^\circ} = \eta G_{12}^{(3)} + (1-\eta)\, G_{12}^{(2)}\,, \tag{3}$$

with $\eta = 0.2$.

The upper index indicates the material class and the lower index indicates the mainly involved direction of the orthotropic material.

By means of the Young's moduli and the shear modulus from the characterization tests the engineering constants of the uni-directional material is calculated using an optimization algorithm. In the uni-directional state all fibers are aligned in one direction and the largest eigenvalue is $\lambda_1 = 1$. For material classes 1, 2 and 3 values of $\lambda_1^{(1)} = 0.9$, $\lambda_1^{(2)} = 0.7$ and $\lambda_1^{(3)} = 0.5$ are used. With the normalized moduli of $E_{0^\circ} = 3.74$ MPa, $E_{90^\circ} = 2.05$ MPa and $G_{0^\circ/90^\circ} = 1$ MPa the normalized engineering constants for each defined material class are obtained as shown in Table 1. Because the optimization algorithm could not calculate reasonable values for the Poisson ratios, the values of $\nu_{12} = 0.34$, $\nu_{23} = 0.2$ and $\nu_{31} = 0.21$ are assumed.

Material class	λ_1	E_1 [MPa]	E_2 [MPa]	G_{12} [MPa]
1	0.9	6.20	1.14	0.61
2	0.7	4.01	1.90	0.98
3	0.5	2.64	2.64	1.10

Table 1: Engineering constants for the three defined material classes.

For each of the three material classes one MAT_108 material card is created. While the previously calculated engineering constants can be inserted directly for the elastic parameters, further calibrations have to be done to generate the plastic parameters. Hence, a structural simulation model of the tensile test in 0°-, 45°- and 90°-direction at nominal strain rate of 100/s is created. The shell meshes used for the structural simulations and the main alignment of the fibers in each direction are presented in Figure 7.

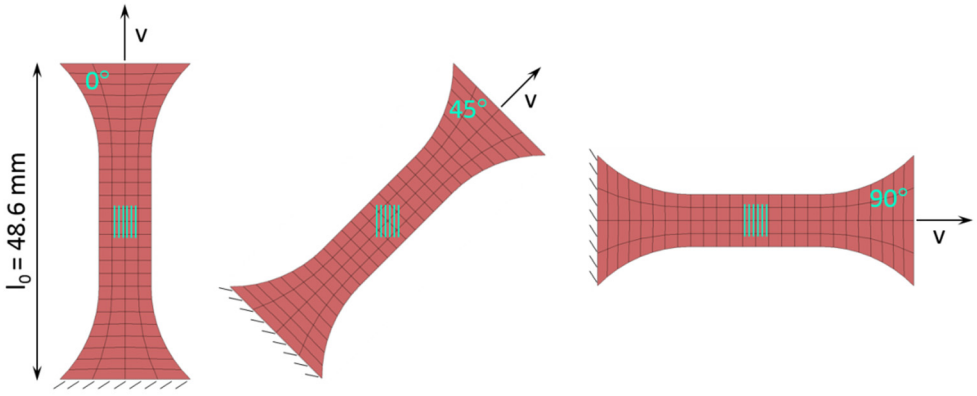


Figure 7: Shell meshes used for the simulations of the tensile test at nominal strain rate of 100/s in 0°-, 45°-, and 90°-direction.

To account for the fiber orientation distribution, the injection molding simulation result of the sample plate is mapped to the shell mesh of the structural simulation models.

By means of the software LS-OPT by LS-DYNA the plastic parameters are calibrated and the same values are used for all material cards. For this the respective smooth force-displacement curves of the tensile tests in 0°-, 45°-, and 90°-direction at a nominal strain rate of 100/s (haul-off speed of 9720 mm/s) are compared to the simulated curves.

During the calibration of the plastic parameters with LS-OPT the elastic parameters for each material class are held constant.

Failure is modeled by assigning one MAT_ADD_EROSION card to each MAT_108 material card, whereby the defined failure criteria are identical for all three MAT_ADD_EROSION cards. The failure parameters are obtained from the tensile tests at a nominal strain rate of 100/s. As failure criteria a value for the maximum stress and maximum strain is inserted. The value for the maximum strain is the average failure strain from the tensile tests at haul-off speed of 9720 mm/s in 90°-direction. The maximum stress is generated by reverse engineering. To prevent the failure of the integration points under compression loading a value of -20 MPa (pressure is negative for tensile) for the minimum pressure is inserted as well. Variable *NCS* (number of failure conditions to satisfy before failure occurs) is set to 2. In this manner failure occurs if two failure criteria are met, e.g. maximum stress and minimum pressure.

As can be seen in Figure 8 and Figure 9 the force-displacement and stress-strain curves from the tensile tests at haul-off speed of 9720 mm/s in 0°-, 45°-, and 90°-direction can be represented well with the generated elastic and yield parameters. Also, failure can be simulated sufficiently.

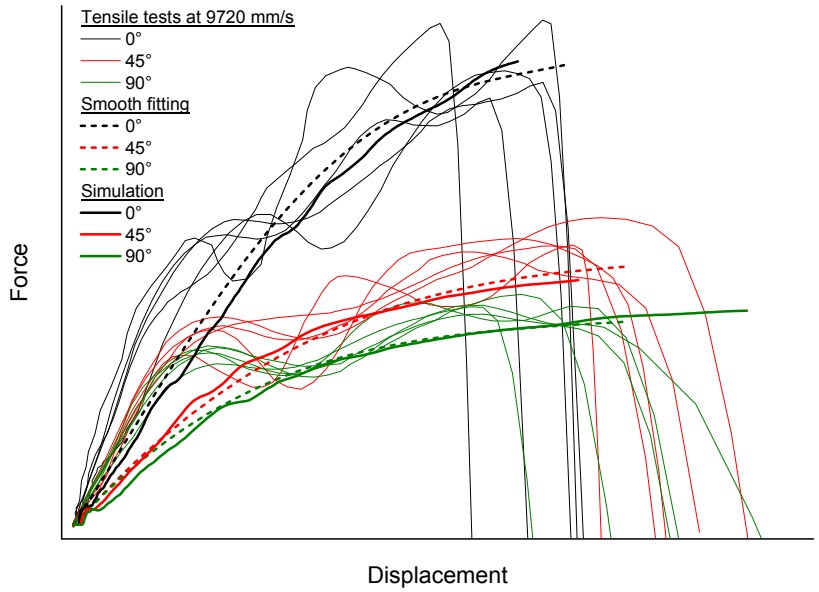


Figure 8: Comparison of the simulated and experimentally determined force-displacement curves from the tensile tests at a haul-off speed of 9720 mm/s in 0°-, 45°-, and 90°-direction.

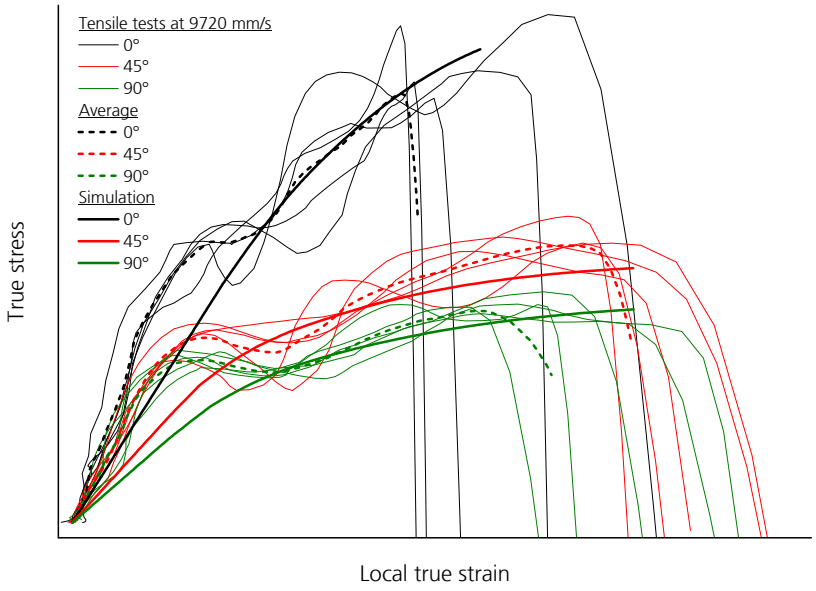


Figure 9: Comparison of the simulated and experimentally determined stress-strain curves from the tensile tests at a haul-off speed of 9720 mm/s in 0°-, 45°-, and 90°-direction.

6 VALIDATION OF THE METHOD

To validate the developed approach comprising the mapping tool, the calculation of engineering constants for various degrees of anisotropy, and the generation of the material cards, three-point bending tests are performed on a component part made of a PPGF30 material. The test set-up on the servo-hydraulic high strain rate testing machine is shown in Figure 10. For better tracking of the deformation the edges of the components are painted with a white marker. Furthermore, a stop on the left side and on the back is used to align the components and thus to minimize the scatter.

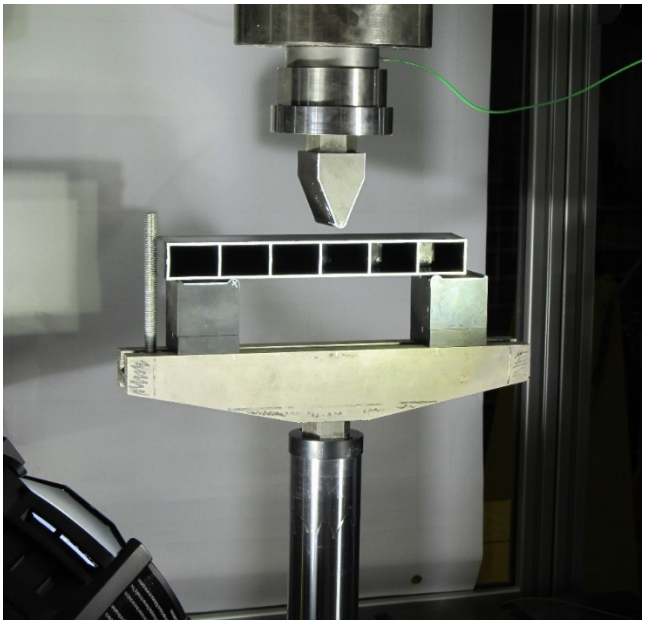


Figure 10: Three-point bending test set-up on the VHS testing machine.

Due to the draft angle the outer ribs of the component form a trapezoid and also the thickness of the ribs increases in the direction of the bottom. The trapezoidal shape is considered in the simulation model and the component is aligned accordingly. The simulation model of the three-point bending test is shown in Figure 11. Here it can be seen, that the ribs are divided in 19 sections, or 19 shell element rows, over the height. To represent the varying thickness of the ribs, the average thickness of each section is assigned to the shell elements within the sections.

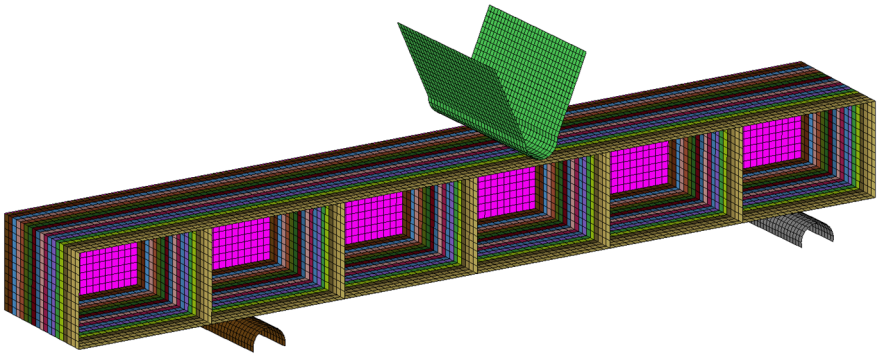


Figure 11: Simulation model of the three-point bending test.

The generated structural simulation model of the component consists of 8155 shell elements and 8200 nodes and the calculations are performed with LS-DYNA version R7.1.1 revision 88541. The edge length of the elements is approximately 2 mm and fully integrated shell elements (*ELFORM* = 16) are used. The supports and the fin are represented by surfaces and are defined as rigid bodies. While the supports move upwards, the fin is fixed. The velocity boundary condition assigned to the supports is derived from the velocity of the piston during the test. Between the component and the rigid bodies, a contact is defined, with an assumed static friction value of $FS = 0.2$, a dynamic friction value of $FD = 0.15$ and a decay of $DC = 0.5$. Also a self-contact is defined for the component and the same friction values as mentioned before are assumed.

The force-displacement curves of the five valid tests and the simulation are shown in Figure 12. The progressions of the test curves are similar and only a small scatter can be observed. It is assumed that the large scatter at larger displacements can be explained by different post failure behavior of the component in each test. Due to the oscillations and the fact, that only the front side of the component can be observed it is difficult to trace variations of the force signal to occurring deformations of the part. However, four pictures, which show the deformation of the components at significant incidents during the tests, are selected exemplarily from Test 1. The selected pictures and the associated moment in the force-displacement curve (marked as blue rectangles in the diagram) of the experiment and the simulation are presented in Figure 12. The first picture shows the initial contact between the fin and the part. The total failure of the upper rib right below the fin can be seen in the second picture. After a drop of the force-displacement curve the force increases again. As can be seen in the third picture, this is where the edge between the left transverse and left upper rib strikes the fin. The last picture shows the total failure of the lower rib, but due to the varying post failure behavior in each test this picture is of no significance.

The simulation result agrees well with the experiments regarding the curve progression, the force level and also the deformation behavior and thus the predictive power of the method can be illustrated. However, it should be noted that the simulation result strongly depends on the number of failed integration points prior to element deletion *NUMFIP* in the

MAT_ADD_EROSION card and on the friction values in the contact definitions. For the present simulation the previously mentioned contact friction values are inserted and for *NUMFIP* a value of -105 (When *NUMFIP* < -100, elements erode when |*NUMFIP*|-100 integration points fail. Also, for *NUMFIP* < -100, the stress at an integration point immediately drops to zero when failure is detected at that integration point) is applied.

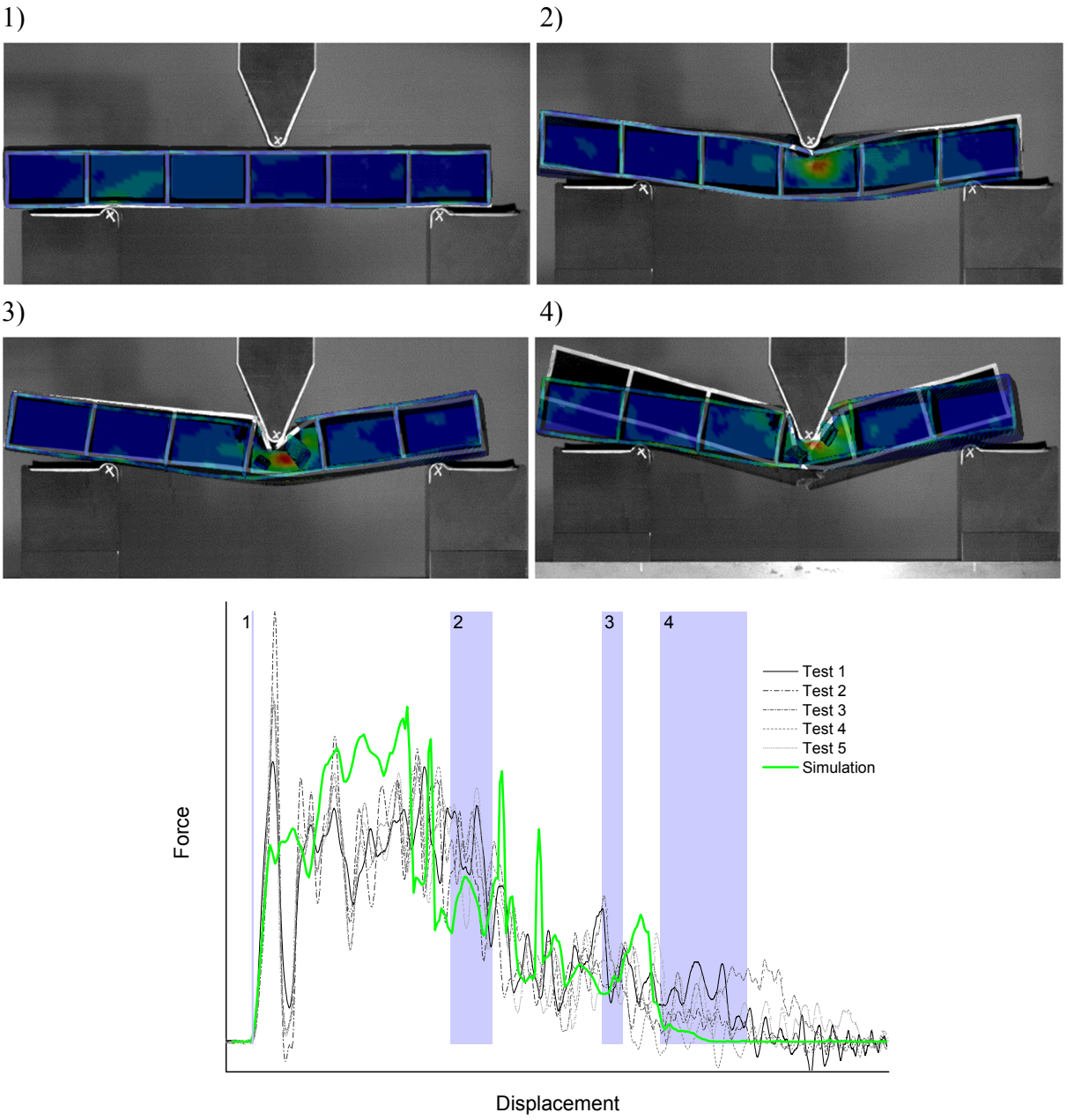


Figure 12: Simulated and experimentally determined force-displacement curves from the dynamic three-point bending tests and pictures of significant incidents during the test and simulation.

7 FIBER ANALYSIS BASED ON CT-SCANS

To further improve the results of an integrative simulation, the injection molding simulation results have to be validated. For this, the fiber orientation in the real material has to be determined via CT-scans and a fiber analysis. Such a fiber analysis was done using the software VGStudio MAX 2.2. The exemplary result of a fiber analysis of the investigated

PPGF30 material is shown in Figure 13 together with the fiber orientations determined from the injection molding simulation. On the bottom, cross sectional views of the material with thickness *t* are shown. The different colors in the middle layer of the cross sectional view of the fiber analysis indicate a different orientation of the fibers in this layer compared to the outer layers. Sectional views through the middle layer (A) and the outer layer (B) confirm this result: The fibers in the outer layer are oriented mainly in flow direction (0°), whereas in the middle layer the fibers are deflected from the flow direction by angles between 40°-60°. The same can be seen in the results from the injection molding simulation. However, the deflection of the fibers in the middle layer is much lower (10°-15°) than in the real material.

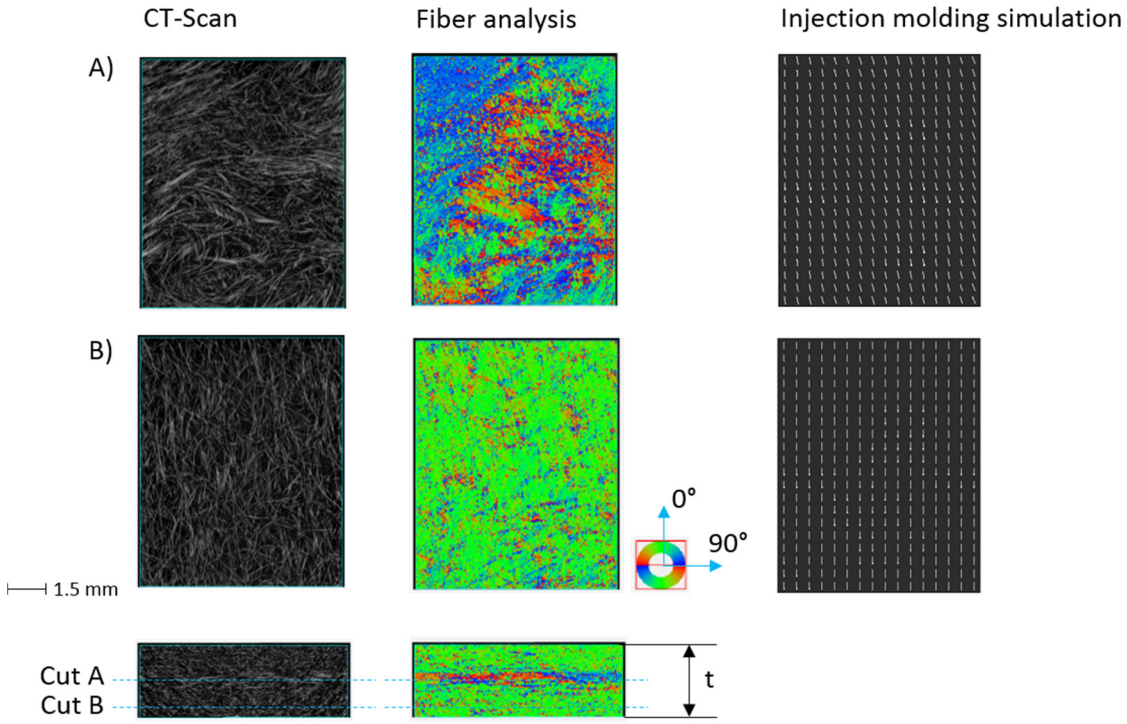


Figure 13: Fiber analysis of an injection molded fiber-reinforced plastic based on CT-scans compared to an injection molding simulation.

8 IMPACT OF CONSIDERUNG THE DEGREE OF ANISOTROPY

In the presented methodology of an integrative simulation the degree of anisotropy is considered in a gradual manner by defining several material classes, each covering different ranges of the greatest eigenvalue. But considering the degree of anisotropy in this way is a rather time consuming approach, because for each material class, one set of parameters has to be calibrated iteratively using the software LS-OPT. Due to this fact it is reasonable to ask about the advantages of considering the degree of anisotropy compared to less complex modelling approaches. To answer this, the effect of considering the degree of anisotropy was further investigated by performing two additional simulations of the dynamic three-point bending test. In the first simulation only the fiber orientation was considered. Furthermore, the values for the elastic and plastic parameters in the longitudinal and transversal material direction are derived solely from the dynamic tensile tests in 0°- and 90°-direction, respectively. In the second simulation an isotropic material behavior was used. For both simulations all other parameters, e.g. friction values, are adopted from the structural simulation model presented in section 6.

In Figure 14 the simulation considering only the fiber direction, and in Figure 15 the simulation using an isotropic material behavior is compared to the experiment and the simulation considering the degree of anisotropy.

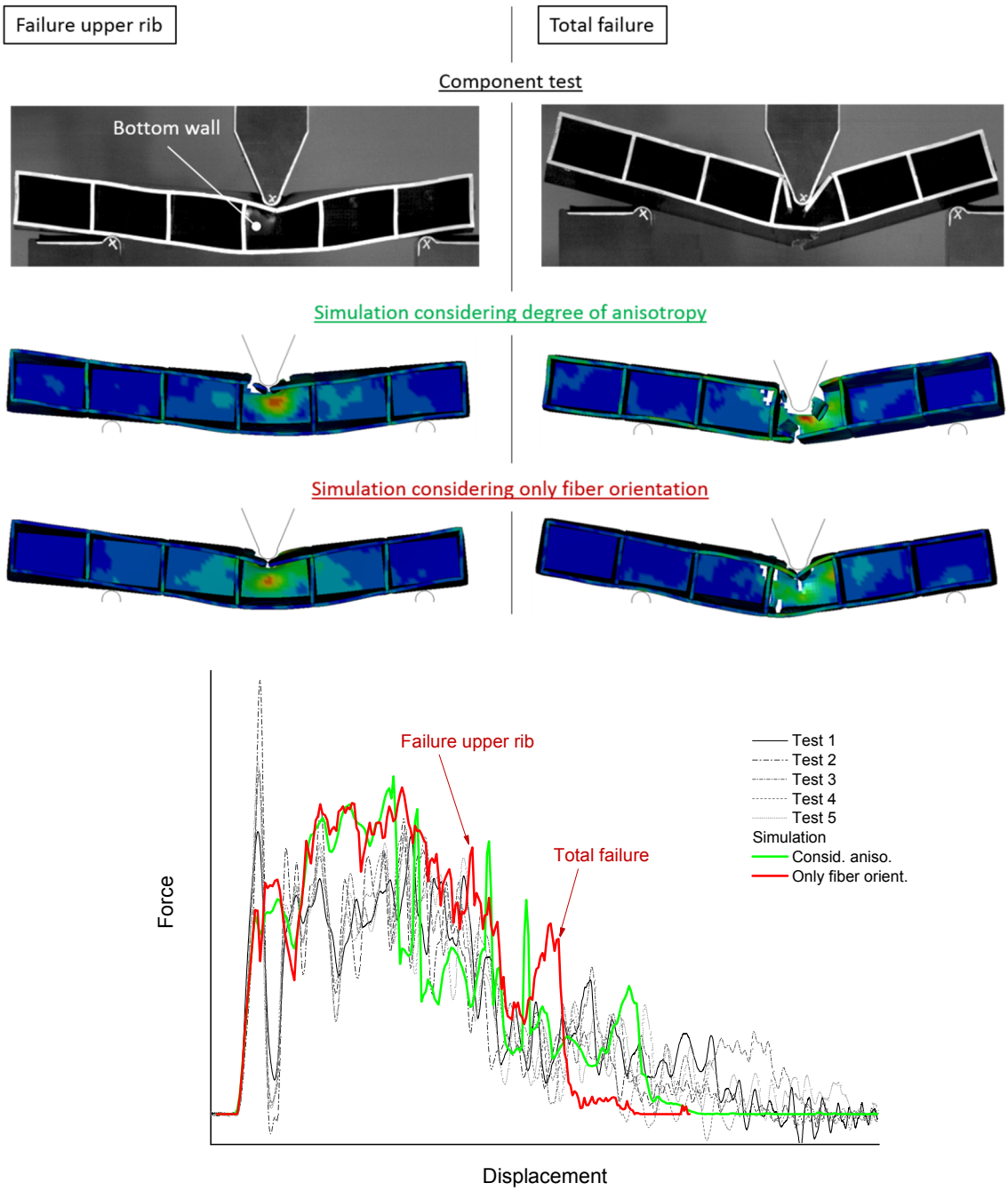


Figure 14: Comparison of the simulation considering only the fiber orientation with the simulation considering the fiber orientation and the degree of anisotropy.

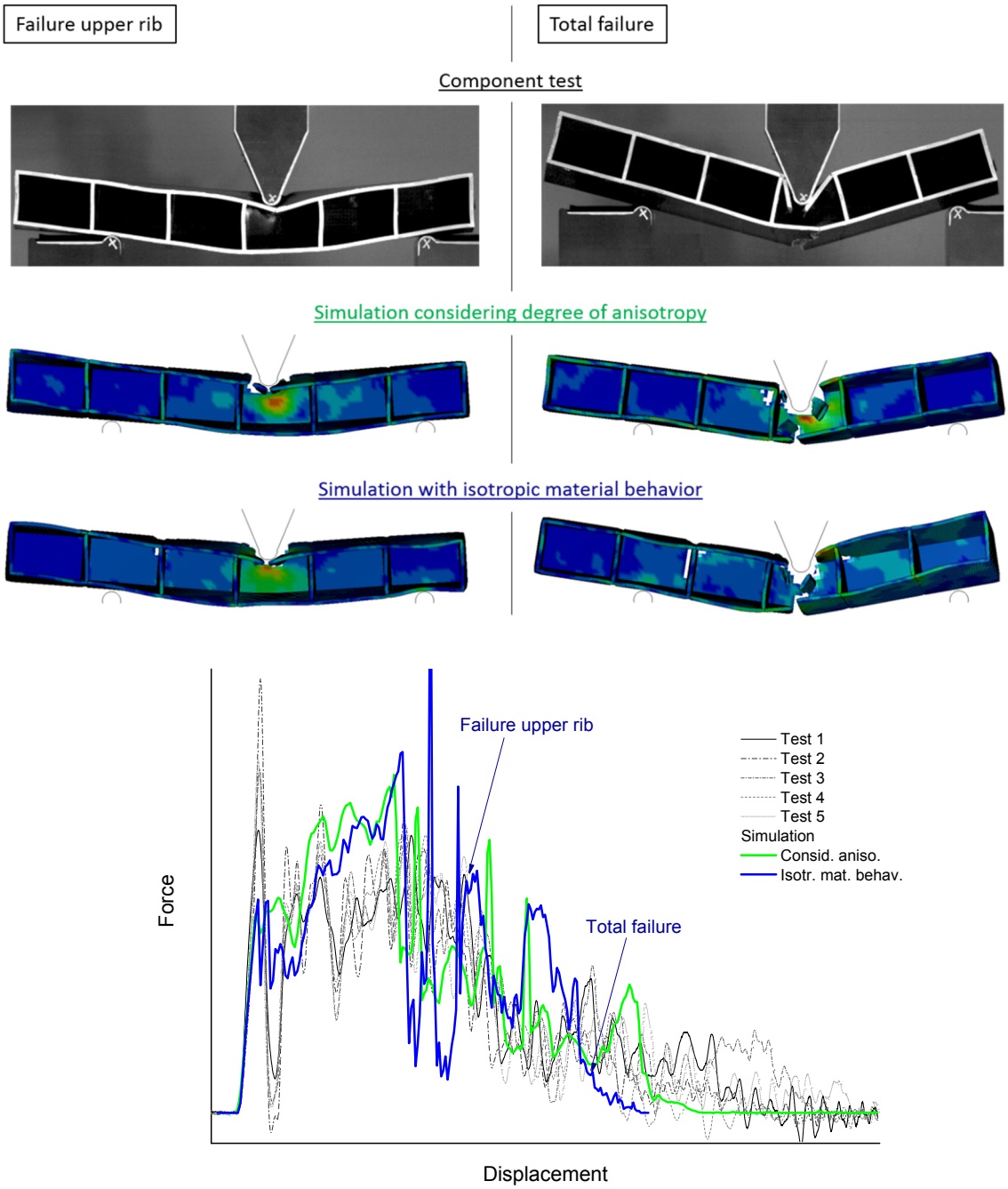


Figure 15: Comparison of the simulation using an isotropic material behavior with the simulation considering the fiber orientation and the degree of anisotropy.

Regarding the initial curve progression and the force level, both simulations show equally good agreement with the experiment. But in contrast to the experiment in the simulation considering only the fiber orientation the upper part of the bottom wall remains intact longer. In addition, the upper rib is not bent. For this reason the force increases significantly and the total failure occurs much earlier than in the experiments. The same can be observed in the simulation using an isotropic material behavior. Here the upper part of the bottom wall fails in a larger region, but the edge between the upper and bottom wall is not impressed. Latter can be observed in all other simulations and the experiment.

In conclusion the simulation considering the degree of anisotropy shows the best agreement with the experiment throughout the entire length of the test, regarding the curve progres-

sion, the force level and the deformation behavior. When considering only the fiber orientation or using an isotropic material behavior, the deformation behavior cannot be represented well and in the present study failure occurs too early. However, regarding the curve progression and force level only, the simulations show comparable good results.

9 CONCLUSIONS

- The mechanical behavior of a long fiber reinforced thermoplastic material is characterized and modeled in detail by taking into account the fiber orientation distribution within the material
- For this purpose, the fiber distribution is thoroughly investigated by performing injection molding simulations and CT-scans of a ribbed component and a sample plate, from which the specimens for the characterization tests are extracted. Together with the results of injection molding simulations, characterization tests and additional structural simulations of the characterization tests, parameters for the material modeling are identified.
- In the present study this approach, also known as integrative simulation, is shown by means of a PPGF30 material, a long glass fiber reinforced thermoplastic. But it should be generally applicable for all long fiber reinforced plastics.
- The characterization of the orientation dependent material behavior includes tensile tests at different strain rates as well as tensile-unloading, compression and shear tests in 0°-, 45°-, and 90°-direction
- Injection molding simulations of the sample plate and the component have been carried out and provide information about the fiber orientation state in terms of the second order orientation tensor
- The largest eigenvalue of the orientation tensor is a measure for the degree of anisotropy and the associated eigenvector delivers the principle fiber direction. The classification of the degree of anisotropy is made in terms of several material classes, which cover different ranges of the largest eigenvalue. For each material class one MAT_108 material card is defined. The determination of the material parameters is done in two steps. At first the elastic parameters are calculated by means of an optimization algorithm. The underlying equation is derived from an approach proposed by Advani and Tucker to calculate the engineering constants of a material with arbitrary fiber distribution by means of the engineering constants in the uni-directional state. In a second step a structural simulation model of the tensile tests at haul-off speed of 9720 mm/s in 0°-, 45°-, and 90°-direction is created.
- In order to also account for the fiber orientation state in the material a mapping tool is developed, which automatically calculates quantities from the injection molding simulation results and integrates them in the structural simulation model. With this mapping tool the fiber orientation and the material card is specified for each layer of the shell element meshes of the tensile specimens with the keyword ELEMENT_SHELL_COMPOSITE in LS-DYNA.
- The calibration of the plastic material parameters is carried out using the software LS-OPT. In addition, the failure behavior is modeled by assigning one MAT_ADD_EROSION card to each MAT_108 material card. The failure parameters are generated by reverse engineering.

- With the elastic, plastic and failure parameters obtained in this way the force-displacement and stress-strain curves of the tensile tests at haul-off speed of 9720 mm/s can be represented well in 0°-, 45°-, and 90°-direction.
- For further validation of the mapping tool and the generated material parameters dynamic three-point bending tests are performed on a ribbed component. Judging from the force-displacement curves and the analysis of the deformation of the component the experiments can be represented well by the generated structural simulation model.
- CT-scans are made from selected areas of the sample plate and the component. Based on the data obtained from these CT-scans fiber analyses are performed and serve as validation of the injection molding simulation results.
- The impact of considering the degree of anisotropy was investigated in detail and a comparison with a simulation considering only the fiber orientation and a simulation using an isotropic material behavior was made. The simulation considering the degree of anisotropy showed the best agreement with the experiment throughout the entire length of the test, regarding the curve progression, the force level and the deformation behavior. The other two simulations could not represent the deformation behavior well, but regarding the curve progression and force level, comparable good results can be shown.

10 ACKNOWLEDGEMENT

This work was developed in coorperation between the EMI and Fuji Heavy Industries Ltd. and the authors wish to acknowledge the support and contribution of all project partners.

REFERENCES

[1] G. Gruber, A. Haimerl, S. Wartzack, Consideration of Orientation Properties of Short Fiber Reinforced Polymers within Early Design Steps. *12th International LS-DYNA Conference*, 2012.

[2] Livermore Software Technology Cooperation, Livermore, LS-DYNA Keyword User's Manual: Volume I, Version R7.1., Cal., May, 2014.

[3] Livermore Software Technology Cooperation, Livermore, LS-DYNA Keyword User's Manual: Volume II Material Models, Version R7.1., Cal., May, 2014.

[4] S.G. Advani, C.L. Tucker, The Use of Tensors to Describe and Predict Fiber Orientation in Short Fiber Composites. *Journal of Rheology*, 31(8), 751-784, 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CONVERGENCE ANALYSIS OF STRESS FIELDS TO HOMOGENIZATION PREDICTIONS IN OPTIMAL PERIODIC COMPOSITE DESIGN

Pedro G. Coelho[1], Rui A. Reis[1], and José M. Guedes[2]

[1] UNIDEMI, DEMI, FCT-UNL
2829-516 Caparica
{pgc@fct.unl.pt, rf.alvesreis@gmail.com}

[2] IDMEC, Instituto Superior Técnico, Universidade de Lisboa
1049-001 Lisboa
jmguedes@tecnico.ulisboa.pt

Keywords: Homogenization, Composites, Cellular, Stress, Topology, Optimization.

Abstract. *Homogenization assumes that a unit-cell of a periodic composite material is infinitely small and it has periodic boundary conditions. In practice, such material comprises a finite number of measurable unit-cells and the stress fields are not periodic near the structure boundary. It is thus critical to investigate in the scope of the present work whether the optimized unit-cell topologies obtained are affected when applied in the context of real composites. This is done here by scaling the unit-cell an increasing number of times and accessing the micro (or local) stresses of the resulting composite by means of standard numerical experiments and comparing them to the homogenization predictions. The outcome indicates that it is sufficient to have a low scale factor to replace the non-homogeneous composite by the equivalent homogeneous material with the stress field computed by homogenization.*

1 INTRODUCTION

Stress-based topology optimization problems are more realistic to engineering practice than the conventional maximum stiffness approaches because one may guarantee not only very efficient designs but also feasible ones. However, dealing with stress criteria is specially challenging due to the: (1) "singularity" problem; (2) local nature of the stress constraint; (3) highly non-linear stress behavior. Several contributions have been made to handle these issues to the point that fine results are now obtained for macroscopic structures [1-4]. The present paper expands the analysis of stresses to material microstructures which may lay the groundwork for the optimal topology design of material "unit-cells" including stress criteria [5]. An "unit-cell" represents here the smallest periodic heterogeneity of periodic composite/cellular media. In general, periodic homogenization models are used to compute the elastic properties and local stresses of periodic composite materials based on the shape/periodicity of a given material unit-cell [6,7]. Conversely, in material design, the unit-cell is not known a priori, and the goal is to design it to attain specific properties values – inverse homogenization problem [8]. This design problem is solved here by formulating it as an optimization problem. However, homogenization assumes that the unit-cell is infinitely small and it has periodic boundary conditions (BC's). In practice, the composite material comprises a finite number of measurable unit-cells and the stress fields are not periodic near the structure boundary. It is thus critical to investigate in the scope of the present work whether the obtained unit-cell topologies are affected when applied in the context of real composites. This is done here by scaling the unit-cell an increasing number of times and accessing the micro (or local) stresses of the resulting composite by means of standard numerical experiments and comparing them to the homogenization predictions. This furthers previous work [9] related to compliance and elastic coefficients convergence to homogenization predictions. Here, the outcome also indicates that it is sufficient to have a low scale factor to replace the non-homogeneous composite by the equivalent homogeneous material with the stress field computed by homogenization.

2 MATERIAL MODEL

Figure 1 presents the periodic material model mixing strong, $E^{(1)}$, and weak, $E^{(2)}$, materials (see also [9]). The unit-cell volume Y (of feature size d) is cubic and one repeats it n^3 times resulting in volume Ψ ($D = |\Psi| = 1$). Therefore, the ratio $n = D/d$ can be seen as a scale factor.

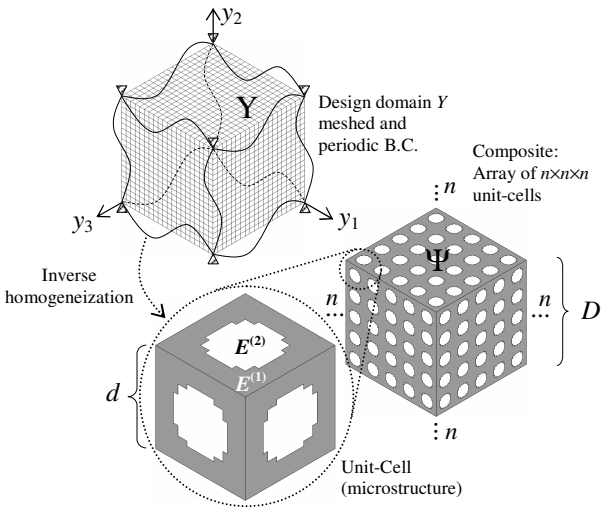


Figure 1: Material model. Finite element discretization of Y with periodic BC's. Array of 5×5×5 unit-cells of global size D (macroscale) and one unit-cell of size d (microscale).

The asymptotic homogenization model used implies that stresses are calculated when $n \to \infty$ and thus no size is given to the unit-cell, i.e. $d \to 0$. The stiffness ratio, $E^{(1)}/E^{(2)}$, is here equal to 10^1 or 10^{12} representing a composite or cellular material, respectively. Both phases are solid isotropic. One keeps phase 1 as the reference, i.e. $E^{(1)} = 300\text{MPa}$ and $\nu = 0.3$ (properties of a biodegradable polymer currently used in scaffolds for tissue engineering, see [10]).

3 OPTIMIZATION PROBLEM

The unit-cell topology shown in Figure 2 is obtained by inverse homogenization solving the following compliance minimization problem subject to a volume fraction constraint [11],

$$\min_{\boldsymbol{\mu}} \quad \frac{1}{2}\boldsymbol{\sigma}^0 \mathbf{C}^{\mathrm{H}}(\boldsymbol{\mu})\boldsymbol{\sigma}^0 \qquad \text{s.t.} \quad \int_Y \mu(\mathbf{y})\,\mathrm{d}Y \le V^* \tag{1}$$

Here $\boldsymbol{\sigma}^0$ is a macroscopic hydrostatic stress state ($\sigma = 1\text{MPa}$), volume $|\mathrm{Y}| = 1$ and $\mathbf{C}^{\mathrm{H}}$ is the homogenized compliance tensor computed as the inverse of the stiffness tensor $\mathbf{E}^{\mathrm{H}}$ through,

$$E^{\mathrm{H}}_{ijmn}(\mu) C^{\mathrm{H}}_{mnkl}(\mu) = \frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) \tag{2}$$

where δ_{ij} are Kronecker deltas and,

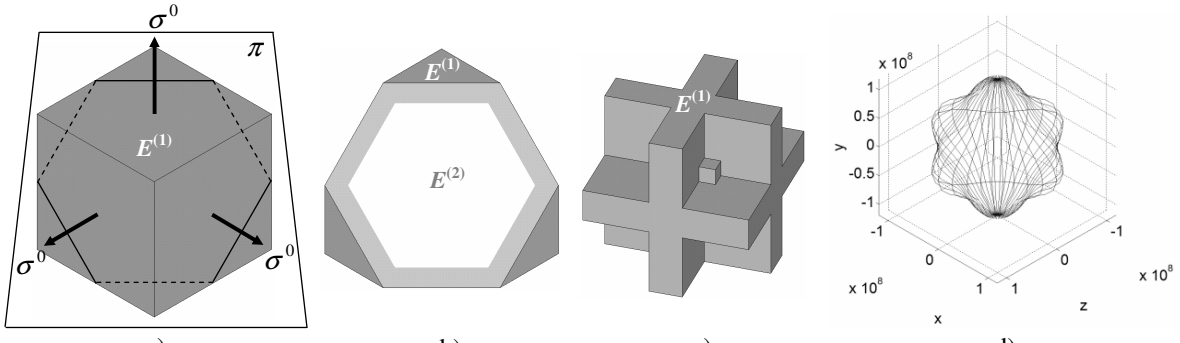
$$E^{\mathrm{H}}_{ijkm}(\mu) = \frac{1}{|\mathrm{Y}|}\int_{\mathrm{Y}} E_{pqrs}(\mu)\left(\delta_{pk}\delta_{qm} - \frac{\partial \chi^{km}_p}{\delta y_q}\right)\left(\delta_{ri}\delta_{sj} - \frac{\partial \chi^{ij}_r}{\delta y_s}\right)\mathrm{dY} \tag{3}$$

with the microstructure material elastic properties, E_{pqrs}, depending on density design variables, $\boldsymbol{\mu}$, in order to interpolate between two base materials according to the power law,

$$E_{pqrs}(\mu) = \mu^p E^{(1)}_{pqrs} + \left(1 - \mu^p\right)E^{(2)}_{pqrs}, \quad p \in \mathbf{N} \tag{4}$$

and the homogenized tensor depends on the material unit cell deformation modes or micro-displacements χ^{kl} (Y-periodic), which are solution of the set of equilibrium equations defined in Y (six equations in three dimensions),

$$\int_{\mathrm{Y}} E_{ijpq}(\mu)\frac{\partial \chi^{kl}_p}{y_q}\frac{\partial w_i}{\partial y_j}\mathrm{d\,Y} = \int_{\mathrm{Y}} E_{ijkl}(\mu)\frac{\partial w_i}{\partial y_j}\mathrm{d\,Y}, \quad \forall_w \ \mathrm{Y-periodic} \tag{5}$$



a) b) c) d)

Figure 2: Unit-cell. a) Box-type (isometric view); b) Section view using cutting plane π normal to the octahedral direction, c) Shifted design results from translation along octahedral direction (material 2 is unselected only for design comprehension, see also [10]); d) Graphical representation of anisotropy as in [9], units in [Pa].

4 STRESS ANALYSIS

4.1 Homogenization prediction

The micro-stresses σ_{ij} (in the level of material microstructure) are obtained from asymptotic homogenization theory, by considering the displacement test fields χ_k^{rs} from Eq. (5) as,

$$\sigma_{ij} = E_{ijkm}\left(\delta_{kr}\delta_{ms} - \frac{\partial \chi_k^{rs}}{\partial y_m}\right)\varepsilon_{rs}^0 \tag{6}$$

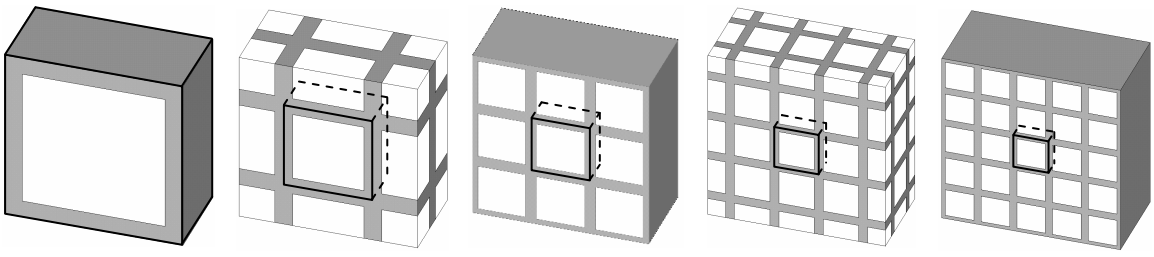
where ε_{rs}^0 is the macroscopic (average) strain field related to the macroscopic stress in (1),

$$\varepsilon_{rs}^0 = C_{rspq}^H \sigma_{pq}^0 \tag{7}$$

The σ_{ij} results represent the three-dimensional stress states varying throughout the unit-cell domain Y (see Figure 4a). Here, the methodology used for obtaining such stresses is based on the software POSTMAT [12,13].

4.2 Numerical experiments

The commercial finite element code ANSYS@ is used here to perform standard numerical testing procedures with specific BC's to calculate the stress field inside the domain Ψ of the periodic composite/cellular material as the scale factor *n* increases from 1 to 5, see periodic patterns in Figure 3. It is shown half the domain, $|\Psi|/2$, in order to highlight the central unit-cell where the stress field is measured (see also Figure 4b,c) and compared to the results given by homogenization running POSTMAT (Figure 4a). Shifted design (as shown in Figure 2c) must be used in some arrays (for *n* = even number) to guarantee that the unit-cell topology extracted from the center of the periodic pattern matches exactly the one (no-shifted) evaluated by POSTMAT. Finite element meshes for each periodic pattern are not shown in Figure 3 because they would become so refined with increasing *n* that elements wouldn't be clearly observed. Anyway, each unit-cell comprises a regular mesh 10×10×10 of 8-node isoparametric hexahedral finite elements. This means that a total of $1000\times n^3$ finite elements are used in the numerical model associated with scale factor *n*.



1×1×1 2×2×2 3×3×3 4×4×4 5×5×5

Figure 3: Volume Ψ (half represented) containing arrays *n*×*n*×*n* of unit-cells where *n* varies from 1 to 5. Unit-cell design is shifted when arrays take even numbers. The unit-cell located at the center of Ψ is highlighted.

One considers the Dirichlet and Neumman-type BC's (see [9,14]) given by, respectively,

$$\mathbf{u}(\mathbf{y})\big|_{\partial\Psi} = \boldsymbol{\Theta}\cdot\mathbf{y}\big|_{\partial\Psi} \tag{8}$$

$$\boldsymbol{\sigma}^0(\mathbf{y})\cdot\mathbf{n}\big|_{\partial\Psi} = \boldsymbol{\Theta}\cdot\mathbf{n}\big|_{\partial\Psi} \tag{9}$$

$$\text{where } \boldsymbol{\Theta} = \begin{bmatrix} \beta & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \beta \end{bmatrix} \tag{10}$$

and $\partial\Psi$ is the boundary of Ψ, **u** is the displacement vector, **y** is the spatial position vector, **n** is the outward normal unit vector and β is the constant characterizing the hydrostatic tensor $\boldsymbol{\Theta}$. On one hand, Eq. (8) applies in $\partial\Psi$ a displacement field **y** linearly dependent such that Ψ is tested at a uniform macroscopic strain β, i.e.

$$\varepsilon_{ij}^{0} = \frac{1}{2}\left(\frac{\partial u_i}{\partial y_j} + \frac{\partial u_j}{\partial y_i}\right) = \beta\delta_{ij} \text{ in } \partial\Psi \text{ and on average } \left\langle \varepsilon \right\rangle_{\Psi} = \frac{\int_{\Psi} \varepsilon \, \mathrm{d}\Psi}{|\Psi|} = \beta\delta_{ij} \tag{11}$$

On the other hand, Eq. (9) means that the test is carried out on Ψ at a uniform macroscopic stress β (also on average holds $\left\langle \sigma \right\rangle_{\Psi} = \beta$). In both tests β must be chosen such that consistency with POSTMAT is ensured, see Eq. (6) and (7). However, BC's in Eq. (9) can't be applied to the analysis of cellular (or porous) material, $E^{(1)}/E^{(2)} = 10^{12}$, because the pressure on the top of the finite elements modeling the void phase results in excessive compliance. Nevertheless, Dirichlet-type BC's can be applied to both ratios, 10^{1} (composite) and 10^{12} (cellular).

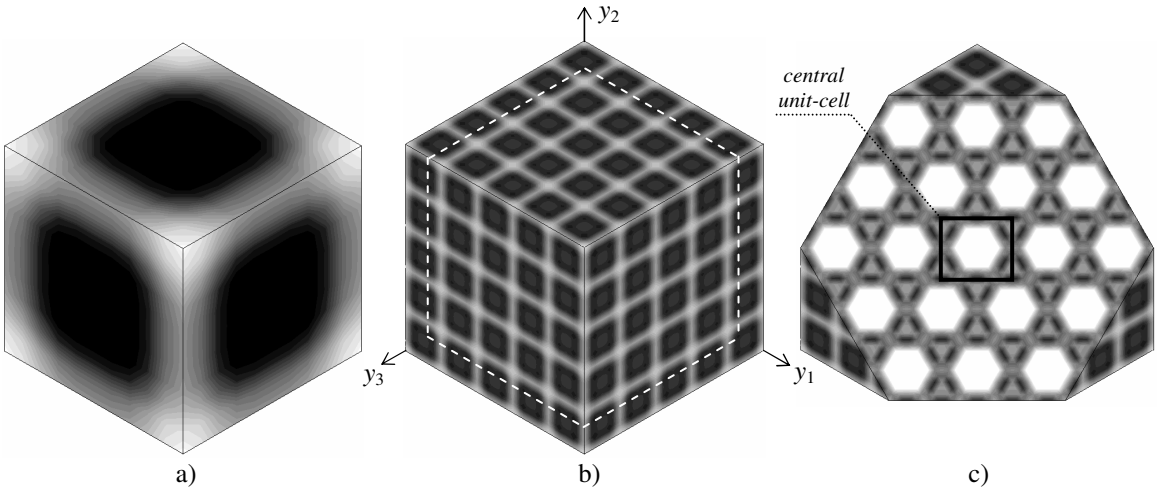


Figure 4: Von-Mises stress is shown for the ratio $E^{(1)}/E^{(2)} = 10^{12}$ (averaged values are shown only for the sake of getting a smoother display). a) POSTMAT results; b) ANSYS results for the array 5×5×5 of unit-cells (undeformed design in dashed lines); c) Section view of the array using plane π as used in Figure 2.

5 CONVERGENCE ANALYSIS

The purpose of this study is the convergence analysis of stress fields to homogenization predictions. These predictions are obtained running the POSTMAT code. Then a battery of ANSYS analyses is carried out such that stresses can be read from the post-processor for each array presented in Figure 3 with specific BC's and stiffness ratio.

ANSYS code provides individual stress components as well as the equivalent stress (Von-Mises) at each node and element. To simplify the comparative analysis between POSTMAT and ANSYS one chooses to compare element equivalent stresses that in ANSYS are computed as follows,

$$\sigma_{eq} = \sqrt{\frac{1}{2}\left[\left(\overline{\sigma}_{11} - \overline{\sigma}_{22}\right)^2 + \left(\overline{\sigma}_{22} - \overline{\sigma}_{33}\right)^2 + \left(\overline{\sigma}_{33} - \overline{\sigma}_{11}\right)^2\right] + 3\left(\overline{\sigma}_{12}^2 + \overline{\sigma}_{23}^2 + \overline{\sigma}_{13}^2\right)} \tag{12}$$

where

$$\overline{\sigma}_{ij} = \frac{\sum_{k=1}^{N} \sigma_{ij}^{k}}{N} \text{ with } N = 8 \text{ (number of Gauss points)} \tag{13}$$

POSTMAT post-processor is properly adapted to do these same calculations of equivalent stress. Finally, one proceeds evaluating the deviation of ANSYS stress results toward POSTMAT according to the formula,

$$\text{Deviation}\,[\%] = \frac{\left(\sigma_{\text{eq}}^{\text{ANSYS}} - \sigma_{\text{eq}}^{\text{POSTMAT}}\right)}{\sigma_{\text{eq}}^{\text{POSTMAT}}} \times 100 \tag{14}$$

This stress deviation measure is evaluated at each finite element of the unit-cell (in the ANSYS model is the central unit-cell, see also Figure 3). Using the mesh 10×10×10, the resulting data for plotting comprises 1000 points. However, due to the symmetry seen in the resulting distribution only half the points are of interest. Therefore, the first five charts presented in Figures 5 to 7 show the elements number (1 to 500) in the abscissa axis and the deviation is represented along the ordinate axis [%]. Each one of these charts is obtained for an increasing number of the scale factor *n*, 1 to 5. Looking only at clusters of points here doesn't help much drawing some conclusions. So, the last chart shown in Figures 5 to 7 is an attempt to provide the scale-size effect analysis in a nutshell and should be seen along with Table 1. This table presents some statistical analysis of the plotted deviations in terms of maximum and minimum values, mean (simple average) and standard deviation. Table 1 and Figures 5 to 7 present data in the same order, i.e. first the composite material case ($E^{(1)}/E^{(2)} = 10^{1}$) is treated and the results with Neumann-type BC's precede the Dirichlet's results. Finally, the cellular material case ($E^{(1)}/E^{(2)} = 10^{12}$) is presented with the respective results considering Dirichlet-type BC's only (as explained in section 4.2).

B.C.	ratio	Measures	1×1×1	2×2×2	3×3×3	4×4×4	5×5×5
Neumann	10^{1}	*max*	3445,707	747,230	70,011	10,301	2,872
		$\overline{e}$	**647,289**	**156,114**	**11,965**	**0,659**	**-0,836**
		min	-47,988	-28,788	-28,786	-5,818	-1,665
		s	**930,933**	**215,734**	**20,560**	**3,174**	**0,540**
Dirichlet	10^{1}	*max*	209,803	21,159	2,770	1,413	1,282
		$\overline{e}$	**2,994**	**-6,564**	**-0,421**	**0,805**	**0,389**
		min	-50,748	-44,030	-5,102	-0,391	-0,226
		s	**32,650**	**12,517**	**1,474**	**0,274**	**0,196**
	10^{12}	*max*	178,462	4,407	1,154	1,536	0,701
		$\overline{e}$	**3,697**	**2,846**	**1,072**	**1,504**	**0,675**
		min	-18,911	-6,465	0,581	1,314	0,656
		s	**30,799**	**1,351**	**0,073**	**0,026**	**0,009**

Table 1: Statistical summary of deviations plotted in Figures 5 to 7. Mean $\overline{e}$, standard deviation **s** and extreme deviations (*max* and *min*) attained [%].

One discusses first the composite stress results in Figures 5 and 6. Here one makes the distinction between the stress deviations measured in the material 1 which is stronger (see the black dots plotted) and material 2 which is weaker (see the empty dots plotted). As a result of separating things this way one concludes that the higher deviations are typically identified with the elements where stress values are lower (material 2), see Figures 5a,b,c and 6a,b,c. In regions where stress values are higher or critical (material 1) the stress values from ANSYS and POSTMAT compare much better which is, in practice, a significant result because ulti-

mately one would compare the maximum value of POSTMAT with the material Yield Stress (σ_Y) for design feasibility. The relative importance of the composite phases in the magnitude of the deviations is far more noticed for low scale factors n, typically between 1 and 3. For $n \geq 4$ the deviations measured in materials 2 become eventually as big as the deviations seen in material 1 and typically both come to be below 2% which is here an excellent convergence result (see Figures 5e and 6e).

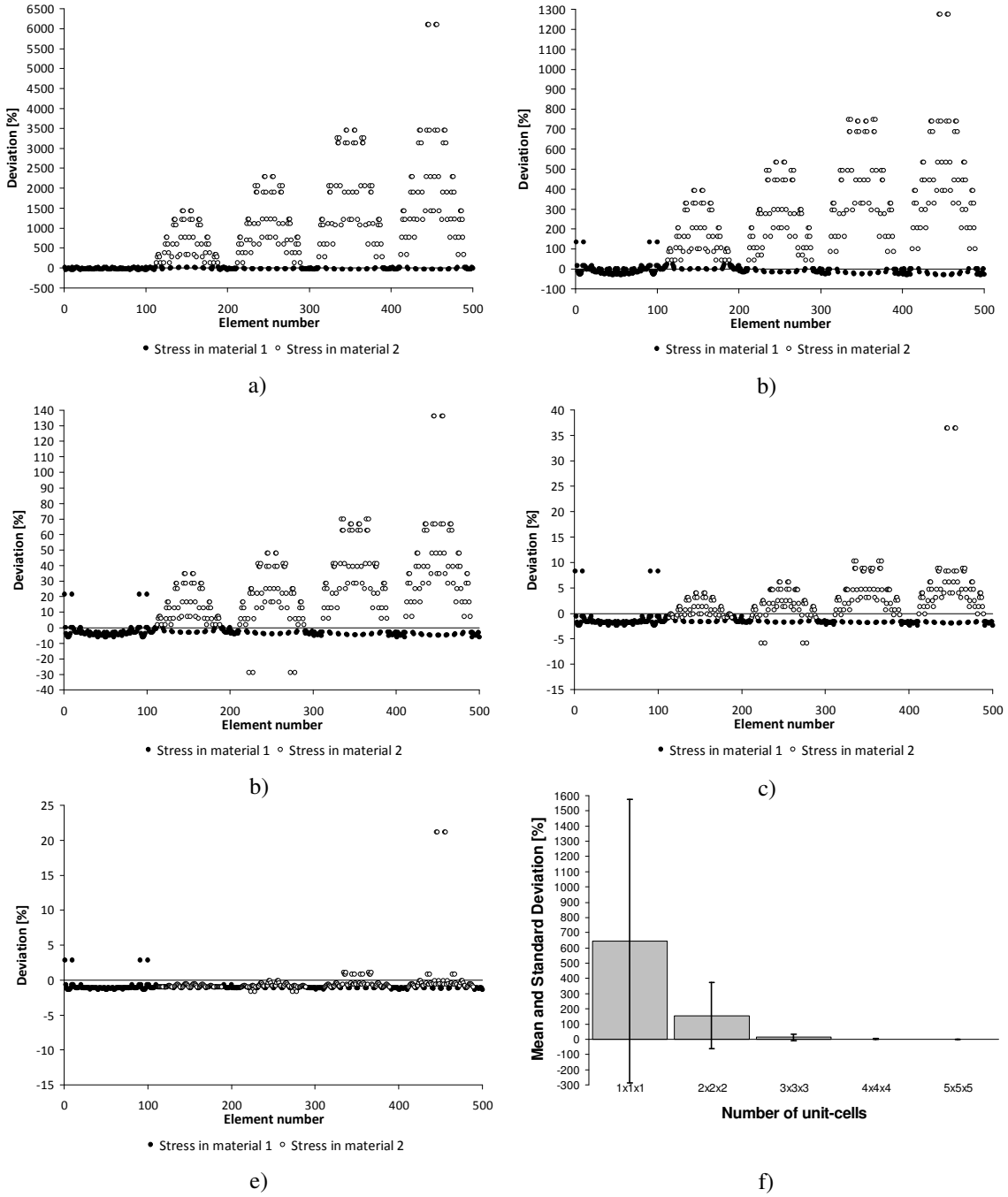


Figure 5: Scale-size effects analysis for **composite $E^{(1)}/E^{(2)} = 10^1$** with **Neumann B.C.** Deviation computed through Eq. (14) in all elements of the central unit-cell taken from different arrays: a) 1×1×1; b) 2×2×2; c) 3×3×3; d) 4×4×4; e) 5×5×5. f) Statistical summary in terms of mean and standard deviation.

However, despite higher *n*, one observes that deviations computed in only four elements (numbers 445, 446, 455, 456) are comparatively much higher, reaching 22%. These elements are located right in the middle of the unit-cell domain Y (in material 2) and their stress value (order of 10^2) is the minimum found in Y whether by POSTMAT or ANSYS (the order of the maximum stress found is 10^6). These deviations are classified here as outliers and the statistical analysis carried out in Table 1 excludes them.

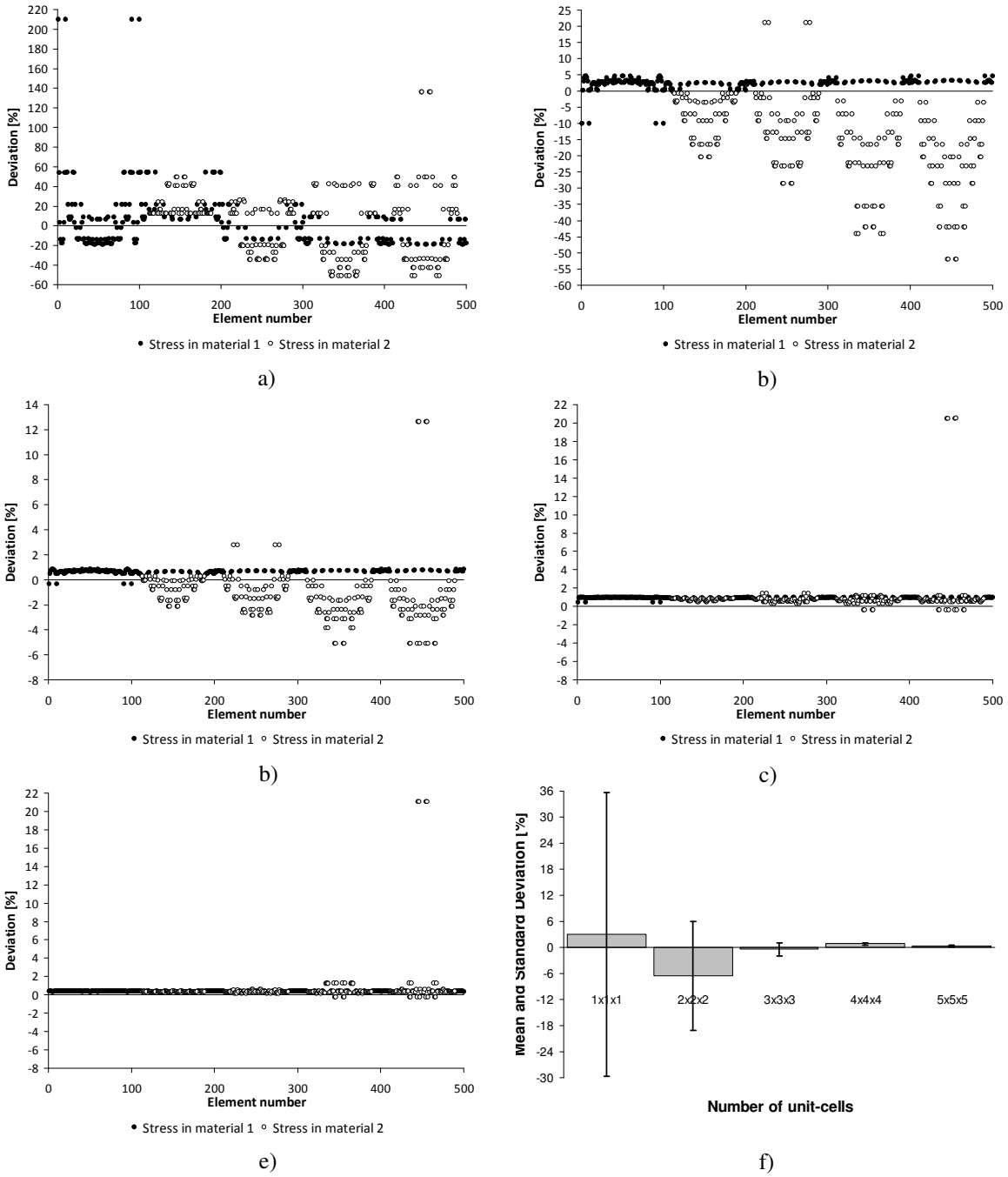


Figure 6: Scale-size effects analysis for **composite $E^{(1)}/E^{(2)} = 10^1$** with **Dirichlet B.C.** Deviation computed through Eq. (14) in all elements of the central unit-cell taken from different arrays: a) 1×1×1; b) 2×2×2; c) 3×3×3; d) 4×4×4; e) 5×5×5. f) Statistical summary in terms of mean and standard deviation.

Considering now the cellular material case, one notices that the stresses are basically zero (e.g. 10^{-6}) in elements representing the void phase that's why one decides overlooking them when calculating deviations and doing statistics (Figure 7 contains then comparatively lesser points). Due to this simplification one notices the standard deviation decreasing faster than the other cases. The *min*, *max* and mean values in Table 1 are all below 1% for $n = 5$ which is impressive for such a low scale factor.

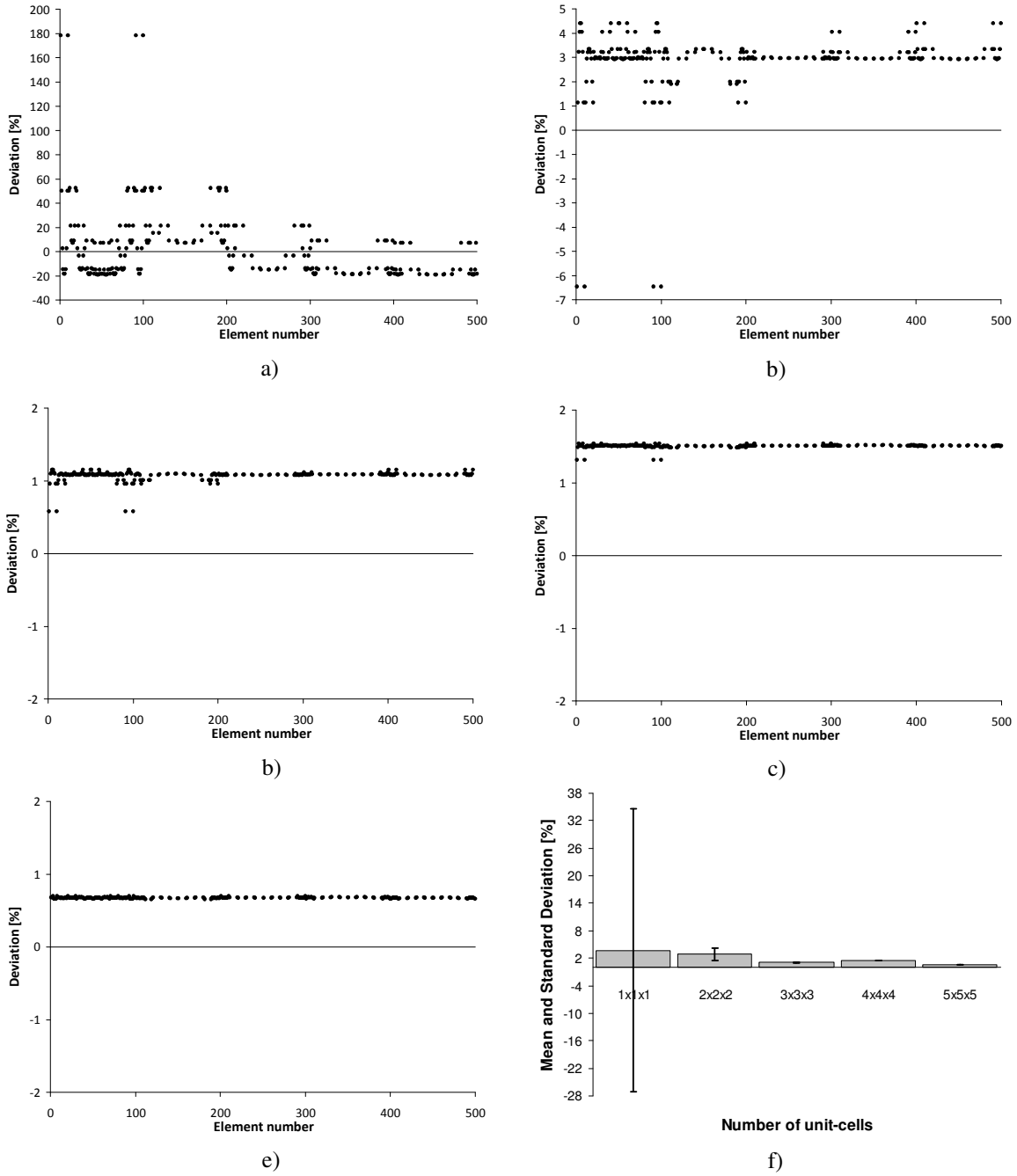


Figure 7: Scale-size effects analysis for **cellular material $\mathbf{E^{(1)}/E^{(2)} = 10^{12}}$** with **Dirichlet B.C.** Deviation computed through Eq. (14) in all elements of the central unit-cell taken from different arrays: a) 1×1×1; b) 2×2×2; c) 3×3×3; d) 4×4×4; e) 5×5×5. f) Statistical summary in terms of mean and standard deviation.

In general, one perceives a fair convergence to the homogenizations predictions. Since the shifted design is used in arrays with even number of unit-cells, it is more fair to compare results among n=1,3,5 and then between n=2,4 because the material distribution in the boundary, $\partial\Psi$, where loads are applied, differs depending on whether the design is shifted or not (see Figure 3). The standard deviation measure gives probably more insight in this analysis than the mean value. The mean may go up and down (changing sign as well) as n increases because ANSYS may target POSTMAT either from above or below. The resulting trend for the mean is not so consistent when compared to the standard deviation trend which always decrease with increasing n. Anyway, for n=5 the mean deviation is below 1% in all cases. The maximum and minimum deviation values are also quite good as a result of excluding outliers as aforementioned.

6 CONCLUSIONS

- Stress-based topology optimization is quite appealing for engineering practice due to the fact that an optimal design must be a feasible one too.
- The inverse homogenization method is straightforward to generate unit-cell designs for periodic composites. However the assumptions of ideal periodicity as well as dimensionless unit-cell have to be checked in the context of real composites which motivates the scale-size effect analysis presented in this work.
- As an outcome the present study indicates that it is sufficient to have a low scale factor (n=5) to replace the non-homogeneous composite by the equivalent homogeneous material with the stress field computed by homogenization.
- A single unit-cell topology design was investigated in this work and a coarse finite element mesh was used to discretize the stress field that is in general highly non-linear. These simplifications may motivate further detailed analyses on scale-size effects.

REFERENCES

[1] J.D. Deaton. R.V. Grandhi. A survey and multidisciplinary continuum topology optimization: post 2000. *Struct Multidisc Optim*. **49**. 1-38. 2014.

[2] M.P. Bendsøe. O. Sigmund. *Topology Optimization: Theory, Methods and Applications.* Springer. 2004.

[3] P. Duysinx. M.P. Bendsøe. Topology optimization of continuum structures with local stress constraints. *Int Journal for Numerical Methods in Engineering*. **43**. 1453-1478. 1998.

[4] M. Bruggi. On an alternative approach to stress constraints relaxation in topology optimization. *Struct Multidisc Optim*. **36**. 125-141. 2008.

[5] R. Lipton. M. Stuebner. Optimization of composite structures subject to local stress constraints. *Computer Methods in Applied Mechanics and Engineering*. **196**. 66-75. 2006.

[6] B. Hassani. E. Hinton. A review of homogenization and topology optimization I - homogenization theory for media with periodic structure. *Computers and Structures*. **69**. 707-717. 1998.

[7] P.W. Chung. K.K. Tamma. R.R. Namburu. Asymptotic expansion homogenization for heterogeneous media: computational issues and applications. *Composites Part A*. **32**. 1291-1301. 2001

[8] O. Sigmund. Materials with prescribed constitutive parameters: An inverse homogenization problem. *International Journal of Solids and Structures*. **31**(17). 2313-2329. 1994.

[9] P.G. Coelho. L.D. Amiano. J.M. Guedes. H.C. Rodrigues. Scale-size effects analysis of optimal periodic material microstructures designed by the inverse homogenization method. *Computers and Structures*. Doi:10.1016/j.compstruc.2015.10.001. In Press.

[10] P.G. Coelho. S.J. Hollister. C.L. Flanagan. P.R. Fernandes. Bioresorbable scaffolds for bone tissue engineering: optimal design, fabrication, mechanical testing and scale-size effects analysis. *Med Eng Phys*. 37(3). 287-296. 2015.

[11] O. Sigmund. On the optimality of bone microstructure. P. Pedersen, M. Bendse eds. *IUTAM Symposium on Synthesis in Bio Solid Mechanics*. Kluwer. 221–234. 1999.

[12] R.T.L. Ferreira. H.C. Rodrigues. J.M. Guedes. J.A. Hernandes. Hierarchical optimization of laminated fiber reinforced composites. *Composite Structures*. **107**. 246-259. 2014.

[13] J.M. Guedes. N. Kikuchi. Preprocessing and postprocessing for materials based on the homogenization method with adaptive finite element method. *Comput Meth Appl Mech Eng*. **83**.143-198. 1990.

[14] T.I. Zohdi, P. Wriggers. *Introduction to computational micromechanics*. Springer. 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds.)
Crete Island, Greece, 5-10 June 2016

A NUMERICAL STUDY FOR THE BUCKLING CAPACITY OF WIND TURBINE BLADES: GEOMETRY, LOADING AND MATERIAL INFLUENCE

Efstathios E. Theotokoglou[1], **Georgios Balokas**[1, 2] **and Evgenia K. Savvaki**[1]

[1] Department of Mechanics, Laboratory of Testing and Materials
School of Applied Mathematical and Physical Sciences, National Technical University of Athens,
Athens, GR-15773 Greece
stathis@central.ntua.gr , euginias@gmail.com

[2] Structure Development Department, ELAN-AUSY GmbH
Harburger Schloßstr. 24, Hamburg, Germany
Georgios.Balokas@elan-ausy.com

Keywords: Composite materials, Wind turbine blades, Fiber-reinforced polymers, Finite element analysis, Nonlinear buckling analysis, Eigenvalue analysis

Abstract. *In this paper the composite load-carrying support structure of wind turbine blades is investigated. The buckling behavior of wind turbine blades is studied numerically by using the finite element method, as experimental experience shows that local buckling is a major failure mode that dominantly influences the total collapse of the blade. Significant advantages are derived from the combination of different fiber-reinforced polymers in hybrid material structures, but also from kevlar-fiber blades.*

1 INTRODUCTION

In recent years, wind energy has noted the highest growth rate in comparison with other renewable energy sources [1]. This is mainly due to the increasing energy needs and the simultaneous depletion of natural resources, but also to the lack of environmental hazards and the practically unlimited resources. Consequently, the exploitation of wind energy systems, e.g. wind turbines, forms an industry domain that gathers strong interest in both research and construction activities. The need to optimize their performance results in higher power wind turbines with significantly longer blades and innovative features in terms of geometry and materials.

The blades are perhaps the most critical structural members of the wind turbine, since the increasing diameter of the rotor brings many challenges to the surface, regarding the design and construction. The wind turbine blade (WTB) is essentially a cantilever beam mounted on a rotating hub with key design criteria the high stiffness, the low weight and the aerodynamic shape. These features are covered by fiber-reinforced composite materials. Such fiber-reinforced polymer materials are implemented in laminates and sandwich structures for the outer aerodynamic shell, but also for the internal load-carrying structure, ensuring the required strength and stiffness of the blade.

In this paper the support structure is investigated for the case of a hollow one-piece construction cross-section (box girder). Through parametric analyses an attempt was made to investigate the buckling behavior of the load-carrying box girder due to flap-wise bending, with respect to the following parameters: the geometry, the loading imposition and the material properties. This study offers a clear perspective about the buckling capacity and its sensitivity on the parameters mentioned above, but also about the post-buckling behavior of the models. Furthermore, this investigation leads to useful conclusions for the material design optimization of the load-carrying box girder, as significant advantages are derived from the combination of different fiber-reinforced polymers in hybrid material structures, but also from kevlar-fiber blades.

2 NUMERICAL MODELING AND ANALYSIS

2.1 Geometrical model

There are currently two major types of configuration for the internal geometry of the blade, which differ in terms of design and connection to the outer shell. The first case is the so-called two-piece construction and consists of two distinct vertical stiffness joints, known as shear-webs, which extend along the blade and provide the required internal support. The second case is the so-called one-piece construction and is extensively studied in the present paper. Here, the internal support of the blade is provided by a single hollow section structure, which extends almost to the entire length of the blade and is usually called box girder. The tension and the compression flange of this structure are welded to the upwind and downwind outer shell respectively. The parts welded to the shell are called spar-caps and they are linked to the shear-webs, which are placed vertically in the cross-section, at 15% and 50% of the length of the chord c, as measured from the leading edge of the blade (Fig.1).

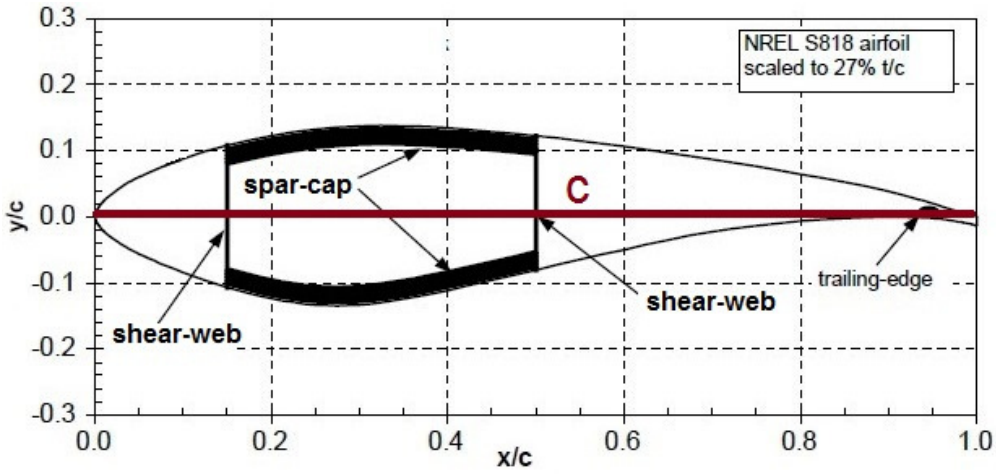


Figure 1: Cross-section of the blade for the one-piece construction case [2]

The box girder provides the required strength and stiffness of the outer shell of the blade, both locally and globally. For this reason, the box girder was chosen to be studied separately, since the dominant deformation mode is almost identical to the one of the blade, while the increased slenderness of the panels makes it sensitive to local buckling phenomena. Experimental data show that local (or shear) buckling is a major failure mode that significantly influences the total collapse of the blade [3-5].

The present study models the internal load-carrying box girder of a horizontal axis wind turbine, with nominal power 1MW, 64.14m tower and 63.04m rotor diameter, for the case of a single box-like, hollow section internal support. The total length of the blade, (from the rotational axis to the blade tip) equals to 30m. The blades are divided into three sections, depending on the deformation levels due to the stress loads [2]. In our case, the root segment is 2.10m long (equal to 7% of the blade length), the transition segment 5.40m (equal to 18% of the blade length) and the main box girder segment 22.50m, inside of which the internal support mechanism is located (hatched area in Fig. 2).

For the numerical analysis, a finite element code was developed, using the ANSYS 14.5 software [6]. The analysis parameters chosen were: a) the geometry of the cross-section, b) the simulation of the flap-wise loading and c) the fiber-reinforced composite material properties. In all cases both linear and nonlinear buckling analyses were performed. From the latter, the respective equilibrium paths (load-displacement curves) were extracted.

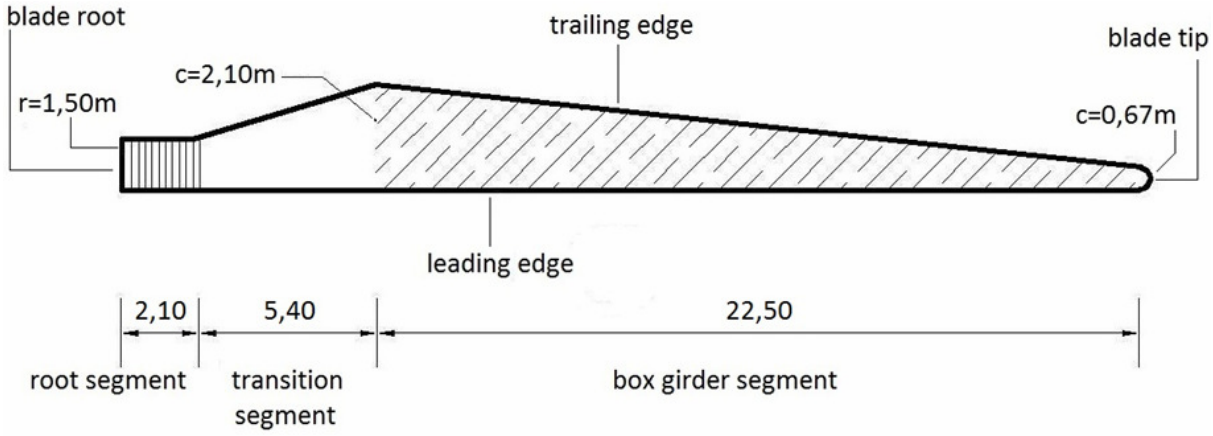


Figure 2: Division of the blade [5].

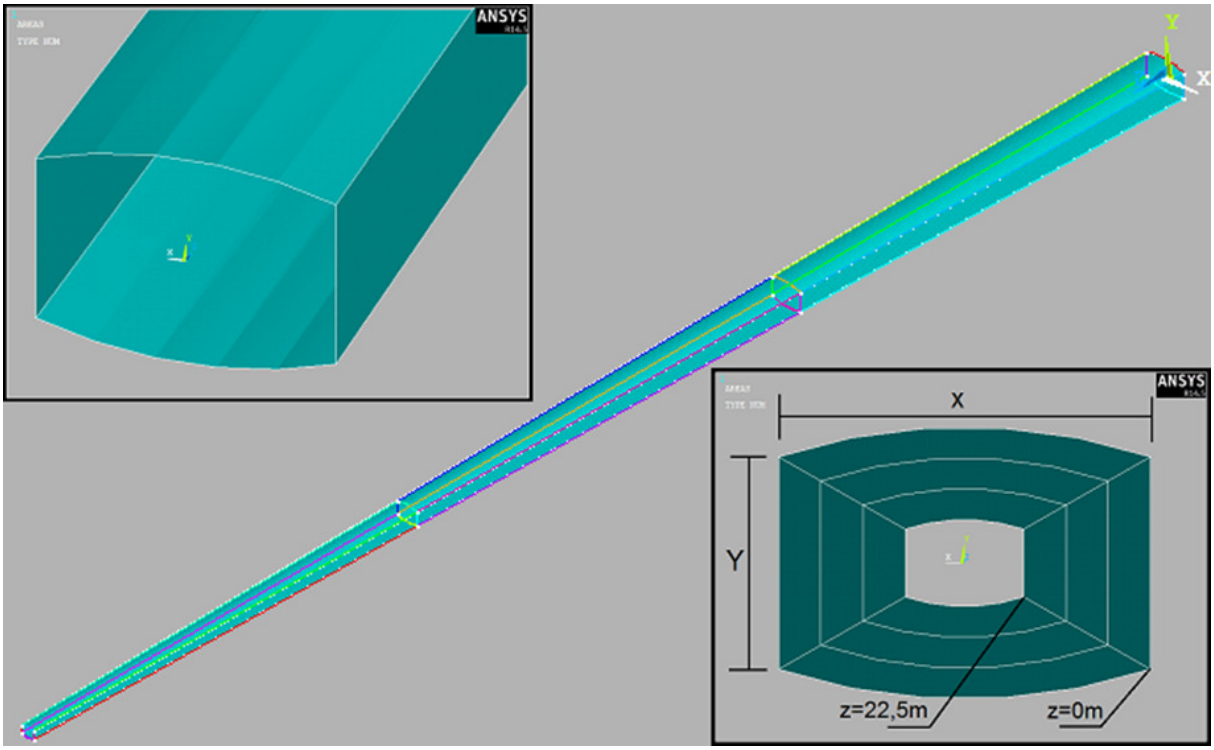


Figure 3: Geometrical model.

The dimensions and the design details of the examined blade correspond to the S818 airfoil model [2]. The coordinates of 24 points (6 for each of the 4 cross-sections) were calculated at the beginning (location z=0.0m), the end (location z=22.5m) and two intermediate positions of the box girder. Then the finite element model was constructed by presuming linear longitudinal tapering (Fig. 3). During the designing process, it is assumed double symmetry conditions for the cross section with respect to the axes X and Y, which in fact is not entirely true, as it can be shown in Figure 1. However it is acceptable, since the same assumption is made in relevant studies of the literature [3].

Shell elements were used (shell281 in ANSYS [6]) for the simulation, which are the most common type of finite elements in WTB analysis. A mesh convergence study was conducted, in order to determine the size of the finite elements for which satisfactory accuracy is achieved within reasonable computing time. So for the discretization of the models, shell elements with side length equal to 300mm were considered.

As it is obvious from Figure 3, the FE model was initially designed with the assumption of an angle formed in the transition area from the spar-cap to the shear-web. This assumption is adopted in similar studies [7]. However, it is not completely realistic since there is some curvature in the transition area. This happens in order to achieve a smooth transition to the thinner shear-web and to avoid high stress concentration. The technique applied in similar cases in composites is called “ply-drop analysis” and is implemented with gradual termination of some layers in the required region that leads to thickness reduction [3, 5, 8].

Ply-drop analysis cannot be applied efficiently in a macro-scale level, so in this study a small curvature was given as a simplistic alternative (Fig. 4a), at the initial corner of the spar-cap (Fig. 4b). This modification in the cross section design was used as an analysis parameter, in order to determine whether a small variation of the geometry of the model could affect the buckling load and the deformation of the panels.

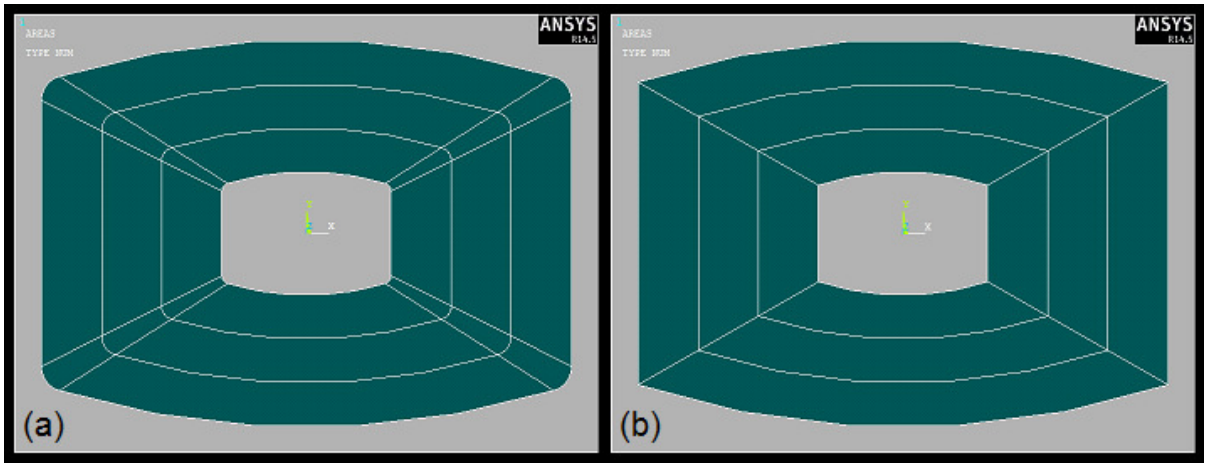


Figure 4: FEA model at the z=0.0m location: a) with curvature and b) without curvature at the transition area.

2.2 Boundary and loading conditions

There are two types of bending for the WTB: the edge-wise bending, which is vertical to one of the two edges of the blade (leading edge and trailing edge) and is caused by the gravity loads of the blade and the flap-wise bending, that is caused by the wind loads and is vertical to the large surface of the blade. Based on the global coordinate system shown in Figure 3, the edge-wise bending takes place around the Y-axis and the flap-wise around the X-axis. The blade also endures centrifugal forces due to the rotation of the rotor, although they are not significant and are usually neglected in the analysis.

In this study, four different loading simulations were examined in order to determine whether there is an influence in the critical buckling load and in the size and shape of the deformations. Moreover, in case of similar results we could reach to a conclusion regarding the most inexpensive performance.

All of our load alternatives cause flap-wise bending to the box girder, because it is the most critical condition and the one that usually leads to failure at local and/or global level [3, 9]. Specifically, the load was simulated with: a) uniform pressure, vertically spaced at the upper spar-cap (red lines matrix in Figure 5a), b) linear load along the model imposed in the middle of the upper spar-cap (Fig. 5b), c) concentrated load at the free end of the model acting on the middle of the spar-cap (Fig. 5c) and d) two concentrated loads of equal magnitude at locations z = 7.5m and z = 15m (Fig. 5d). All the loads are static, imposed by an incremental step-by-step process (Newton-Raphson method) and are applied in a way that the upper spar-cap is tensioned (upwind) and the lower spar-cap is compressed (downwind).

Regarding the boundary conditions, all four sides of the cross section at the z=0 location, are assumed fully fixed (restrained rotation and displacement of the axes X, Y and Z).

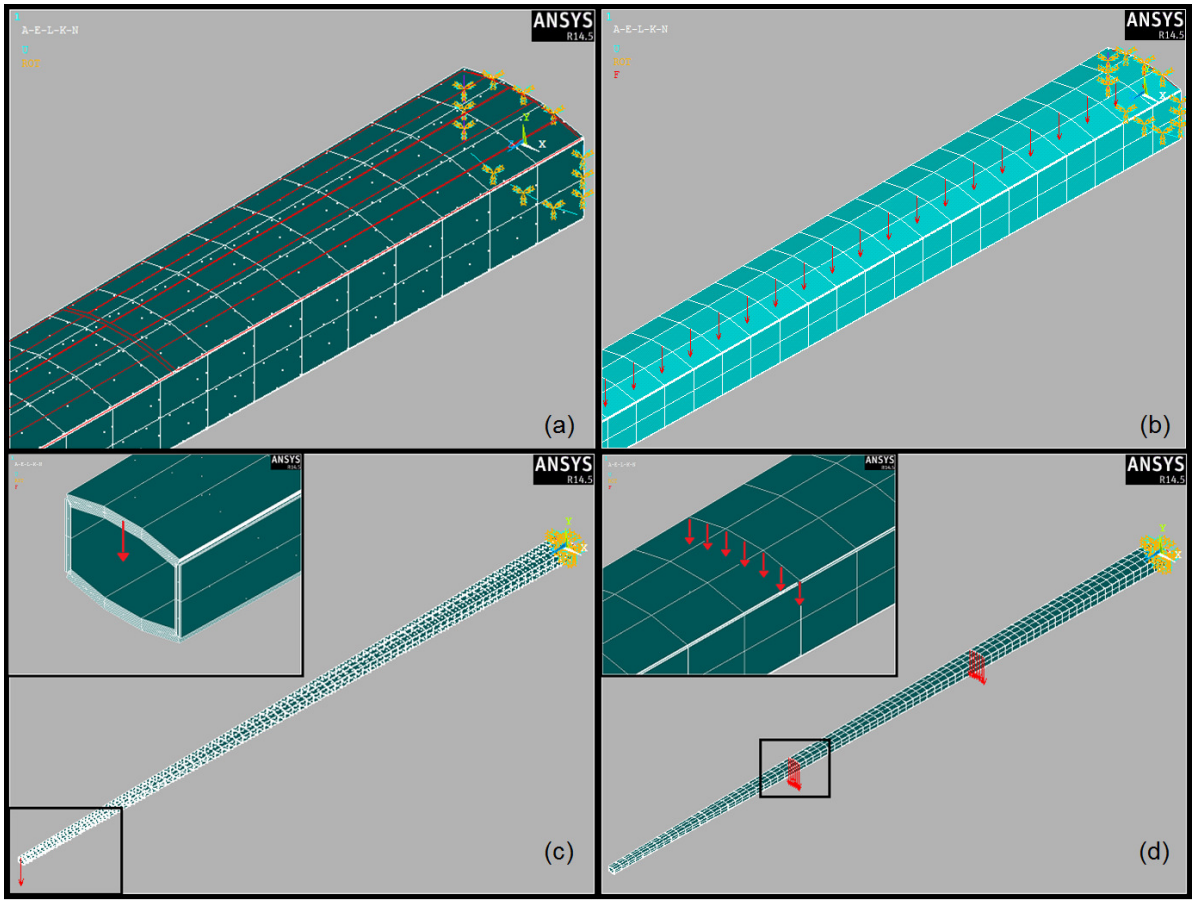


Figure 5: The four different load simulations.

2.3 Design and Material selection

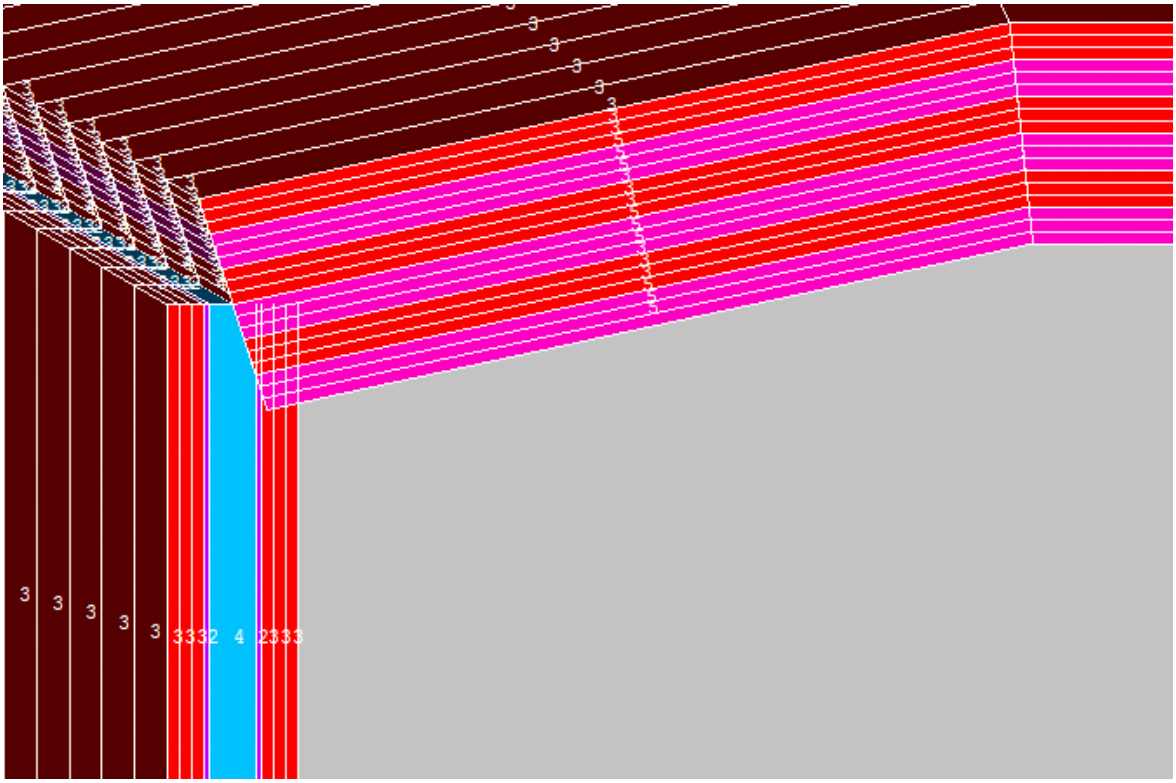
The selection of appropriate materials, which can optimally cover the increasing requirements of safety, efficiency and service life of a modern, high-power, wind turbine, is perhaps the most critical task and also a great challenge for the designers. The selected materials should have three basic properties: i) high material stiffness, to maintain the optimal aerodynamic shape of the blade, but also to prevent contact with the turbine tower and local buckling phenomena, ii) low density in order to minimize the gravity loads and iii) long fatigue life to reduce material degradation during the operation and to ensure at least a 20-year service life [10].

For the blade construction, polymers reinforced with glass fibers (glass fiber reinforced plastics-GFRP) or carbon fibers (carbon fiber reinforced plastics-CFRP) are commonly used. The high-strength fibers operate as reinforcement and are retained by load bearing mean, the matrix, which is usually made of epoxy, because of its excellent properties. Such materials are light and additionally, their fibers have a much higher strength-to-weight ratio and stiffness-to-weight in comparison with steel or wood, which were initially used for blade manufacturing.

The outer airfoil skins are sandwich structures while the internal load-carrying box girder combines laminate with sandwich layup. Specifically, spar-caps are consisted of alternating equal thickness layers of triaxial laminates (-45°/0°/45°), and unidirectional laminates. The unidirectional laminates are providing the required bending stiffness, while laminates with fibers at ±45° direction are providing torsional stiffness and buckling resistance for the surface under compression. Shear-webs are constructed using a sandwich-like material consisting

of triaxial composite laminate face sheets separated by a balsa wood core. The application of a core material increases locally the bending strength and buckling resistance. The layer order is shown in Figure 6, while the identification number of each material is provided in Table 1, along with the thickness of the corresponding layer.

The thickness of the spar-caps is constant along the blade length and equal to 21.6mm, while the thickness of the shear-webs is 17.69mm at the z=0m location and 12.92mm at the z=22.5m location (linear reduction as a function of the chord length).


3
3
3

Glass fibers are currently the most widely used fibers for blade construction. They are available in various types and different chemical compositions, but E-glass fibers are those primarily used for the blades, due to their mechanical performance and resistance to corrosion. However, limited information is available on the structural design process of blades (i.e. properties, layup and layer thickness), since the information remains confidential between the manufacturers. Therefore the material properties of the GFRP model, such as the thickness, sequence and orientation of individual layers were obtained from a previous study [5] and have derived from experimental results for the given fiber orientations and arrangement of materials used in another experimental study [11]. The properties of the GFRP as well as those of the core material are provided in Table 2, where E_{xx} is the axial Young's modulus, E_{yy} is the transverse Young's modulus, G_{xy} is the in-plane shear modulus, ν_{xy} is the Poisson's ratio, v_f is the fiber volume fraction, w_f is the fiber weight fraction, and ρ is the density.

Properties	Uni-axial (#5)	Tri-axial(#3)	Ran-dom (#2)	Balsa (#4)	Gel (#1)	Epoxy adhesive
E_{xx} (GPa)	31.00	24.20	9.65	2.07	3.44	2.76
E_{yy} (GPa)	7.59	8.97	9.65	2.07	3.44	2.76
G_{xy} (GPa)	3.52	4.97	3.86	0.14	1.38	1.10
ν_{xy}	0.31	0.39	0.30	0.22	0.30	0.30
u_f	0.40	0.40	-	-	-	-
w_f	0.61	0.61	-	-	-	-
ρ (g/cm^3)	1.70	1.70	1.67	0.14	1.23	1.15

Table 2: GFRP Material properties.

The size growth and the decreasing cost of carbon fibers have made them quite popular in the last 10-15 years. Carbon fibers present an exceptional combination of high stiffness, high strength and low density. Overall, they have much better performance than glass fibers, but yet they are of limited use because of their higher cost. In this study, the material properties of CFRP were obtained by applying formulas of composite materials theory [12] due to lack of experimental data similar to those used for the GFRP model. For our calculations we used the properties of AS4-D carbon fibers type and the same epoxy-based as the one used in the GFRP model. The properties of CFRP are shown in Table 3.

Properties	Uni-axial	Tri-axial	Random
E_{xx} (GPa)	146.00	65.00	62.47
E_{yy} (GPa)	18.53	22.50	62.47
G_{xy} (GPa)	9.41	13.46	24.19
ν_{xy}	0.27	0.29	0.29

Table 3: CFRP material properties.

In recent years aramid fibers are widely used, replacing metallic and inorganic fibers in composite structures in aerospace, maritime and automobile industry. Aramid fibers not only have better material properties than steel and glass fibers for the same weight level, but also maintain these properties at high temperatures, since they have high thermal insulation and fire resistance. Furthermore, aramid fiber reinforced polymers are proven to have much higher tensile strength and better resistance to fatigue from polymers with glass fibers [13]. Despite the positive characteristics that they present and their wide range of applications, the possibility of using them in WTBs manufacturing has not yet been examined. This paper investigates their buckling capacity, compared to glass and carbon fiber reinforced polymers, as they were presented above.

Aramid presents many valuable properties, depending on the treatment and the application in hand. One of its most popular derivatives is Kevlar, which is the trade name for aramid. There are many Kevlar fiber categories like Kevlar 29, Kevlar 49, Kevlar 68, Kevlar 119, Kevlar 129 and Kevlar 149. Kevlar 149 is one of the most recent categories in the Kevlar family. They have much higher Young's modulus than Kevlar fibers 29 and 49, while having approximately the same density and diameter, but a very low sensitivity to moisture (they are mainly used in aerospace). Because of these properties, Kevlar fibers 149 were chosen to be used for the fiber-reinforced polymer of the internal load carrying box girder. The mechanical properties of AFPR model were obtained by applying the same formulas used in the CFRP model and they are shown in Table 4.

Properties	Uni-axial	Tri-axial	Random
E_{xx} (GPa)	113	50.87	48.88
E_{yy} (GPa)	15.24	18.28	48.88
G_{xy} (GPa)	7.46	10.72	18.63
ν_{xy}	0.37	0.34	0.31

Table 4: AFRP material properties.

3 ANALYSIS RESULTS

3.1 Finite element analysis

We initially present the analysis results for the model considered: the E-GFRP model without curvature at the transition area between the spar-cap and the shear-web, where the load is simulated with uniform pressure vertically spaced at the upper spar-cap. We performed both eigenvalue and nonlinear buckling analysis for the baseline case as well as for all the other cases.

Eigenvalue analysis (linear buckling analysis) offers a quick estimate about the response of the model, while a rough value of the critical buckling load is calculated. The critical buckling load was estimated through linear analysis equal to 12.47KPa. However no information regarding the size of deformations or the post-buckling response of the model can be derived from the linear analysis. Moreover, linear analysis is insufficient for reaching precise conclusions because it is based on the assumption of small displacements and does not take into account the influence of nonlinear phenomena (large deformations, imperfections etc.). The most realistic behavior of the model can be obtained from the nonlinear buckling analysis. For the latter, we took into account the geometrical nonlinearity (where changes in geometry due

to large deformations significantly affect the relationship between the applied load and the displacement) and the results are presented below.

An early conclusion based on the deformed model (Fig. 7), is that the global deflection of the model is not considerably affected by local buckling. In a closer look, we can see that the local buckling area lies in the compressed spar-cap, near the root. Also, shear buckling occurs at the same position in the shear-webs. In Figure 8 the Brazier effect [8] of the blade is confirmed by the deformed shape, as the ovalization of the cross-section during buckling is apparent.

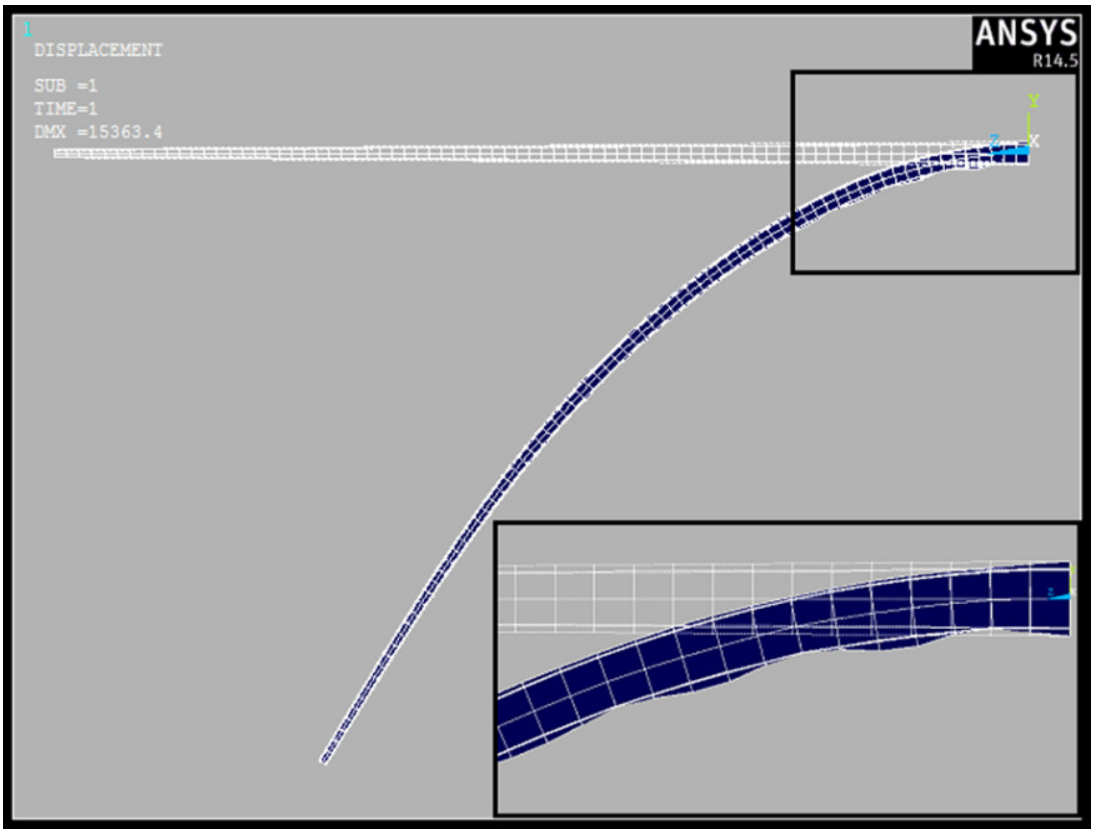


Figure 7: Global and local deformation of the model after the load imposition.

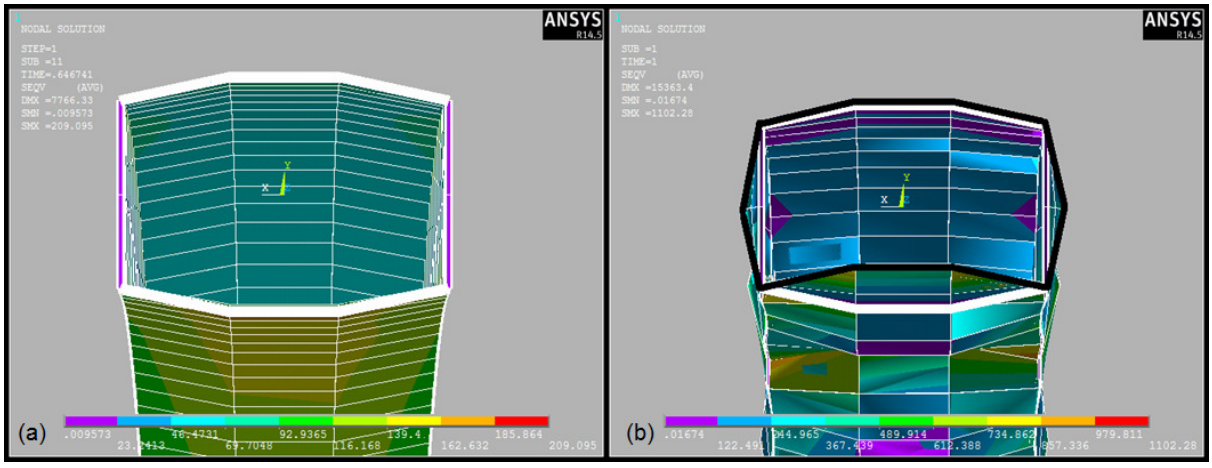


Figure 8: a) The model cross-section just before the local buckling and b) the cross-section ovalization.

The maximum local deformation (in the Y direction) of the compressed spar-cap is detected on D-spot in Fig.9 while the maximum local deformation of shear-web (in the X- direction) occurs on Z-spot in Fig. 10. The equilibrium paths are shown in Figs 11 and 12. The curves exhibit linear behavior (pre-buckling phase) up to 8.41KPa (critical buckling load), where the bifurcation point A is reached. After this point, the model has stable non-linear response, up to 13KPa (point C) (post-buckling phase).

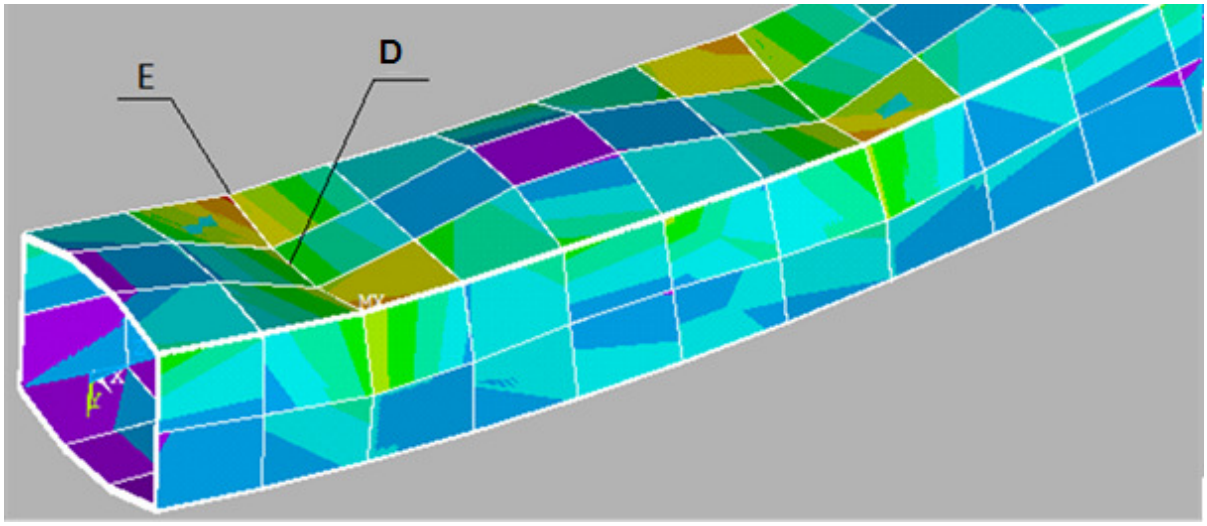


Figure 9: Spar-cap local buckling and von Mises stresses distribution

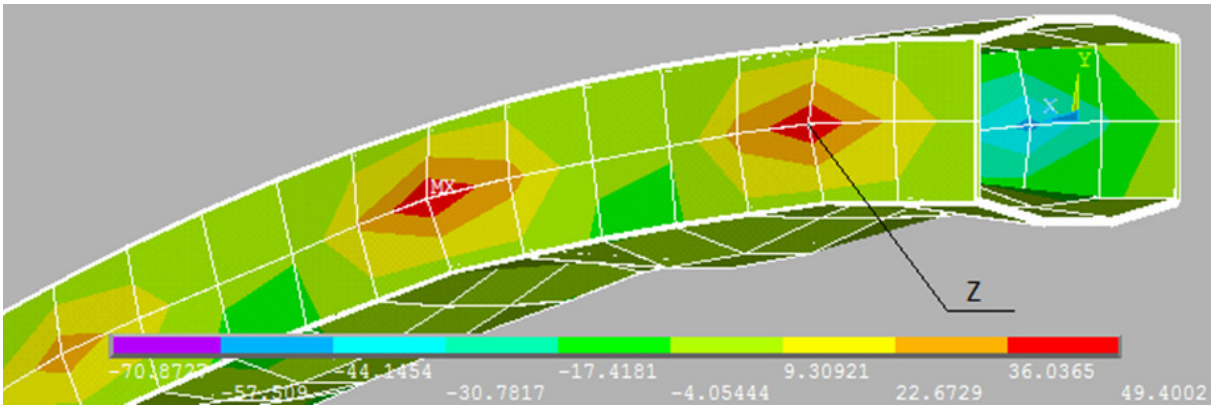


Figure 10: Displacements in the X-direction (in mm)

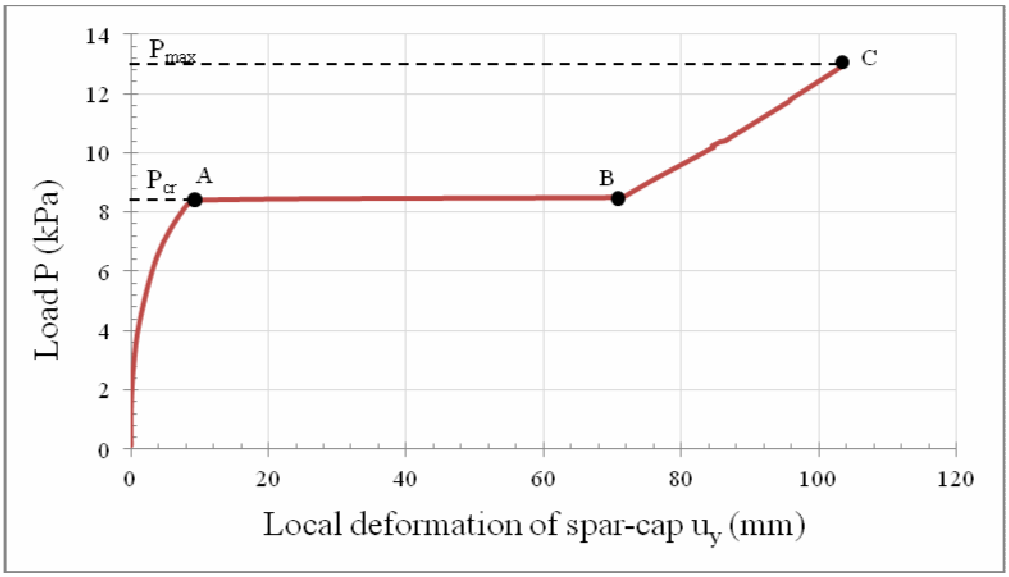


Figure 11: Load-displacement path (displacement at the Y-direction).

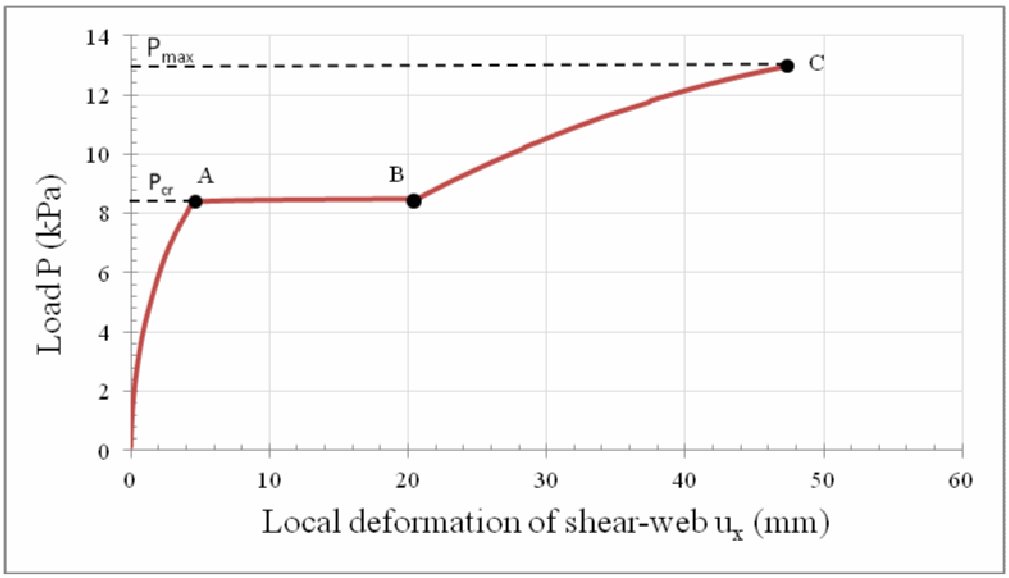


Figure 12: Load-displacement path (displacement at the X-direction).

3.2 Cross-section geometry parameter

We compare two models based on the GFRP material properties, with uniform pressure on the spar-cap and the geometrical discrepancies described in Section 2.1. From the load-displacement paths (Fig. 13), we can see that the two models have similar response both in the pre-buckling and the post-buckling phase, with equal stiffness and about the same critical buckling load. The deformed shape of both models is identical in terms of position and number of folds. However, the post-buckling path for the model with the curvature at the transition area has higher inclination and the size of deformations is smaller.

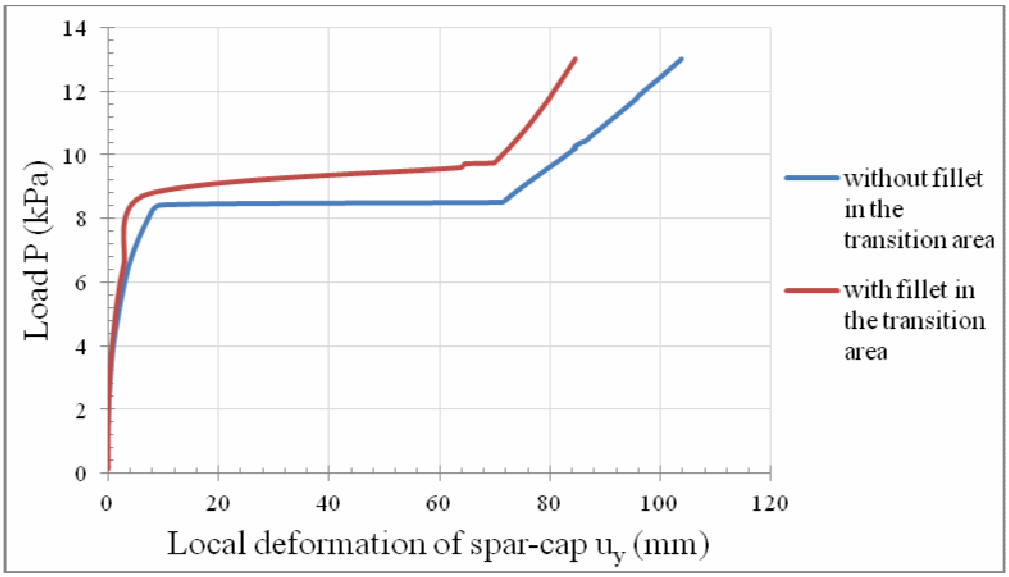


Figure 13: Load-displacement path (displacement at the Y-direction).

3.3 Loading imposition parameter

Based on specific examples of the literature [3, 9, 14], we examined four different ways of the loading simulation, which were described in Section 2.2. More specifically, except uniform pressure it is imposed: linear load in the middle of upper spar-cap of maximum value P_{max} = 6.32KN/m, concentrated load at the free end of the model of maximum value P_{max}=60KN and two concentrated loads of equal magnitude, with maximum value 63KN each. To allow comparison between pressure (KPa), linear load (KN/m) and concentrated loads (KN), we convert everything into equivalent bending moment at the support of the model for each load step.

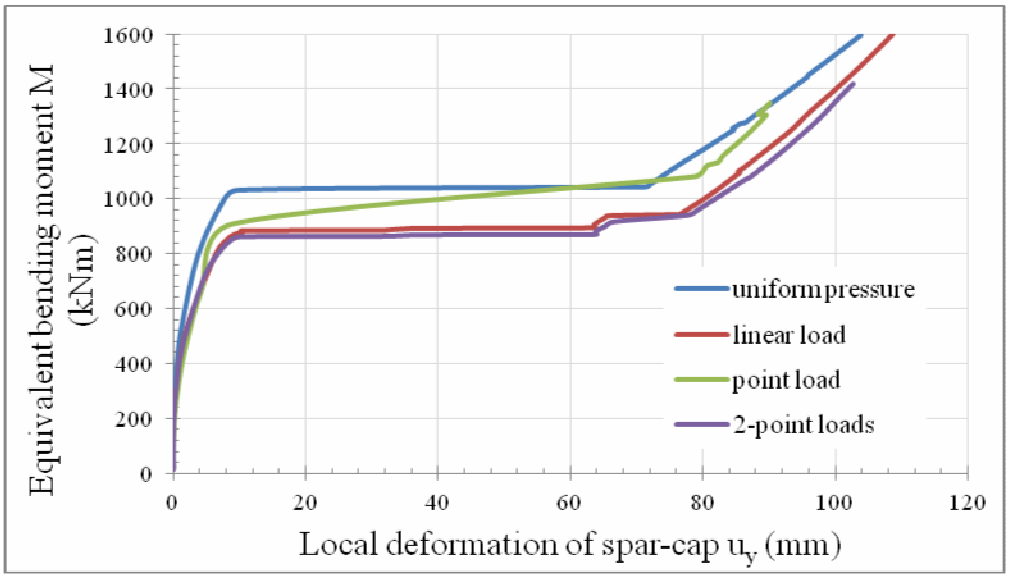


Figure 14: Load-displacement path (displacement at the Y-direction).

From Figure 14, it is clear that the models have similar behavior both in the pre-buckling and post-buckling phase. The pre-buckling paths are in general very close, while they are identical for the cases of the linear load and the two concentrated loads. However, there are

differences in the critical buckling load, where the maximum value emerged when we imposed uniform pressure. On the other hand, the critical load for the concentrated load at the free-end case is closer to the average value, while the size of the deformations is smaller. It is noted that there were no significant savings regarding computational efficiency in any of the examined cases.

Based on the above and the relatively small deviations observed, it can be deduced that the different loading simulations may indeed affect the distribution of the stresses across the model but they do not affect crucially its load carrying capacity and response. Therefore we can neither recommend nor exclude any of these load simulation methods based on our results.

3.4 Material properties parametric analysis

- **Comparison between glass, carbon and kevlar-fiber models**

The fiber-reinforced composite material properties are the last and most important analysis parameters in our study. The equilibrium paths (Fig. 15) indicate that the CFRP model has more than twice the buckling capacity of the GFRP model and also much higher stiffness (pre-buckling path with higher slope). This was expected, due to the remarkable properties of carbon. The AFRP model also has high critical buckling load and stiffness in comparison with the GFRP. Additionally, a significant restriction of deflections is observed prior the bifurcation point.

However, the AFRP model exhibits unstable post-buckling behavior, as it is indicated by the displacement values after the bifurcation point is reached. This instability can also be observed in the deformed shape of the model, as multiple deformation peaks emerge during loading history.

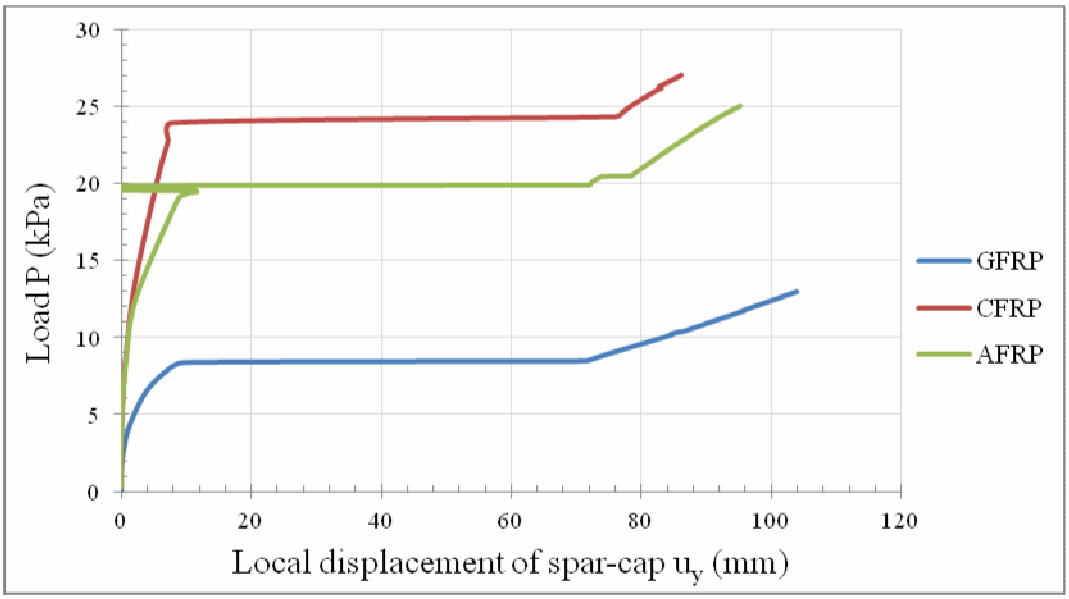


Figure 15: Load-displacement path (displacement at the Y-direction).

- **Comparison between hybrid models of GFRP and CFRP**

The need to further increase the strength-to-weight ratio and the stiffness-to-weight ratio of the blades has turned wind turbine industry towards hybrid structures. Hybrid models are likely to have higher strength and lower density compared to blades that are exclusively constructed by glass fibers. In this section, the following 3 cases are investigated: i) replacement

of the GFRP by CFRP material in the shear-webs (the GFRP_spar-cap/ CFRP_shear-web model), ii) replacement of the GFRP by CFRP material in spar-caps (the CFRP_spar-cap/ GFRP_shear-web model) and iii) replacement of the GFRP by CFRP material in the uni-axial layers in spar-caps (the CFRP_uni-axial in spar-cap model).

The nonlinear curves (Fig 16) indicate that the GFRP_spar-cap/ CFRP_shear-web has almost the same response as the GFRP model. The hybrid model has a slightly greater stiffness. The CFRP_spar-cap/ GFRP_shear-web model has exactly the same pre-buckling path as the CFRP model up to the critical buckling load. However, this model does not have an increasing post-buckling section (stiffening). The lack of sufficient post-buckling strength is a significant disadvantage. The most reasonable combination is that of the carbon fibers placed only in the uni-axial layers of the spar-caps. In this case, the hybrid model is found to have a sufficient combination of load carrying capacity and post-buckling strength, but also minimum use of the expensive carbon material. This conclusion is also reached in relevant studies of the literature [15, 16].

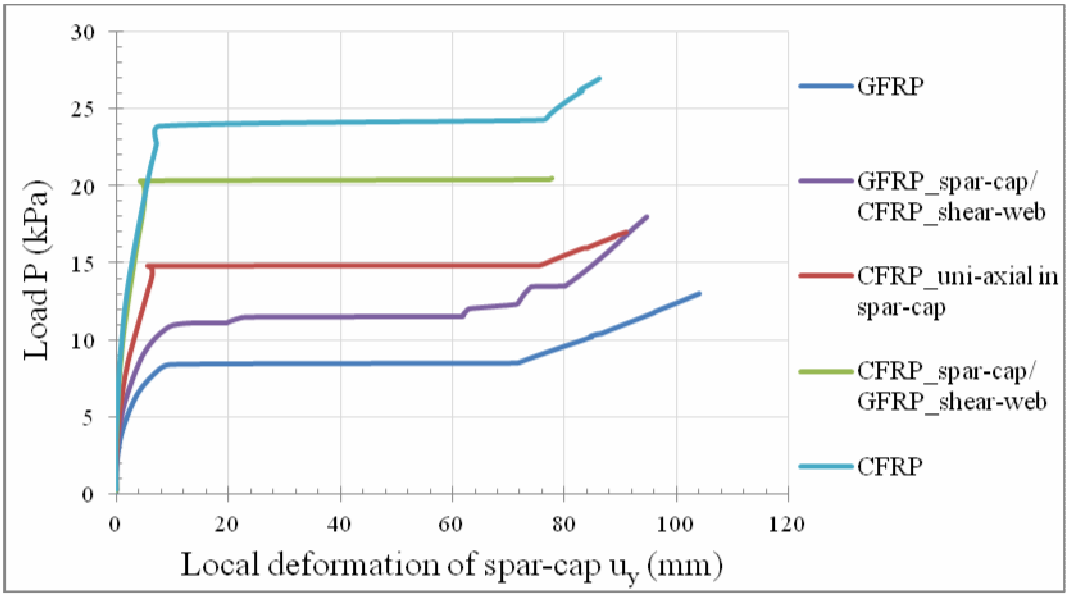


Figure 16: Load-displacement path (displacement at the Y-direction).

- **Comparison between hybrid models of CFRP and AFRP**

The response of hybrid models constructed by carbon and aramid fibers was also investigated. In particular, the following cases were studied: i) the replacement of the CFRP material by the AFRP material in spar-caps (the AFRP_spar-cap/ CFRP_shear-web model) and ii) the replacement of the CFRP by AFRP material in shear-webs (the CFRP_spar-cap/ AFRP_shear-web model).

The AFRP_spar-cap/ CFRP_shear-web model has the same curve with the AFRP model, without the unstable post-buckling behavior that the latter exhibits. This result is characterized as a positive contribution. The CFRP_spar-cap/ AFRP_shear-web model has exactly the same response, stiffness and stable post-buckling behavior as the CFRP model (their load-displacement paths are identical). Furthermore, the deformed models differ slightly (Fig. 18). Thus, the following useful conclusion results: when CFRP material is used in the spar-caps and AFRP material in shear-webs, a model can be created with the same load carrying capacity and buckling strength as the CFRP model, with significant savings in cost and weight (the Kevlar fibers 149 are cheaper and lighter than carbon fibers).

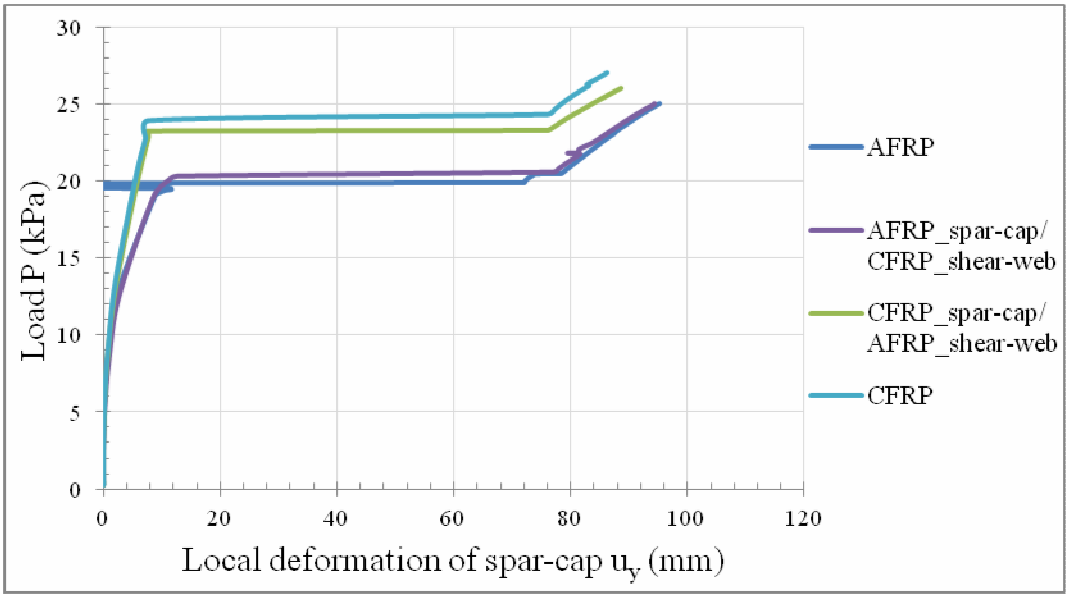


Figure 17: Load-displacement path (displacement at the Y-direction).

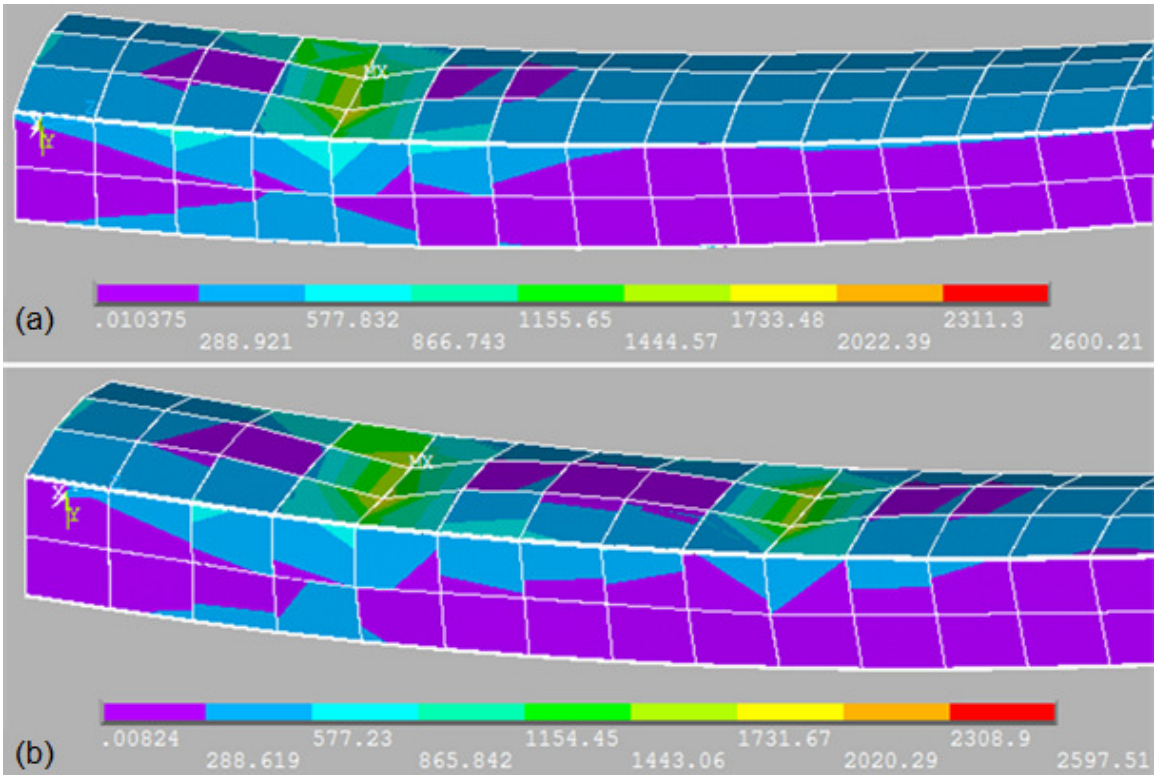


Figure 18: Local deformations and von Mises stress distribution at the critical buckling load for: a) the CFRP model and b) the CFRP_spar-cap/ AFRP_shear-web model

- **Comparison between hybrid models GFRP and AFRP**

Finally, the response of hybrid models, constructed by glass and aramid fibers, was investigated. The following cases were studied: i) the replacement of the AFRP material by the GFRP material in the spar-caps (the GFRP_spar-cap/ AFRP_shear-web model model), ii) the replacement of the GFRP by AFRP material in the spar-caps (the AFRP_spar-cap/

GFRP_shear-web model) and iii) replacement of the GFRP by AFRP material in the uni-axial layers of the spar-caps (the AFRP_uni-axial in spar-cap model).

The GFRP_spar-cap/ AFRP_shear-web model has slightly higher stiffness and critical buckling load than the GFRP model (Fig. 19). The AFRP_spar-cap/ GFRP_shear-web model exhibits sufficient load carrying capacity, without the unstable post-buckling behavior that was observed at the AFRP model analysis. The load-displacement curve for the AFRP_uni-axial in spar-cap model is in the middle of the rest, presenting a satisfatory combination of load carrying capacity and post-buckling strength. As a total conclusion from all the hybrid cases considered, it should be noted that the response and the stiffness of the hybrid models are defined by the spar-cap material.

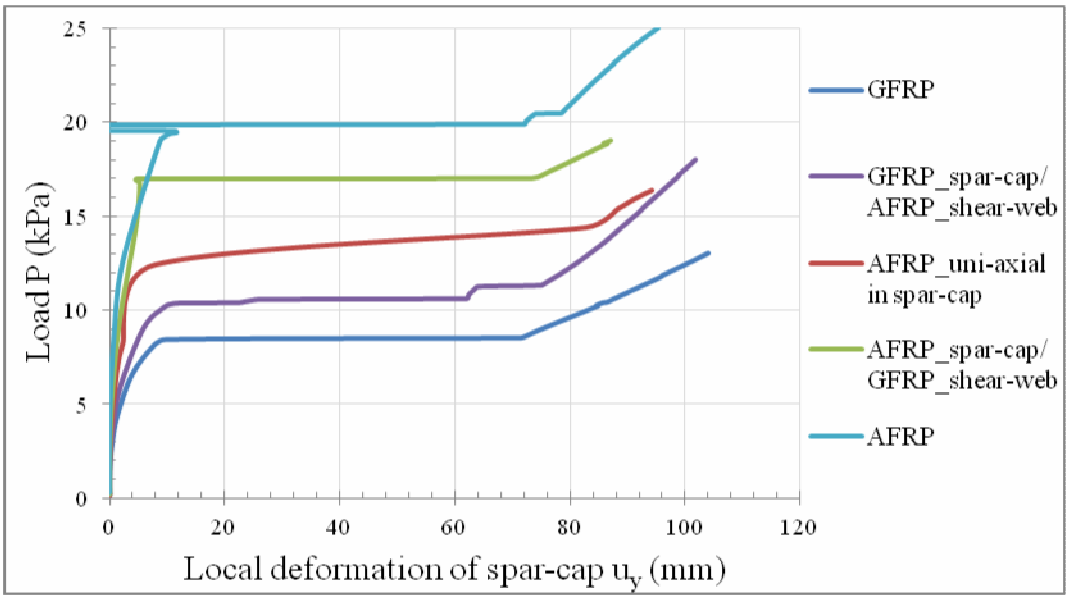


Figure 19: Load-displacement path (displacement at the Y-direction).

4 CONCLUSIONS

This paper studies the buckling capacity of the internal support of WTBs, by performing parametric analyses with respect to geometry, loading and material properties. The nonlinear buckling analysis results are consistent with a related study of the literature [3]. On the other hand, the eigenvalue analysis was proved extremely conservative, since the critical buckling load values are overestimated in all cases examined. The above confirms the necessity of the nonlinear analysis, despite the increased computational cost required. Nonlinear analysis leads to more realistic results and should always be performed in cases where changes in the geometry, due to large deformations, significantly affect the relationship between the applied load and the displacement.

The comparison between the GFRP, CFRP and AFRP models, showed that the CFRP model has greater stiffness and strength compared with the GFRP model and approximately double critical buckling load. Positive results are obtained from the use of Kevlar fibers, since the stiffness and the critical buckling load of the AFRP model is quite larger than those of the GFRP model. However, the unstable post-buckling behavior of the AFRP model is negatively evaluated. This instability is eliminated immediately, in cases where aramid fibers were used in conjunction with glass or carbon fibers in hybrid models. This is a strong prompt towards further research in the use of Kevlar fibers in wind turbine blades. Additionally, it is of great interest that if the CFRP material is used in the spar-caps and the AFRP material in shear-

webs (instead of CFRP material), a model can be produced with the same strength and buckling capacity as the CFRP material, but significantly cheaper and lighter, since the Kevlar 149 fibers have a lower cost and weight.

This study offers a clear perspective about the buckling capacity of the blade and its sensitivity when the material parameters are changed, but also about the post-buckling behavior and strength of the models. Nevertheless, there is space for improvement in the FEM simulation (e.g. more realistic configuration of the geometrical model, without the assumption of double symmetry at the X and Y axes, with more detailed ply-drop design in the transition area etc.). Further studies should also include material nonlinearity, by testing various models of plasticity for the composite fiber reinforced polymers. Moreover, the significant advantages that seemed to arise from the use of aramid fibers should be further tested, in order to examine their response in experimental models.

REFERENCES

[1] Renewable Energy Policy Network for the 21st Century, Renewables 2015 Global Status Report, Paris, France.

[2] D.A. Griffin, WindPact Turbine Design Scaling Studies Technical Area 1 - Composite Blades for 80- to 120-Meter Rotor, NREL, Washington, 2001.

[3] B.F. Sorensen, E. Jorgensen, C.P. Debel, F.M. Jensen, H.M. Jensen, T.K. Jacobsen and K.M. Halling, Improved design of large wind turbine blade of fiber composites based on studies of scale effects (Phase 1)-Summary Report, Risø-R-1390(EN), Risø National Laboratory, Roskilde, Denmark, 2004.

[4] O.T. Thomsen, Sandwich materials for wind turbine blades - present and future", Journal of Sandwich Structures and Materials, Vol.11, p.p. 7-27, 2009.

[5] E.E. Theotokoglou and G.A. Balokas, Computational Analysis and Material Selection in Cross-Section of a Composite Wind Turbine Blade, Journal of Reinforced Plastics and Composites, 34(2): 101-115, 2015.

[6] ANSYS Engineering Analysis System. User's manual, Swanson Analysis System Inc., Houston, USA, 2007.

[7] F.M. Jensen, B.G. Falzon, J. Ankersen, H. Stang, Structural Testing and Numerical Simulation of a 34m Composite Wind Turbine Blade, Composite Structures, 76, 52-61, 2006.

[8] E.E. Theotokoglou and G.A. Balokas, A Micro-Scale Structural Response Comparison Between GFRP and CFRP Wind Turbine Blades, Proceedings of the 8th GRACM International Congress on Computational Mechanics, Volos, Greece, July 12-15 2015.

[9] N. Gaudern. and D. Symons, Comparison of Theoretical and Numerical Buckling Loads for Wind Turbine Blade Panels, Wing Engineering, Volume 34, No. 2, PP 193-206, 2010.

[10] T. Burton, Sharpe D., Jenkins N. and Bossanyi E., "Wind Energy Handbook", Wiley, USA, 1991.

[11] M. Grujicic, G. Arakere, E. Subramanian, V. Sellappan, A. Vallejo, M. Ozen, Structural-Response Analysis, Fatigue-Life Prediction, and Material Selection for 1MW

Horizontal-Axis Wind Turbine Blades, Journal of Materials Engineering and Performance, 19: 790-801, 2009.

[12] J.E. Zimmer, J.R. Cost, Determination of the Elastic Constants of a Unidirectional Fiber Composite using Ultrasonic Velocity Measurements, J.Acoust. Soc. Am. 47, 795, 1970.

[13] C. Praveen Shaju, T. Manikandan, S. Sai Balaji, Experimental Study on Environmental Exposure of Kevlar Epoxy Composites, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 10, October 2013.

[14] X. Chen, W. Zhao, X.L. Zhao and J.Z. Xu, Failure Test and Finite Element Simulation of a Large Wind Turbine Composite Blade under Static Loading, Energies 2014, 7, 2274-2297.

[15] Ashwill, Materials and Innovations for Large Blade Structures: Research Opportunities in Wind Energy Technology, 50th AIAA Structures, Structural Dynamics & Materials Conference Palm Springs, May, 2009.

[16] K. Cox, A. Echtermeyer, Structural design and analysis of a 10MW wind turbine blade, Energy Procedia 24: 194-201, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A SEMI-IMPLICIT MICROPLAR DISCRETE-TO-CONTINUUM METHOD FOR GRANULAR MATERIALS

Kun Wang[1], **WaiChing Sun**[1]

[1]Department of Civil Engineering and Engineering Mechanics, Columbia University
614 SW Mudd, Mail Code: 4709, New York, NY 10027
e-mail: {kw2534,wsun}@columbia.edu

Keywords: Micropolar Continua, Discrete-to-Continuum, Granular Matters, Length Scale, Strain Localization.

Abstract. *A micropolar discrete-continuum coupling model is proposed to link the collectively particulate mechanical simulations at high-order representative elementary volume to field-scale boundary value problems. By incorporating high-order kinematics to the homogenization procedure, contact moment and force exerted on grain contacts are homogenized into a non-symmetric Cauchy stress and higher-order couple stress. These stress measures in return become the constitutive updates for the macroscopic finite element model for micropolar continua. Unlike the non-lcoal weighted averaging models in which the intrinsic length scale must be a prior knowledge to compute the nonlocal damage or strain measures, the proposed model introduces the physical length scale directly through the higher-order kinematics. As a result, there is no need to tune or adjust the intrinsic length scale. Furthermore, since constitutive updates are provided directly from micro-structures, there is also no need to calibrate any high-order material parameters that are difficult to infer from experiments. These salient features are demonstrated by numerical examples. The classical result from Mindlin is used as a benchmark to verify the proposed model.*

1 INTRODUCTION

A granular material is a conglomeration of interacting solid particles. Collectively, these particles may store elastic strain energy while they are bonded or in contact with each other [1, 2], and dissipate energy due to grain sliding, rotation and fracture or fragmentation. The applications and handling of granular materials are central to numerous industries, such as geotechnical engineering, pharmaceutical and food processing. When the particles of the granular assembly are neither bonded nor subjected to confining pressure, granular flow may occur [3]. On the other hand, the collective macroscopic responses of these interacting particles may exhibit traits that are similar to those of a solid continuum [4, 5, 6, 1, 7].

In the past decades, theoretical and computational models designed to model granular materials in solid-like state have achieved great success. In particular, the critical state theory has been widely used as a tool to predict granular materials in both dense and loose states. Nevertheless, the prediction of constitutive responses of granular materials after the onset of strain localization remains a challenging task. One ongoing problems is that numerical methods, such as finite element or finite volume method requires the usage of macroscopic elasto-plastic constitutive law to approximate nonlinear responses with incremental updates. Without proper the so-called regularization limiter, the incremental numerical solution may exhibit spurious dependence on the mesh size when material bifurcation described in Rudnicki and Rice [8] occurs. In additional to introduce rate dependence in the constitutive responses, one possible remedy is to incorporate higher-order kinematics into the constitutive laws [9, 10, 11]. One example of the higher-order theory is the Cosserat theory originated by the Cosserat brothers in their treatise. [12]. The Cosserat theory has been re-discovered and studied by researchers that model granular continua with micropolar kinematics. A micropolar continuum is different than the classical Cauchy-Boltzmann continuum in the sense that material points are associated with microstructures that have orientations characterized by micro-rotation [12, 13]. This additional degree of freedom allows characteristic length or physical length scale to be introduced into the micropolar constitutive law and hence can be used as a mean to resolve the mesh sensitivity issue and captured the material size effects [14, 15, 16].

Nevertheless, developing phenomenological constitutive laws for micropolar continua is not a trivial task. This difficulty is not only due to the increased complexity of the higher-order constitutive laws, but also the extra burden to identify material parameters corresponding to the micropolar effect. While it is true that the pathological mesh dependence at the post-bifurcation regime can be resolved by explicitly modeling the interactions among grains with discrete element simulations, such an approach is not feasible for field-scale problem where large amount of particles are involved.

The purpose of this study is to propose a new multi-scale model that combines the strengths of the detailed grain-scale discrete simulations and the efficient high-order continuum-based finite element methods. The offspring discrete-continuum model able to (1) establish information exchange paths that prescribe higher order deformation (curvature) on DEM representative volume elements (RVEs) and extract higher order stress (couple stress) from RVEs, (2) naturally incorporate material length scale (related to particle size) into the multiscale model and thus improves the performance on capturing the material size effect in poly-disperse granular materials, (3) provide the physical underpinning from DEM simulations, and (4) resolve the problems associated with the mesh sensitivity when strain localization occurs.

As for notations and symbols, bold-faced letters denote tensors; the symbol '$\cdot$' denotes a single contraction of adjacent indices of two tensors (e.g. $\boldsymbol{a} \cdot \boldsymbol{b} = a_i b_i$ or $\boldsymbol{c} \cdot \boldsymbol{d} = c_{ij} d_{jk}$); the

symbol ':' denotes a double contraction of adjacent indices of tensor of rank two or higher (e.g. $\boldsymbol{C} : \boldsymbol{\epsilon}^e = C_{ijkl}\epsilon^e_{kl}$); the symbol '$\otimes$' denotes a juxtaposition of two vectors (e.g. $\boldsymbol{a} \otimes \boldsymbol{b} = a_i b_j$) or two symmetric second order tensors (e.g. $(\boldsymbol{\alpha} \otimes \boldsymbol{\beta}) = \alpha_{ij}\beta_{kl}$). As for sign conventions, we consider the direction of the tensile stress and dilative pressure as positive. We impose a superscript $(\cdot)^{\text{DEM}}$ on a variable to emphasize that such variable is inferred from DEM.

2 THEORETICAL BASIS FOR MICROPOLAR DISCRETE-CONTINUUM MODEL

Previous work on hierarchical DEM-FEM coupling models have been mainly focused on establishing coupling between discrete element simulations and Cauchy-type-continuum finite element analysis via first-order homogenization procedure [17, 6, 18, 19]. This first order homogenization is, nevertheless, only valid if the separation of scales exists [20, 21]. A separation of scales means that the characteristic length of macroscale problem l_{macro}, the size of RVEs l_{meso} and particle size l_{micro} fulfills the following inequality, i.e.

$$l_{micro} \ll l_{meso} \ll l_{macro}. \tag{1}$$

However, there are situations (e.g. presence of defects, inclusions, crack tips, dislocation) in which the meso-scale characteristic length l_{meso} is of comparable size of the macroscopic counterpart l_{marco}. In those cases, it is important to properly reflect the size effect in the homogenization procedure.

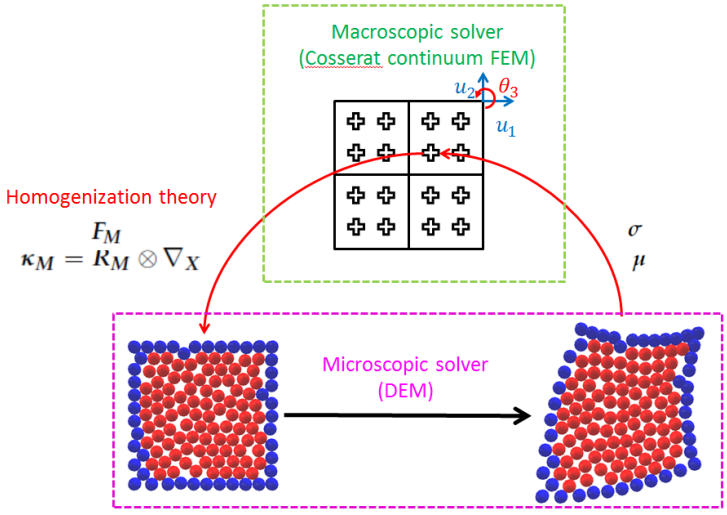


Figure 1: Multiscale discrete-micropolar continuum model

This can be done by using a high-order homogenization procedure to incorporate the size effect. The higher-order discrete-continuum coupling model proposed in this paper requires three building blocks (as shown in Fig.1): (1) the macroscopic finite element solver that provides incremental kinematic updates to the DEM solver, (2) the second order homogenization scheme that acts as an interface between the DEM and FEM solvers by converting macroscopic strain measures into constraints for RVE and up-scaling stress measures from the RVE to the integration points of the FEM solver and, (3) the DEM solver that calculate the first- and high-order stress based on the force and moment exerted on the grain contacts. To provide a proof of concepts, we have implemented a 2D micropolar DEM-FEM model for small strain problems. The following sections provide a brief outline of the the key ingredients of the numerical model, in

particular the infinitesimal micropolar theroy in 2D [22] and the second order homogenization theory [23].

2.1 2D small strain micropolar balance principle for macroscopic problems

Here we limit our attention on a prototype model that assumes the strain is infinitesimal and under plane strain condition. Recall that the kinematics of the micropolar continua can be defined by two sets of kinematic degrees of freedom that describe the changes of the position and orientation, e.g. The Cosserat strain $\boldsymbol{\gamma}$ that takes account of the higher-order kinematics is defined as a function of both the deformation gradient $u_{i,j}$ and the curvature tensor $\boldsymbol{\kappa}$, i.e.,

$$\begin{aligned}\gamma_{ij} &= u_{i,j} - \epsilon_{ijk}\theta_k \\ &= e_{ij} + w_{ij} - \epsilon_{ijk}\theta_k \\ &= e_{ij} - \epsilon_{ijk}(\theta_k - \omega_k)\end{aligned} \tag{2}$$

$$\kappa_{ij} = \theta_{i,j} \tag{3}$$

where $e_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i})$ and $w_{ij} = \frac{1}{2}(u_{i,j} - u_{j,i})$ are the components of the infinitesimal strain and infinitesimal rotation tensors. ϵ_{ijk} is the component of the Levi-Civita tensor $\overset{3}{\boldsymbol{E}}$. ω_k is the axial vector of the skew-symmetric infinitesimal rotation tensor w_{ij}.

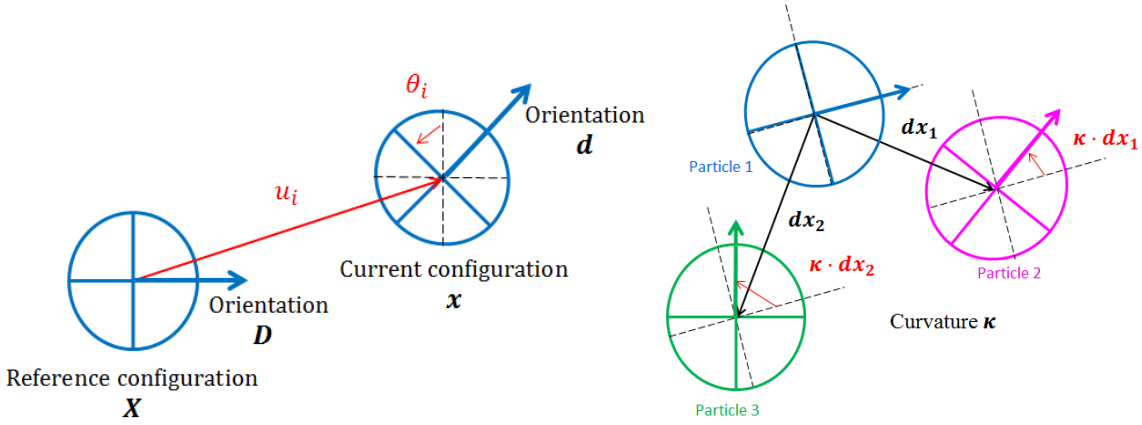


Figure 2: Micropolar kinematics. Left: displacement and rotation of a particle. Right: difference between the orientations of neighbor particles

The stress measures energy-conjugate to the micropolar strain and curvature are Cauchy stress $\boldsymbol{\sigma}$ and couple stress $\boldsymbol{\mu}$, respectively. $\boldsymbol{\mu}$ takes into the account of the moment induced by curvature. Note that, in the context of micropolar theory, the Cauchy stress $\boldsymbol{\sigma}$ no longer holds symmetry and is sometime referred as the force stress tensor [16]. Neglecting the inertia effect, the balance principle of linear and angular momentum for the static equilibrium reads,

$$\frac{\partial \sigma_{ik}}{\partial x_i} + f_k = 0, \tag{4}$$

$$\frac{\partial \mu_{ik}}{\partial x_i} + \epsilon_{imk}\sigma_{im} = 0. \tag{5}$$

For two-dimensional problem, Eq.(5) can be simplified as,

$$\frac{\partial \mu_{13}}{\partial x_1} + \frac{\partial \mu_{23}}{\partial x_2} + \sigma_{12} - \sigma_{21} = 0, \tag{6}$$

2.2 Micropolar homogenization procedure on DEM unit cells

The procedure that applies macroscopic deformation gradient $\boldsymbol{F}_M$ and curvature $\boldsymbol{\kappa}_M$ to microscopic problem is developed by Larsson and Diebels [23]. This scheme is extended to DEM RVEs based on [17]. Note the position of the center of a particle in reference configuration as $\boldsymbol{X}$ and in current configuration as $\boldsymbol{x}$. The mapping between $\boldsymbol{X}$ and $\boldsymbol{x}$ reads [23],

$$\boldsymbol{x} = \boldsymbol{F}_M \cdot \boldsymbol{X} + \frac{1}{2}((\boldsymbol{R}_M \otimes \nabla_{\boldsymbol{X}}) \cdot \boldsymbol{X}) \cdot \boldsymbol{X} + \boldsymbol{u}^f(\boldsymbol{X}), \tag{7}$$

where $\boldsymbol{u}^f(\boldsymbol{X})$ is the microstructural fluctuation of the particle center. The Hill-Mandel micro-heterogeneity condition requires that the macroscopic deformation gradient $\boldsymbol{F}_M$ is the volume average of the microscopic deformation gradient $\boldsymbol{F}$. This imposes two constraints on the DEM RVE problem: (1) the origin of the local coordinate system is set to the center of gravity of the grain assembly; (2) $\boldsymbol{u}^f(\boldsymbol{X}) = 0$ for all particles on the boundary. This corresponds to the uniform displacement boundary condition on DEM RVE.

The deformed RVEs by first-order and second-order homogenization are shown in Fig. 3. Note that in first-order the deformation is composed of extension/compression and simple shear modes, while in second-order there are additionally curvature modes.

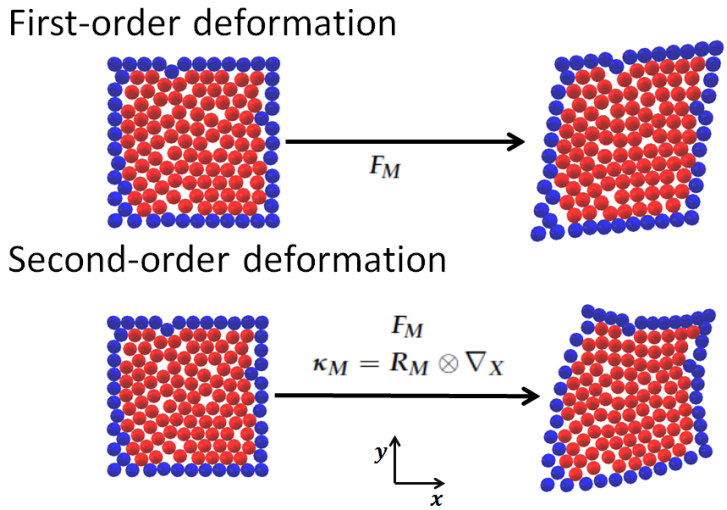


Figure 3: Deformation of DEM RVE by first-order and second-order theory

Upon deforming the RVE to a new equilibrium configuration, the macroscopic Cauchy stress is homogenized from DEM:

$$< \boldsymbol{\sigma}(\boldsymbol{x}, t) >_{\text{RVE}} = \frac{1}{\text{V}_{\text{RVE}}} \sum_{c}^{Nc} \boldsymbol{f}^c \otimes \boldsymbol{l}^c \tag{8}$$

where $\boldsymbol{f}^c$ is the contact force at the grain contact $\boldsymbol{x}^c$ and $\boldsymbol{l}^c$ is the branch vector that connects the centroids of two grains forming the contact ($\boldsymbol{x}^a$ and $\boldsymbol{x}^b$) [24, 25]. V_{RVE} is the volume of the RVE and N_c is the total number of particles in the RVE. The homogenized macroscopic couple stress can be expressed as [26]:

$$< \boldsymbol{\mu}(\boldsymbol{x}, t) >_{\text{RVE}} = \frac{1}{\text{V}_{\text{RVE}}} \sum_{c}^{Nc} (\overset{3}{\boldsymbol{E}} \cdot \boldsymbol{f}^c \otimes (\boldsymbol{x}^b \otimes (\boldsymbol{x}^c - \boldsymbol{x}^b) - \boldsymbol{x}^a \otimes (\boldsymbol{x}^c - \boldsymbol{x}^a)) + \boldsymbol{m}^c \otimes (\boldsymbol{x}^b - \boldsymbol{x}^a)) \tag{9}$$

The couple stress consists of two contributions: the eccentric contact force $\boldsymbol{f}^c$ and the contact moment $\boldsymbol{m}^c$.

3 NUMERICAL METHODS

The micropolar FEM-DEM model for plane strain problems is formulated in this section. Firstly, to construct the macroscopic boundary-value problem for micropolar continuum, consider a domain $\mathcal{B}$ with its boundary $\partial\mathcal{B}$ composed of Dirichlet boundaries (displacement $\partial\mathcal{B}_{\boldsymbol{u}}$, micro-rotation $\partial\mathcal{B}_{\boldsymbol{\theta}}$) and Von Neumann boundaries (surface traction $\partial\mathcal{B}_{\boldsymbol{t}}$, surface moment $\partial\mathcal{B}_{\boldsymbol{m}}$) satisfying

$$\begin{cases} \partial\mathcal{B} = \overline{\partial\mathcal{B}_{\boldsymbol{u}} \cup \partial\mathcal{B}_{\boldsymbol{t}}} = \overline{\partial\mathcal{B}_{\boldsymbol{\theta}} \cup \partial\mathcal{B}_{\boldsymbol{m}}} \\ \emptyset = \partial\mathcal{B}_{\boldsymbol{u}} \cap \partial\mathcal{B}_{\boldsymbol{t}} = \partial\mathcal{B}_{\boldsymbol{\theta}} \cap \partial\mathcal{B}_{\boldsymbol{m}} \end{cases} \tag{10}$$

The prescribed boundary conditions are

$$\begin{cases} \boldsymbol{u} = \overline{\boldsymbol{u}} \text{ on } \partial\mathcal{B}_{\boldsymbol{u}} \\ \boldsymbol{n}\cdot\boldsymbol{\sigma} = \overline{\boldsymbol{t}} \text{ on } \partial\mathcal{B}_{\boldsymbol{t}} \\ \boldsymbol{\theta} = \overline{\boldsymbol{\theta}} \text{ on } \partial\mathcal{B}_{\boldsymbol{\theta}} \\ \boldsymbol{n}\cdot\boldsymbol{\mu} = \overline{\boldsymbol{m}} \text{ on } \partial\mathcal{B}_{\boldsymbol{m}} \end{cases} \tag{11}$$

where $\overline{(\cdot)}$ denotes prescribed values and $\boldsymbol{n}$ is outward unit normal of surface $\partial\mathcal{B}$.

Following the standard procedures of the variational formulation, we obtain the weak form of the balance of linear momentum and angular momentum

$$\begin{aligned} G &: V_{\boldsymbol{u}} \times V_{\boldsymbol{\eta}} \to \mathbb{R} \\ G(\boldsymbol{u},\boldsymbol{\eta}) &= \int_{\mathcal{B}} \boldsymbol{\nabla}\boldsymbol{\eta} : \boldsymbol{\sigma}^{\text{DEM}} \,\mathrm{dV} - \int_{\mathcal{B}} \boldsymbol{\eta}\cdot\rho\boldsymbol{g} \,\mathrm{dV} \\ &\quad - \int_{\partial\mathcal{B}_{\boldsymbol{t}}} \boldsymbol{\eta}\cdot\bar{\boldsymbol{t}} \,\mathrm{d\Gamma} = 0 \end{aligned} \tag{12}$$

$$\begin{aligned} H &: V_{\boldsymbol{\theta}} \times V_{\boldsymbol{\psi}} \to \mathbb{R} \\ H(\boldsymbol{\theta},\boldsymbol{\psi}) &= \int_{\mathcal{B}} \boldsymbol{\nabla}\boldsymbol{\psi} : \boldsymbol{\mu}^{\text{DEM}} \,\mathrm{dV} - \int_{\mathcal{B}} \boldsymbol{\psi}\cdot\overset{3}{\boldsymbol{E}} : \boldsymbol{\sigma}^{\text{DEM}} \,\mathrm{dV} - \int_{\mathcal{B}} \boldsymbol{\psi}\cdot\rho\boldsymbol{J}\cdot\boldsymbol{c} \,\mathrm{dV} \\ &\quad - \int_{\partial\mathcal{B}_{\boldsymbol{m}}} \boldsymbol{\psi}\cdot\overline{\boldsymbol{m}} \,\mathrm{d\Gamma} = 0 \end{aligned} \tag{13}$$

The displacement and micro-rotation trial spaces for the weak form are defined as

$$V_{\boldsymbol{u}} = \{\boldsymbol{u} : \mathcal{B} \to \mathbb{R}^3 | \boldsymbol{u} \in [H^1(\mathcal{B})]^3, \boldsymbol{u}|_{\partial B_{\boldsymbol{u}}} = \overline{\boldsymbol{u}}\} \tag{14}$$

$$V_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} : \mathcal{B} \to \mathbb{R}^3 | \boldsymbol{\theta} \in [H^1(\mathcal{B})]^3, \boldsymbol{\theta}|_{\partial B_{\boldsymbol{\theta}}} = \overline{\boldsymbol{\theta}}\} \tag{15}$$

and the corresponding admissible spaces of variations are defined as

$$V_{\boldsymbol{\eta}} = \{\boldsymbol{\eta} : \mathcal{B} \to \mathbb{R}^3 | \boldsymbol{\eta} \in [H^1(\mathcal{B})]^3, \boldsymbol{\eta}|_{\partial B_{\boldsymbol{u}}} = \boldsymbol{0}\} \tag{16}$$

$$V_{\psi} = \{\boldsymbol{\psi} : \mathcal{B} \to \mathbb{R}^3 | \boldsymbol{\psi} \in [H^1(\mathcal{B})]^3, \boldsymbol{\psi}|_{\partial B_{\boldsymbol{\theta}}} = \mathbf{0}\} \tag{17}$$

where H^1 denotes the Sobolev space of degree one.

The spatially discretized equations can be derived following the standard Galerkin procedure. Shape functions $\boldsymbol{N_u}(\boldsymbol{x})$ and $\boldsymbol{N_\theta}(\boldsymbol{x})$ are used for approximation of solid motion $\boldsymbol{u}$ and micro-rotation $\boldsymbol{\theta}$ respectively:

$$\begin{cases} \boldsymbol{u} = \boldsymbol{N_u}\boldsymbol{U}, \;\; \boldsymbol{\eta} = \boldsymbol{N_u}\bar{\boldsymbol{\eta}} \\ \boldsymbol{\theta} = \boldsymbol{N_\theta}\boldsymbol{\Theta}, \;\; \boldsymbol{\psi} = \boldsymbol{N_\theta}\bar{\boldsymbol{\psi}} \end{cases} \tag{18}$$

with $\boldsymbol{U}$ being the nodal displacement vector, $\boldsymbol{\Theta}$ being the nodal micro-rotation vector, and $\bar{\boldsymbol{\eta}}$, $\bar{\boldsymbol{\psi}}$ being their variations.

In the DEM-FEM implementation, a Cartesian coordinate system is adopted with axes denoted as $(\boldsymbol{x}_1, \boldsymbol{x}_2\ \boldsymbol{x}_3)$. In the two-dimensional case, we consider the cross section orthogonal to the $\boldsymbol{x}_3$-axis. A equal-order quadrilateral mixed finite element is used to interpolate the displacement and micro-rotation fields with the same standard bilinear shape functions: $\boldsymbol{N} = \boldsymbol{N_u} = \boldsymbol{N_\theta}$. For the two-dimensional problems expressed in Cartesian coordinates, the nodal generalized displacement vector contains three degree of freedoms, two for displacement and one for change of orientation, i.e.,

$$\boldsymbol{d} = [u_1 \;\; u_2 \;\; \theta_3]^{\mathrm{T}} \tag{19}$$

The generalized strain vector for 2D plane strain problems consists of the micropolar strain $\boldsymbol{\gamma}$ and the curvature $\boldsymbol{\kappa}$:

$$\boldsymbol{E} = [\gamma_{11} \;\; \gamma_{22} \;\; \gamma_{33} \;\; \gamma_{12} \;\; \gamma_{21} \;\; \kappa_{31} \;\; \kappa_{32}]^{\mathrm{T}} \tag{20}$$

The Cauchy stress $\boldsymbol{\sigma}$ and the couple stress $\boldsymbol{\mu}$ homogenized from DEM RVEs at each integration point are grouped into the generalized stress vector written as:

$$\boldsymbol{S}^{\mathrm{DEM}} = [\sigma_{11} \;\; \sigma_{22} \;\; \sigma_{33} \;\; \sigma_{12} \;\; \sigma_{21} \;\; \mu_{31} \;\; \mu_{32}]^{\mathrm{T}} \tag{21}$$

Accordingly, the element shape function matrix $\boldsymbol{N}^e$ and the generalized element B matrix $\boldsymbol{B}^e$ can be expressed as,

$$\boldsymbol{N}^e = \begin{bmatrix} N_1 & 0 & 0 & N_2 & 0 & 0 & \dots & \dots \\ 0 & N_1 & 0 & 0 & N_2 & 0 & \dots & \dots \\ 0 & 0 & N_1 & 0 & 0 & N_2 & \dots & \dots \end{bmatrix} \tag{22}$$

$$\boldsymbol{B}^e = \begin{bmatrix} \frac{\partial N_1}{x_1} & 0 & 0 & \dots & \dots \\ 0 & \frac{\partial N_1}{x_2} & 0 & \dots & \dots \\ 0 & 0 & 0 & \dots & \dots \\ \frac{\partial N_1}{x_2} & 0 & N_1 & \dots & \dots \\ 0 & \frac{\partial N_1}{x_1} & -N_1 & \dots & \dots \\ 0 & 0 & \frac{\partial N_1}{x_1} & \dots & \dots \\ 0 & 0 & \frac{\partial N_1}{x_2} & \dots & \dots \end{bmatrix} \tag{23}$$

Finally, the finite element equation for balance of linear momentum and angular momentum in domain $\mathcal{B}$ is written in a compact form as:

$$\underbrace{\int_{\mathcal{B}} \boldsymbol{B}^{\mathrm{T}} \boldsymbol{S}^{\mathrm{DEM}}\, dV}_{\boldsymbol{F}^{int}(\boldsymbol{d})} = \underbrace{\int_{\mathcal{B}} \boldsymbol{N}^{\mathrm{T}} \boldsymbol{F}\, dV + \int_{\partial\mathcal{B}} \boldsymbol{N}^{\mathrm{T}} \boldsymbol{T}\, dA}_{\boldsymbol{F}^{ext}} \tag{24}$$

where $\boldsymbol{T} = [\bar{t}_1 \;\; \bar{t}_2 \;\; \bar{m}_3]^{\mathrm{T}}$ and $\boldsymbol{F} = [\rho g_1 \;\; \rho g_2 \;\; \rho(\boldsymbol{J}\cdot\boldsymbol{c})_3]^{\mathrm{T}}$ are generalized traction and body force vectors, respectively. The non-linear system of equations is solved by a modified implicit-explicit scheme which is originally proposed in Hughes et al. [27] and Prevost [28] to solve single-scale hydro-mechanical transient problems. This solution scheme for DEM-FEM model allows larger time step size compared to explicit scheme and avoids additional computational cost in computing the elasto-plastic tangential stiffness from DEM RVEs in fully implicit scheme. The implicit-explicit predictor-corrector scheme is performed by evaluating a portion of the left hand side forces explicitly using the predicted solution $\tilde{\boldsymbol{d}}$ defined as:

$$\tilde{\boldsymbol{d}} = \boldsymbol{d_n} \tag{25}$$

and by treating the remaining portion implicitly with the solution $\boldsymbol{d_{n+1}}$:

$$\boldsymbol{d_{n+1}} = \tilde{\boldsymbol{d}} + \Delta \boldsymbol{d_{n+1}} \tag{26}$$

The implicit-explicit partition reads,

$$\boldsymbol{F}^{int}_{\mathrm{IMP}}(\boldsymbol{d_{n+1}}) + \boldsymbol{F}^{int}_{\mathrm{EXP}}(\tilde{\boldsymbol{d}}) = \boldsymbol{F^{ext}_{n+1}} \tag{27}$$

To obtain the incremental update of the macroscopic displacement and micropolar rotation from the non-linear system of equations, Newton-Raphson method is employed. As a result, the consistent linearization of the implicit part $\boldsymbol{F}_{\mathrm{IMP}}$ is required. The resulting tangential stiffness matrix depends on what force terms are included in $\boldsymbol{F}^{int}_{\mathrm{IMP}}$.

Since computation of the homogenized $\boldsymbol{K^s}$ from DEM RVEs produces considerable computational cost, in the proposed multi-scale solution scheme, we choose to treat the elastic stiffness $\boldsymbol{K^e}$ implicitly and $\boldsymbol{K^{ep}}$ explicitly. In particular, we assume that the elastic response is linear and isotropic and use a perturbation method to obtain the all the non-polar and micropolar elastic material parameters from the RVEs at the initial step. The resultant operator-split stiffness matrices read,

$$\begin{cases} \boldsymbol{K_T}^{implicit} = \dfrac{\partial \boldsymbol{F}^{int}_{\mathrm{IMP}}}{\partial \boldsymbol{d}} = \boldsymbol{K^e} \\ \boldsymbol{K_T}^{explicit} = -\boldsymbol{K^{ep}} \end{cases} \tag{28}$$

4 NUMERICAL EXAMPLES

The micropolar FEM-DEM model is verified against the problem of stress concentration around a circular hole in a field of uniaxial and uniform tension. Consider a plane stress problem where a circular hole with radius r is streched by an uniform tensile stress field P. Classical elasticity theory states that the stress concentration factor K_c (ratio between the maximum value of tensile stress around the hole and P) is 3. In the realm of microplar theory, however, a reduction of this factor is predicated. The analytical solutions are studied by Mindlin [29] using the couple-stress theory, by Kaloni and Ariman [30] and Neuber [31] using the micropolar

theory, and the solutions are unified by Cowin [32]. These studies conclude that K_c, instead of having the constant value of 3, depends on geometry and material properties, namely the Poisson's ratio ν, the ratio of r to the material characteristic length l and a coupling number N. The micropolar material parameters are explained in details in [33].

The spatial domain, finite element discretization and boundary conditions of the numerical problem are illustrated in Fig. 4. Due to the symmetry of the problem domain, only a quarter of the circular hole is solved. On the two cutting-planes, the normal displacements as well as the rotations are restricted to zero. The circular hole has the radius of $0.05m$ while the patch size is $1m$, ensuring a uniform tension field P applied around the hole. A strongly gradated mesh is used to efficiently capture the stress concentration field around the hole. The DEM sample attached to all Gauss integration points of the finite elements is also shown. The grains with radius of $0.05m$ are arranged in a 5x5 cubic array and a concrete interaction model is used to allow tensile forces between the particles [34]. The DEM sample is initially stress-free and remains in elastic region under the uniaxial tension field.

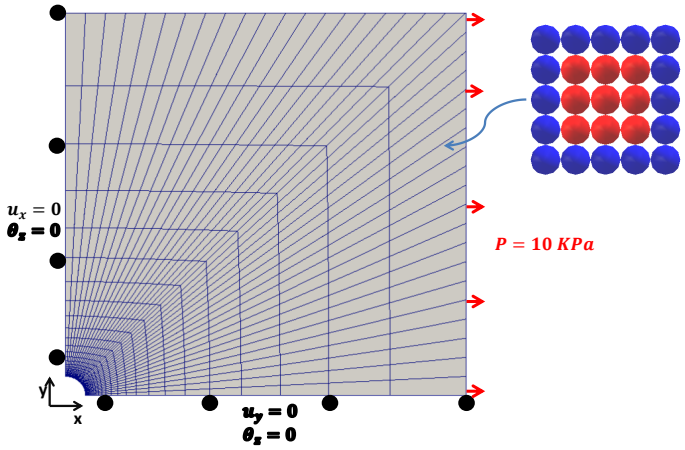


Figure 4: Geometry, mesh, boundary conditions and DEM RVE for stress concentration problem

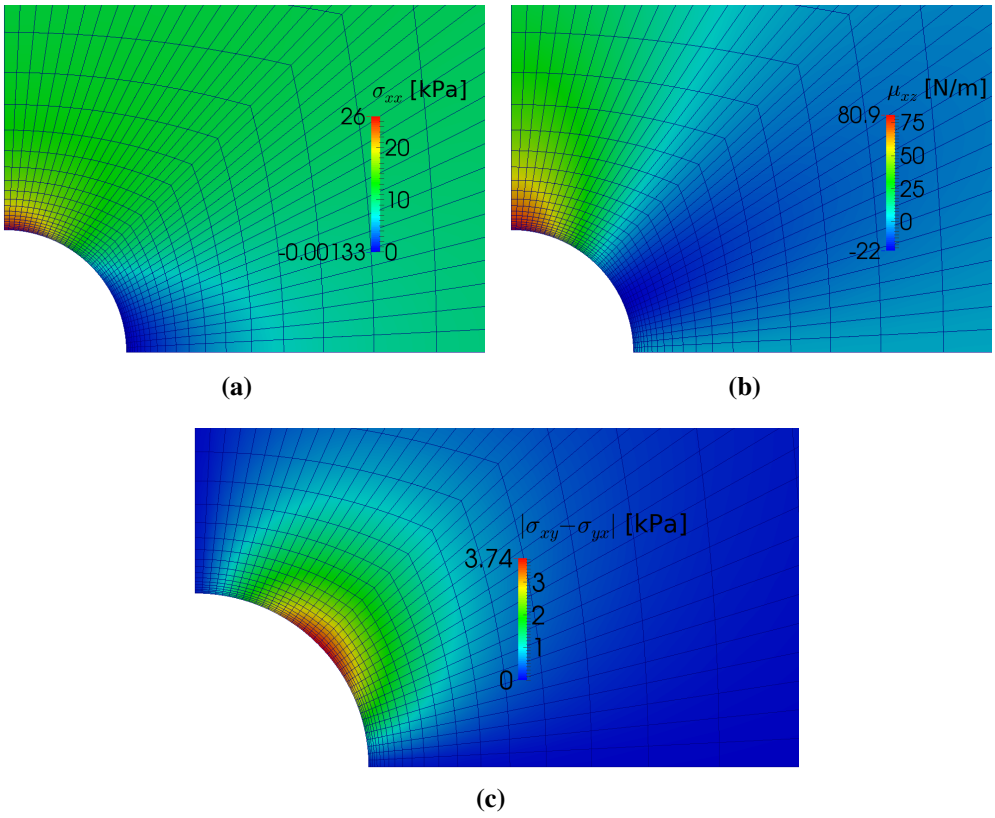
The material parameters in the context of the micropolar elasticity theory and the values identified from the DEM concrete sample using the perturbation method are recapitulated in Table 1. Note that the internal length l is 0.718 of the grain radius, which is in accordance with the Mindlin's statement that l is about 0.75 of the radius of a sphere when the identical elastic spheres are arranged in a simple cubic array [29].

The Cauchy stress and couple-stress fields around the circular hole given by the micropolar FEM-DEM model are presented in Fig.5. The stress concentration factor $K_c = \max(\sigma_{xx})/P$ is about 2.6, smaller than the value of 3 predicted by the classical continuum theory. The distributions of couple-stress μ_{xz} and μ_{yz} are symmetric about the bisectrix line. The existence of these couple-stresses results in the asymmetry of the Cauchy stress, as illustrated by the field of difference of shear stresses $|\sigma_{xy} - \sigma_{yx}|$ (Fig. 5(c)).

To illustrate the effect of material length scale l on the stress concentration factor, the grain radius of the DEM concrete sample is varied ($r = 0.1m,\ 0.05m,\ 0.01m,\ 0.005m,\ 0.001m$). The length scales l are always 0.718 of the corresponding radius and the Poisson's ratio ν and the coupling number N remain the same values. The stress concentration factor K_c is plotted

Micropolar material parameters (notations in consistent with [33])	Value identified from DEM
Micropolar Young's modulus E_m	$4.802 \times 10^9\ Pa$
Poisson's ratio ν	0.235
Micropolar shear modulus G_m	$2.255 \times 10^9\ Pa$
Shear modulus μ^*	$1.916 \times 10^9\ Pa$
Modulus between anti-symmetric parts κ	$6.783 \times 10^8\ Pa$
Modulus of curvature γ	$1.163 \times 10^7\ N$
Coupling number N ($N = \sqrt{\frac{\kappa}{2(\mu^*+\kappa)}}$)	0.362
Internal length scale l ($l = \sqrt{\frac{\gamma}{2(2\mu^*+\kappa)}}$)	$0.0359m$

Table 1: Micropolar elasticity material parameters in the stress concentration problem



(a) **(b)**

(c)

Figure 5: Stress field around a circular hole under tension.

against the length ratio r/l in Fig. 6. The same variation trend of K_c between the multiscale solution and the analytical solution is observed. This study provides evidence that our multiscale model intrinsically incorporates the internal material length scale.

5 CONCLUSIONS

A new higher-order DEM-FEM model for granular materials is established by incorporating higher-order kinematics and second-order homogenization theory. This work demonstrates the possibility of using grain-scale simulations as a replacement of phenomenological constitutive

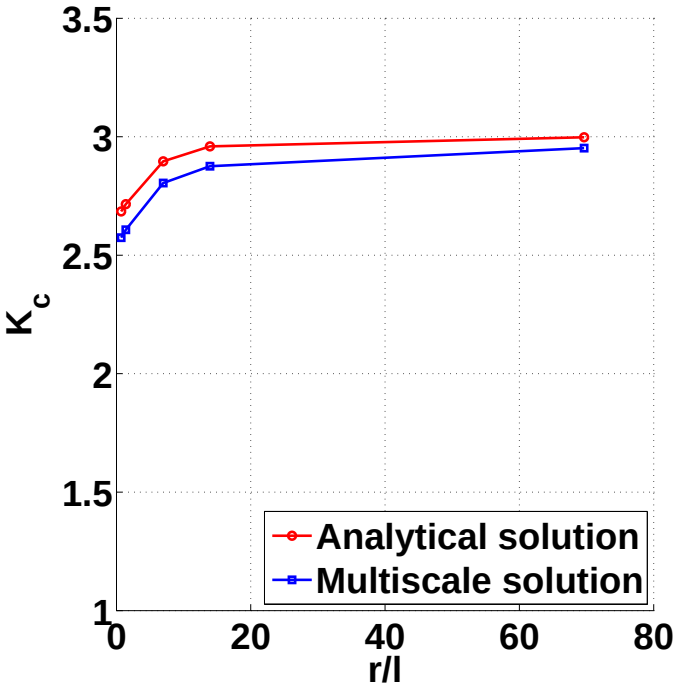


Figure 6: The stress concentration factor K_c as a function of length ratio r/l. Comparison between multiscale model solutions and analytical solutions [32]

law for micropolar continua. The resultant scheme retrieves the characteristic length naturally from the DEM assemblies, without the need to explicitly define one via a nonlocal integral. Another key upshot of the proposed high-order multiscale scheme is that it does not require any additional material parameters other than the one required for DEM. These features greatly simplifies the material identification procedure an allow one to analyze the interplay between evolution of micro-structural attributes and the macroscopic outcome.

6 ACKNOWLEDGMENTS

This research is supported by the Earth Materials and Processes program at the US Army Research Office under grant contract W911NF-14-1-0658 and W911NF-15-1-0581, as well as the Mechanics of Material program at National Science Foundation under grant contract CMMI-1462760 and Provost's Grants Program for Junior Faculty who Contribute to the Diversity Goals of the University at Columbia University. These supports are gratefully acknowledged.

REFERENCES

[1] Kun Wang and WaiChing Sun. Anisotropy of a tensorial bishop's coefficient for wetted granular materials. *Journal of Engineering Mechanics*, 2015. doi: 10.1061/(ASCE)EM.1943-7889.0001005.

[2] Kun Wang and WaiChing Sun. A semi-implicit discrete-continuum coupling method for porous media based on the effective stress principle at finite strain. *Computer Methods in Applied Mechanics and Engineering*, 2016. ISSN 0045-7825. doi: http://dx.doi.org/10.1016/j.cma.2016.02.020.

[3] David L Henann and Ken Kamrin. Continuum modeling of secondary rheology in dense granular materials. *Physical review letters*, 113(17):178001, 2014.

[4] WaiChing Sun, Matthew R Kuhn, and John W Rudnicki. A multiscale dem-lbm analysis on permeability evolutions inside a dilatant shear band. *Acta Geotechnica*, 8(5):465–480, 2013.

[5] WaiChing Sun. A unified method to predict diffuse and localized instabilities in sands. *Geomechanics and Geoengineering*, 8(2):65–75, 2013.

[6] Yang Liu, WaiChing Sun, Zifeng Yuan, and Jacob Fish. A nonlocal multiscale discrete-continuum model for predicting mechanical behavior of granular materials. *International Journal for Numerical Methods in Engineering*, 2015. ISSN 1097-0207. doi: 10.1002/nme.5139.

[7] Matthew R Kuhn, WaiChing Sun, and Qi Wang. Stress-induced anisotropy in granular materials: fabric, stiffness, and permeability. *Acta Geotechnica*, 10(4):399–419, 2015.

[8] John Walter Rudnicki and JR Rice. Conditions for the localization of deformation in pressure-sensitive dilatant materials. *Journal of the Mechanics and Physics of Solids*, 23 (6):371–394, 1975.

[9] WaiChing Sun and Alejandro Mota. A multiscale overlapped coupling formulation for large-deformation strain localization. *Computational Mechanics*, 54(3):803–820, 2014.

[10] WaiChing Sun. A stabilized finite element formulation for monolithic thermo-hydro-mechanical simulations at finite strain. *International Journal for Numerical Methods in Engineering*, 103(11):798–839, 2015.

[11] SeonHong Na and WaiChing Sun. Wave propagation and strain localization in a fully saturated softening porous medium under the non-isothermal conditions. *International Journal for Numerical and Analytical Methods in Geomechanics*, 2016.

[12] Eugène Cosserat and François Cosserat. Théorie des corps déformables. *Paris*, 1909.

[13] A Cemal Eringen and ES Suhubi. Nonlinear theory of simple micro-elastic solidsi. *International Journal of Engineering Science*, 2(2):189–203, 1964.

[14] Masanobu Oda and Kazuyoshi Iwashita. Study on couple stress and shear band development in granular media based on numerical simulation analyses. *International journal of engineering science*, 38(15):1713–1740, 2000.

[15] Wenxiong Huang and Erich Bauer. Numerical investigations of shear localization in a micro-polar hypoplastic material. *International journal for numerical and analytical methods in geomechanics*, 27(4):325–352, 2003.

[16] Jia Lin, Wei Wu, and Ronaldo I Borja. Micropolar hypoplasticity for persistent shear band in heterogeneous granular materials. *Computer Methods in Applied Mechanics and Engineering*, 289:24–43, 2015.

[17] Christian Miehe and Joachim Dettmar. A framework for micro–macro transitions in periodic particle aggregates of granular materials. *Computer Methods in Applied Mechanics and Engineering*, 193(3):225–256, 2004.

[18] N Guo, J Zhao, and WaiChing Sun. Multiscale analysis of shear failure of thick-walled hollow cylinder in dry sand. *Géotechnique Letters*, 6(1):1–6, 2016.

[19] Yang Liu, WaiChing Sun, and Jacob Fish. Determining material parameters for critical state plasticity models based on multilevel extended digital database. *Journal of Applied Mechanics*, 88(1), 2016.

[20] Jacob Bear and Yehuda Bachmat. *Introduction to modeling of transport phenomena in porous media*, volume 4. Springer Science & Business Media, 1990.

[21] Sia Nemat-Nasser and Muneo Hori. *Micromechanics: overall properties of heterogeneous materials*. Elsevier, 2013.

[22] I Vardoulakis. Lecture notes on cosserat continuum mechanics with application to the mechanics of granular media. In *3rd National Meeting on Generalized Continuum Theories and Applications, Thessaloniki, Greece*, 2009.

[23] Ragnar Larsson and Stefan Diebels. A second-order homogenization procedure for multiscale analysis based on micropolar kinematics. *International journal for numerical methods in engineering*, 69(12):2485–2512, 2007.

[24] J Christoffersen, MM Mehrabadi, and S Nemat-Nasser. A micromechanical description of granular material behavior. *Journal of Applied Mechanics*, 48(2):339–344, 1981.

[25] Katalin Bagi. Stress and strain in granular assemblies. *Mechanics of materials*, 22(3): 165–177, 1996.

[26] JP Bardet and I Vardoulakis. The asymmetry of stress in granular media. *International Journal of Solids and Structures*, 38(2):353–367, 2001.

[27] Thomas JR Hughes, Karl S Pister, and Robert L Taylor. Implicit-explicit finite elements in nonlinear transient analysis. *Computer Methods in Applied Mechanics and Engineering*, 17:159–182, 1979.

[28] Jean H Prevost. Implicit-explicit schemes for nonlinear consolidation. *Computer Methods in Applied Mechanics and Engineering*, 39(2):225–239, 1983.

[29] RD Mindlin. Influence of couple-stresses on stress concentrations. *Experimental Mechanics*, 3(1):1–7, 1963.

[30] Puran N Kaloni and Teoman Ariman. Stress concentration effects in micropolar elasticity. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 18(1):136–141, 1967.

[31] H Neuber. On the general solution of linear-elastic problems in isotropic and anisotropic cosserat continua. In *Applied Mechanics*, pages 153–158. Springer, 1966.

[32] Stephen C Cowin. An incorrect inequality in micropolar elasticity theory. *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, 21(3):494–497, 1970.

[33] S Bauer, M Schäfer, P Grammenoudis, and Ch Tsakmakis. Three-dimensional finite elements for large deformation micropolar elasticity. *Computer Methods in Applied Mechanics and Engineering*, 199(41):2643–2654, 2010.

[34] Václav Šmilauer. *Cohesive particle model using the discrete element method on the yade platform*. PhD thesis, Université de Grenoble; Czech Technical University in Prague (http://www. cvut. cz), 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MODELING OF LOW-ALLOYED TRIP-STEELS BASED ON DIRECT MICRO-MACRO SIMULATIONS

S. Prüger[1], A. Gandhi[1], and D. Balzani[1,2]

[1]TU Dresden
Faculty of Civil Engineering
Institute of Mechanics and Shell Structures
01062 Dresden, Germany
e-mail: stefan.prueger,ashutosh.gandhi,daniel.balzani@tu-dresden.de

[2] Dresden Center for Computational Materials Science, Germany

Keywords: Computational homogenization, Low-alloyed TRIP-steel, Martensitic phase transformation, Transformation kinetics

Abstract. *Low-alloyed TRIP steels are often used in the automotive industry due to their favorable mechanical properties such as high ductility and strength and their moderate production costs. These steels possess a heterogeneous multiphase microstructure, initially consisting of ferrite, bainite and retained austenite which is responsible for the mechanical properties. Upon deformation, a diffusionless, stress-induced, martensitic phase transformation from austenite to martensite is observed, enhancing ductility and strength.*
*We focus on multi-scale methods in the sense of FE^2 to describe the macroscopic behavior of low-alloyed TRIP-steels, because this approach allows for a straightforward inclusion of various influencing factors such as residual stress distribution, graded material properties which can hardly included in phenomenological descriptions of these heterogeneous multiphase materials. In order to allow for efficient computations, a simplified microstructure is used in an illustrative direct micro-macro simulation. The inelastic processes in the austenitic inclusions involve the phase transformation from austenite to martensite and the inelastic deformation of these two phases. The isotropic, rate-independent, hyperelastic-plastic material model of Hallberg et al. (IJP, **23**, pp.1213–1239, 2007), originally proposed for high-alloyed TRIP steel, is adopted here for the inclusion phase. Minor modifications of the model are proposed to improve its implementation and performance. The influence of various material parameters associated with the phase transformation on the evolution of retained austenite is studied for different homogeneous deformation states. The non-monotonic stress-state dependence observed in experiments is clearly captured by the model. A numerical two-scale calculation is carried out to enlighten the ductility enhancement in low-alloyed TRIP-steels due to the martensitic phase transformation.*

1 INTRODUCTION

Low-alloyed TRIP-steels offer a favorable combination of both high strength and pronounced ductility, leading to a high energy absorption capacity. This is desired especially for automotive applications, because it allows for light-weight and crashworthy constructions and therefore leads to safer and more efficient automobiles. These favorable mechanical properties can be achieved at lower costs compared to high-alloyed TRIP-steels, due to the reduced amount of alloying elements required. In contrast to high-alloyed, initially fully austenititc TRIP-steels, low-alloyed TRIP-steels possess a multiphase microstructure, consisting of ferrite, bainite and retained austenite. In order to obtain the favorable mechanical properties the microstructure is optimized by a sophisticated heat treatment procedure and a smart alloy design, cf. [1]. Micromechanical considerations [2] lead to the conclusion that the high ductility of these steels cannot be solely attributed to the deformation-induced martensitic phase transformation from the metastable, retained austenite to martensite, because its volume fraction is typically in the range of 10-15% and therefore simply too low. Therefore, the multiphase character of the microstructure is responsible for the pronounced ductility. However, the phase transformation in the retained austenite plays an important role in this type of steel as it dynamically influences the stress and strain partitioning between different phases during deformation [3, 4, 5] and delays microcracking [6], leading to composite type microstructure with adaptive properties. The rate of the austenite to martensite transformation depends on various factors: (i) ambient temperature cf. [7], (ii) stress-state cf. [8], (iii) austenite grain size and morphology [9], (iv) neighboring constituents [9].

A variety of constitutive models has been proposed for the description of the deformation and transformation behavior of low-alloyed TRIP-steels. Phenomenological approaches such as [10, 11, 12, 13] mainly focus on capturing the influences of temperature and stress-state on the transformation kinetics. They are mostly used in single scale (macroscopic) simulations due to their low computational costs. Modeling approaches that incorporate analytical homogenization schemes, like for instance [14, 15, 16, 17, 18, 20] are better suited to account for the strengthening effect of the evolving martensite and multiphase character of the microstructure. However, simplifying assumptions regarding the microstructure morphology and field fluctuations within the phases have to be made, which may lead to inaccurate assessment of failure initialization. Direct microstructural simulations allow for a straightforward incorporation of all of the above mentioned influencing factors (i)-(iv). However, they are computationally very demanding as the macroscopic constitutive response is obtained by the solution of microscopic boundary value problems for a suitable representative volume element. In the context of low-alloyed TRIP-steels a two-dimensional microstructural section [21] or artificial inclusion type microstructure [22] as well as polyhedral inclusions representing single crystalline phases [23] have been used as representative volume elements.

In the current contribution, we follow the direct micro-macro simulation approach as this framework is in general sufficient to study the above mentioned influencing factors on the phase transformation behavior of low-alloyed TRIP-steels. The paper is organized as follows. In section 2, the basic equations for the direct micro-macro scale transition are summarized, whereas in section 3 a suitable material model for the phase transformation from austenite to martensite is adopted from the literature. In section 4, the influence of model parameters on the phase transformation behavior is illustrated and the effective mechanical behavior of a simplified three-dimensional microstructure for low-alloyed TRIP-steel is presented. Section 5 summarizes the main findings.

2 DIRECT MICRO-MACRO SCALE TRANSITION

Employing the direct micro-macro scale transition approach enables the computation of effective, macroscopic material behavior for a representative volume element with arbitrary complex microstructure by means of the solution of microscopic boundary value problems and appropriate averaging schemes of the microscopic fields. Here, we focus on a purely mechanical setting and state the basic relations, required for the computations. According to Hill [24] the stress power per unit reference volume at the macroscopic scale, expressed in terms of average quantities, should be equal to the microscopic counterpart, yielding

$$\langle \boldsymbol{P} : \dot{\boldsymbol{F}} \rangle = \langle \boldsymbol{P} \rangle : \langle \dot{\boldsymbol{F}} \rangle \,, \tag{1}$$

where $\boldsymbol{P}$ denotes the 1st Piola-Kirchhoff stress tensor, $\dot{\boldsymbol{F}}$ is the material time derivative of the deformation gradient and the volume average over the reference configuration is defined as $\langle \cdot \rangle = \frac{1}{V} \int_{\mathcal{B}_0} \cdot \, \mathrm{d}V$. As the current study relies on representative volume elements with a periodic microstructure, it is convenient to apply periodic boundary conditions in order to enforce eq. (1). Based on the decomposition of the deformation field in a homogeneous deformation and fluctuation as

$$\dot{\boldsymbol{x}} = \dot{\overline{\boldsymbol{F}}} \cdot \boldsymbol{X} + \dot{\widetilde{\boldsymbol{w}}} \,, \tag{2}$$

the periodic boundary conditions require a periodic fluctuation field $\widetilde{\boldsymbol{w}}$ and antiperiodic traction vectors $\boldsymbol{t}_0$ along the boundary, namely

$$\dot{\widetilde{\boldsymbol{w}}}^{+} = \dot{\widetilde{\boldsymbol{w}}}^{-} \qquad \text{and} \qquad \boldsymbol{t}_0^{+} = -\boldsymbol{t}_0^{-} \text{ on } \partial\mathcal{B}_0 \,. \tag{3}$$

3 MATERIAL MODEL AT THE MICRO-SCALE

In order to obtain reliable predictions of the macroscopic, effective deformation and transformation behavior of low-alloyed TRIP-steels, suitable material models that capture the essential features of inelastic processes are required at micro-scale. The aim of the present contribution is to describe the overall behavior of low-alloyed TRIP-steels under isothermal conditions and excluding rate-dependent effects. Therefore, rate-independent models that incorporate the two main inelastic processes, plasticity and the austenite to martensite phase transformation, should be chosen. Furthermore, assuming a polycrystalline microstructure at the micro-scale, the constitutive model proposed by Hallberg [25] for high-alloyed TRIP-steels is a suitable choice, because it is contains only a relatively small number of material parameters, but still includes plasticity and phase transformation as two independently evolving inelastic processes and their interaction in a simplified manner. In contrast to the originally proposed model, we apply some minor modifications which allow for a more convenient implementation, but the essential features of the model remain and in particular the thermodynamic consistency is still guaranteed. The material model is based on the multiplicative split of the deformation gradient

$$\boldsymbol{F} = \boldsymbol{F}^{\mathrm{e}} \cdot \boldsymbol{F}^{\mathrm{in}} \tag{4}$$

into an elastic and an inelastic part where the latter is associated with any inelastic process occurring in the material. Employing the assumption of isotropic elastic behavior and an additive split of the free energy according to

$$\rho_0 \Psi(\boldsymbol{b}^{\mathrm{e}}, \alpha_1, \alpha_2, \ldots) = \rho_0 \Psi^{\mathrm{e}}(\boldsymbol{b}^{\mathrm{e}}) + \rho_0 \Psi^{\mathrm{in}}(\alpha_1, \alpha_2, \ldots) \,, \tag{5}$$

with α_i representing the internal variables, the Kirchhoff-stress is defined as

$$\boldsymbol{\tau} = 2\rho_0 \frac{\partial \Psi}{\partial \boldsymbol{b}^{\mathrm{e}}} \cdot \boldsymbol{b}^{\mathrm{e}} \,. \tag{6}$$

Herein, the elastic left Cauchy-Green tensor is given as $\boldsymbol{b}^{\mathrm{e}} = \boldsymbol{F}^{\mathrm{e}} \cdot \boldsymbol{F}^{\mathrm{e}T}$. In particular, the elastic energy function

$$\rho_0 \Psi^{\mathrm{e}} = \frac{\lambda}{2} \left[\sum_A \ln(\lambda_A^{\mathrm{e}}) \right]^2 + \mu \sum_A \left[\ln(\lambda_A^{\mathrm{e}}) \right]^2 \tag{7}$$

is utilized, which is described in terms of principle elastic stretches λ_A^{e} and the Lamé constants λ, μ. The principle elastic stretches are obtained by the eigenvalue decomposition of $\boldsymbol{b}^{\mathrm{e}}$. The elastic left Cauchy-Green tensor can be related to the inelastic right Cauchy-Green tensor $\boldsymbol{C}^{\mathbf{in}} = \boldsymbol{F}^{\mathbf{in}T} \cdot \boldsymbol{F}^{\mathbf{in}}$ via

$$\boldsymbol{b}^{\mathrm{e}} = \boldsymbol{F} \cdot \boldsymbol{C}^{\mathbf{in}^{-1}} \cdot \boldsymbol{F}^T \,, \tag{8}$$

whereas the temporal evolution of $\boldsymbol{b}^{\mathrm{e}}$ and $\boldsymbol{C}^{\mathbf{in}}$ is expressed as

$$\mathcal{L}(\boldsymbol{b}^{\mathrm{e}}) = \boldsymbol{F} \cdot \left(\boldsymbol{C}^{\mathbf{in}^{-1}} \right)^{\cdot} \cdot \boldsymbol{F}^T \tag{9}$$

where $\mathcal{L}$ is the Lie derivative. An alternative representation of eq. (9) solely in terms of $\boldsymbol{C}^{\mathbf{in}}$ and $\boldsymbol{F}$ is obtained if the additive split of the rate of deformation $\boldsymbol{l} = \dot{\boldsymbol{F}} \cdot \boldsymbol{F}^{-1} = \boldsymbol{l}^{\mathrm{e}} + \boldsymbol{l}^{\mathbf{in}}$ is used in conjunction with the assumption of isotropic elastic and inelastic behavior, which is considered here, i.e.

$$\left(\boldsymbol{C}^{\mathbf{in}^{-1}} \right)^{\cdot} = -2\boldsymbol{F}^{-1} \cdot \underbrace{\sum_p \dot{\lambda}_p \frac{\partial \Phi_p}{\partial \boldsymbol{\tau}}}_{=\boldsymbol{d}^{\mathbf{in}}} \cdot \boldsymbol{F} \cdot \boldsymbol{C}^{\mathbf{in}^{-1}} \,. \tag{10}$$

As proposed by Hallberg [25] the inelastic rate of deformation $\boldsymbol{d}^{\mathbf{in}}$ is additively split into multiple contributions stemming from different inelastic processes, characterized by independent limit surfaces Φ_p and the corresponding consistency parameters $\dot{\lambda}_p$. Similar to small strain elastic-plastic formulations, the antimetric part of the inelastic velocity gradient remains undetermined.

For the rate-independent plasticity a limit surface of VON MISES-type with nonlinear isotropic hardening is chosen.

$$\Phi^{\mathrm{pl}} := \sqrt{3J_2} - \sigma^{\mathrm{y}}(\alpha^{\mathrm{pl}}, f_{\mathrm{m}}) \tag{11}$$

Herein the second invariant of the Kirchhoff stress deviator is defined as $J_2 = \frac{1}{2}\mathrm{dev}(\boldsymbol{\tau}) : \mathrm{dev}(\boldsymbol{\tau})$, whereas α^{pl} and f_{m} are the isotropic hardening variable and the martensite volume fraction, respectively. The yield stress is computed from an exponential hardening law and a nonlinear rule of mixture, according to

$$\sigma^{\mathrm{y}}(\alpha^{\mathrm{pl}}, f_{\mathrm{m}}) = \sigma_0^{\mathrm{y}} \bar{m}(f_{\mathrm{m}}) + \mathrm{H}\kappa(\alpha^{\mathrm{pl}}) \tag{12}$$

with

$$\bar{m}(f_{\mathrm{m}}) = 1 + (\exp(\mathrm{f}_1 f_{\mathrm{m}}) - 1)\mathrm{f}_2 \tag{13}$$

and

$$\kappa(\alpha^{\mathrm{pl}}) = \frac{\mathrm{R}_\infty}{\mathrm{H}} \left[1 - \exp\left(-\frac{\mathrm{H}}{\mathrm{R}_\infty} \alpha^{\mathrm{pl}} \right) \right] \,, \tag{14}$$

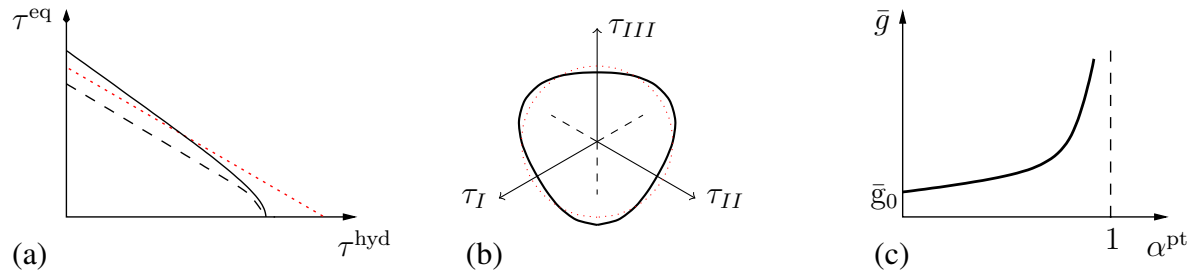


Figure 1: Limit surface for phase transformation: (a) section containing the hydrostatic axis with tensile $(--)$ and compressive meridian $(—)$ and the corresponding cone $(\cdots)$, (b) non-circular cross-section in the π-plane, (c) evolution of the transformation barrier.

where the hardening variable evolves according to $\dot{\alpha}^{\mathrm{pl}} = \dot{\lambda}^{\mathrm{pl}}$. The equations (12), (13), (14) introduce three material parameters $(\sigma_0^{\mathrm{y}}, \mathrm{H}, \mathrm{R}_\infty)$ for the hardening and two material parameters $(\mathrm{f}_1, \mathrm{f}_2)$ in the rule of mixture.
The limit surface for phase transformation is taken here as a hyperbolic approximation of Hallberg's conical transformation surface and is inspired by the proposal given in [26]. It possesses the form

$$\Phi^{\mathrm{pt}} := \sqrt{(\tau^{\mathrm{eq}})^2 \left(1 + \mathrm{k}\frac{J_3}{J_2^{3/2}}\right)^2 \mathrm{R}^2 + \Delta_{\mathrm{v}}^2(\kappa_1^2(\alpha^{\mathrm{pt}}, T) - 1)(\tau^{\mathrm{hyd}})^2} + \Delta_v \kappa_2(\alpha^{\mathrm{pt}}, T)\tau^{\mathrm{hyd}} - \mathrm{R}\kappa_1(\alpha^{\mathrm{pt}}, T)\,, \quad (15)$$

with

$$\kappa_1(\alpha^{\mathrm{pt}}, T) = \sqrt{\left(\frac{1}{\Delta_{\mathrm{v}}}\left(\bar{g}(\alpha^{\mathrm{pt}}) - \Delta g^{\mathrm{a}\to\mathrm{m}}(T)\right)\right)^2 - \mathrm{c}^2} \quad (16)$$

$$\kappa_2(\alpha^{\mathrm{pt}}, T) = \frac{1}{\Delta_{\mathrm{v}}}\frac{\bar{g}(\alpha^{\mathrm{pt}}) - \Delta g^{\mathrm{a}\to\mathrm{m}}(T)}{\kappa_1}\,. \quad (17)$$

The transformation surface introduces the material parameters $\Delta_{\mathrm{v}}, \mathrm{R}$ which correspond to the volumetric and deviatoric transformation strain, as well as the shape parameters k and c. The parameter k controls the deviation from the circular cross-section in the π-plane, i.e. the deviatoric plane that contains the origin and the parameter c determines the transition from the hyperbolic surface to the conical surface (see fig. 1). The latter parameter, however, is of minor importance in the studies considered in this paper. Although it guarantees a continuous differentiable limit surface under purely hydrostatic loading conditions, for the stress states reached in the loading scenarios considered, the hyperbolic approximation of the transformation surface is very close to the original conical surface. Furthermore, $\tau^{\mathrm{hyd}}, \tau^{\mathrm{eq}}, \Delta g^{\mathrm{a}\to\mathrm{m}}(T)$ and $\bar{g}(\alpha^{\mathrm{pt}})$ denote the hydrostatic part of the Kirchhoff stress, the VON MISES equivalent Kirchhoff stress, the temperature dependent energy difference between the austenitic (a) and the martensitic (m) phase and the transformation barrier, respectively. The evolution equations for the internal variables associated with the transformation, namely the hardening variable for transformation hardening $\alpha_1 := \alpha^{\mathrm{pt}}$ and the martensite volume fraction $\alpha_2 := f_{\mathrm{m}}$ are derived following the procedure for generalized standard materials [27]. Therefore, the portion of the free energy function corresponding to inelastic processes, introduced in eq. (5), is further specified as

$$\rho_0\Psi^{\mathrm{in}}(\alpha^{\mathrm{pt}}, \alpha^{\mathrm{pl}}, f_{\mathrm{m}}) = \rho_0\Psi^{\mathrm{pl}}(\alpha^{\mathrm{pl}}) + \rho_0\Psi^{\mathrm{pt}}(\alpha^{\mathrm{pt}}) + \rho_0\Psi^{\mathrm{chem}}(f_{\mathrm{m}}; T) \quad (18)$$

with

$$\rho_0\Psi^{\mathrm{chem}}(f_{\mathrm{m}}; T) = (1 - f_{\mathrm{m}})\rho_0\Psi_{\mathrm{a}}^{\mathrm{chem}}(T) + f_{\mathrm{m}}\rho_0\Psi_{\mathrm{m}}^{\mathrm{chem}}(T) \quad (19)$$

and

$$\rho_0 \Psi^{\mathrm{pt}}(\alpha^{\mathrm{pt}}) = \bar{\mathrm{g}}_1 \left(\left(1 - \alpha^{\mathrm{pt}}\right) \left(\ln(1 - \alpha^{\mathrm{pt}}) - 1 \right) + 1 \right) . \tag{20}$$

Herein, $\Psi^{\mathrm{pl}}(\alpha^{\mathrm{pl}})$ is linked to the isotropic hardening of the VON MISES criterion and chosen identical to the proposal of Hallberg [25] and is not included here for brevity. Upon defining the driving forces $K_{\alpha^{\mathrm{pt}}} = \rho_0 \frac{\partial \Psi^{\mathrm{pt}}}{\partial \alpha^{\mathrm{pt}}}$ and $K_{f_{\mathrm{m}}} = \rho_0 \frac{\partial \Psi^{\mathrm{chem}}}{\partial f_{\mathrm{m}}} = -\Delta g^{\mathrm{a}\to\mathrm{m}}(T)$, the evolution equations are defined as

$$\dot{f_{\mathrm{m}}} = -\dot{\lambda}^{\mathrm{pt}} \frac{\partial \Phi^{\mathrm{pt}}}{\partial K_{f_{\mathrm{m}}}} \tag{21}$$

$$\dot{\alpha}^{\mathrm{pt}} = -\dot{\lambda}^{\mathrm{pt}} \frac{\partial \Phi^{\mathrm{pt}}}{\partial K_{\alpha^{\mathrm{pt}}}} . \tag{22}$$

From the structure of the transformation surface it can be readily verified that $\alpha^{\mathrm{pt}} = f_{\mathrm{m}}$. The transformation barrier evolves according to the transformation hardening law

$$\bar{g}(\alpha^{\mathrm{pt}}) = \bar{\mathrm{g}}_0 - \bar{\mathrm{g}}_1 \ln(1 - \alpha^{\mathrm{pt}}) \tag{23}$$

with the initial transformation barrier $\bar{\mathrm{g}}_0$ and the initial hardening modulus $\bar{\mathrm{g}}_1$. As can be seen from fig. 1c the barrier is progressively increasing, limiting the hardening variable to $\alpha^{\mathrm{pt}} \leq 1$. The material model described above is implemented into the Finite Element Program FEAP and employs a combination of the operator split and a general return mapping algorithm proposed by Aurrichio et al. [28] to integrate the set of nonlinear evolution equations. Furthermore, the return mapping algorithm has been extended to handle non-smooth intersections of multiple limit surfaces as discussed in [29, p.206ff].

4 RESULTS

4.1 Parameter study

The material model described in section 3 contains 12 material parameters, which need to be adjusted in order to capture the deformation and transformation behavior of the retained austenite and the evolving martensite. The parameters associated with elasticity and plasticity of the austenitic/martensitic material can be chosen rather easily, if the individual hardening curves are known. In the present paper the parameters are set such that the austenite and martensite hardening curves of an experimental TRIP-steel presented in [19] are reproduced. For later reference they are listed in table 1. The influence of three of the remaining parameters, namely the

E [MPa]	ν	σ_0^{y} [MPa]	H [MPa]	R_∞[MPa]	f_1	f_2	Δ_{v}	R
200000	0.3	300	3500	420	1.65	1.1	0.04	0.07

Table 1: Material parameters of the austenitic/martensitic material

initial transformation barrier $\bar{\mathrm{g}}_0$, the initial transformation hardening modulus $\bar{\mathrm{g}}_1$ and the shape parameter k is studied in order to get a reasonable estimate for the range of parameters and their impact on the transformation kinetics. It is well known from experiments that the phase transformation in low-alloyed TRIP-steels is stress-state dependent. In particular, this dependence is non-monotonic with respect to the stress-triaxiality measure $h = \frac{\tau^{\mathrm{hyd}}}{\tau^{\mathrm{eq}}}$, based on the experimental results presented in [8] and depicted in fig.2c. Herein, 4 mechanical tests (simple shear, uniaxial tension, biaxial tension and the Marchiniak test) are carried out, where each of

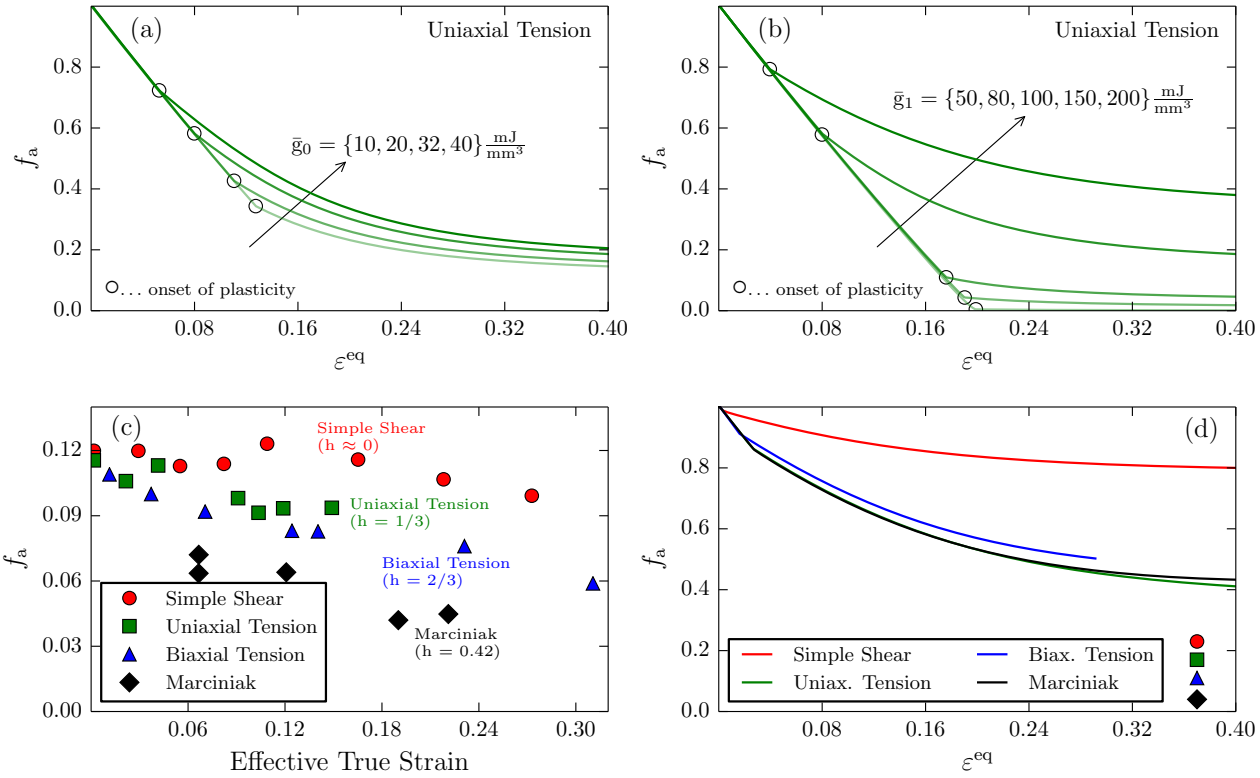


Figure 2: Influence on transformation kinetics: (a) variation of initial transformation barrier, (b) variation of transformation hardening modulus, (c) experimentally determined non-monotonic stress-state dependence cf. [8], (d) simulated non-monotonic stress-state dependence (symbols indicate the experimental trends: lowest transformation rate in simple shear and highest rate in Marciniak test)

these tests is characterized by a different stress state. These four tests are simulated employing deformation controlled one element tests with an initially fully austenitic microstructure and analyzing the evolution of retained austenite as a function of the overall equivalent strain ε^{eq}. It can be seen from fig. 2a,b that initially only phase transformation and no plastic deformation is predicted by the model for the given choice of parameters. While varying the initial transformation barrier changes the onset of plasticity, i.e. both inelastic deformation mechanisms are active at the same time and keeps the asymptotically reached value of retained austenite more or less unaltered, a modification of the hardening modulus influences both the onset of plasticity and the asymptotically reached value of retained austenite. In both studies a value of k=0.24 and $\bar{g}_1 = 150 \frac{mJ}{mm^3}$ and $\bar{g}_0 = 32 \frac{mJ}{mm^3}$ is kept while varying the other parameters. Although only the results for the uniaxial tensile test are presented, the same trend can be observed for all the homogeneous tests.

In a third study the shape parameter k is altered in the range of $-0.24 \leq k \leq 0.24$ leading to convex, non-circular cross-sections of the transformation in the π-plane. The initial transformation barrier and the transformation hardening are chosen in this study as $\bar{g}_0 = 40 \frac{mJ}{mm^3}$ and $\bar{g}_1 = 200 \frac{mJ}{mm^3}$. It can be seen from fig. 2d that the non-monotonic stress-state dependency is clearly captured for the choice k$= 0.24$, as the transformation rate in biaxial tension ($h = 2/3$) is lower than in the Marciniak test ($h = 0.42$). However, the transformation rate predicted by the material model in uniaxial tension contradicts the experimentally observed trend, which indicates the need for further parameter studies. For lower values of k the highest rate of transformation is observed in biaxial tension, leading to a proportional increase in transformation rate with increasing stress triaxiality.

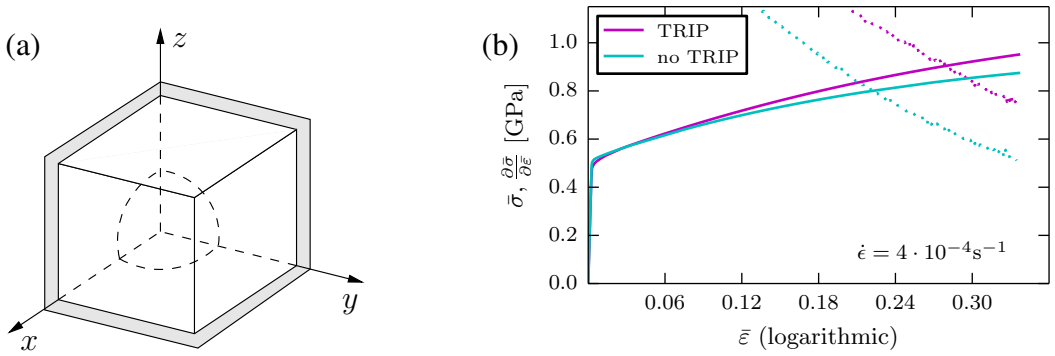


Figure 3: Simplified microstructure (a) and overall true stress-strain curve for uniaxial tension (b)

4.2 Simplified microstructure

In order to study the influence of the martensitic phase transformation on the macroscopic deformation behavior of low-alloyed TRIP-steels a simplified microstructure is considered which consists of a regular cubic arrangement of spherical, retained austenite inclusions embedded in ferritic/bainitic matrix. Note that more sophisticated but still efficient statistically similar representative volume elements may be constructed following the approach in Balzani et al. [30], see also Brands et al. [31] for an analysis in the context of DP steels. Consistent with experimental data, the initial volume fraction of retained austenite is chosen $f_{\mathrm{a}} = 0.12$. In the matrix a mixture of ferrite and bainite is employed with the volume fractions set to $f_{\mathrm{f}} = 0.5$ and $f_{\mathrm{b}} = 0.38$, corresponding to the phase composition of an experimental TRIP-steel presented in [19]. Both in the matrix and in the inclusion the material model described in section 3 is utilized. In the inclusion the material parameters given in table 1 and $\bar{\mathrm{g}}_0 = 40 \frac{\mathrm{mJ}}{\mathrm{mm}^3}$, $\bar{\mathrm{g}}_1 = 200 \frac{\mathrm{mJ}}{\mathrm{mm}^3}$ and k$= 0.24$ have been selected. As no phase transformation is observed in the matrix a rather high initial transformation barrier is used to switch off the transformation criterion. The material parameters associated with the nonlinear hardening law are computed to reasonable approximate the ferrite and bainite hardening curves given in [19]. All the material parameters required for the ferrite/bainite mixture are listed in table 2.
The representative volume element (RVE) containing the simplified microstructure is subjected to a displacement controlled uniaxial tensile test prescribing an axial true strain of $\bar{\varepsilon} = 0.34$. Due to the symmetry intrinsic to the boundary value problem, only $1/8$ of the RVE as depicted

$\mathrm{E}_{\mathrm{Matr}}$ [MPa]	ν_{Matr}	$\sigma^{\mathrm{y}}_{0,\mathrm{Matr}}$ [MPa]	H [MPa]	$\mathrm{R}_{\infty,\mathrm{Matr}}$[MPa]	$\mathrm{f}_{1,\mathrm{Matr}}$	$\mathrm{f}_{2,\mathrm{Matr}}$
200000	0.3	548	1800	550	1.65	1.1

Table 2: Material parameters for the ferritic/bainitic matrix

in fig. 3a is simulated. The overall true stress-strain curve is shown in fig. 3b for the cases of a transforming and a non-transforming retained austenite inclusion. The latter case corresponds to conditions of higher ambient temperatures where the retained austenite is stabilized leading to negligible transformation rates. Although the increase in the overall stress due to transformation is less than 100 MPa, the ductility is enhanced substantially ($+0.05$ true strain). The Considère criterion is employed as a measure of ductility, which estimates the limit of uniform elongation under uniaxial tension by the intersection of the flow curve and the instantaneous hardening modulus, i.e. $\bar{\sigma} = \frac{\partial \bar{\sigma}}{\partial \bar{\varepsilon}}$. The increase in ductility cannot be attributed solely to the transformation strains, but rather to the change in the yield behavior, because the inclusion has only transformed partly (see fig. 3b), indicating a dynamic stress and strain redistribution during

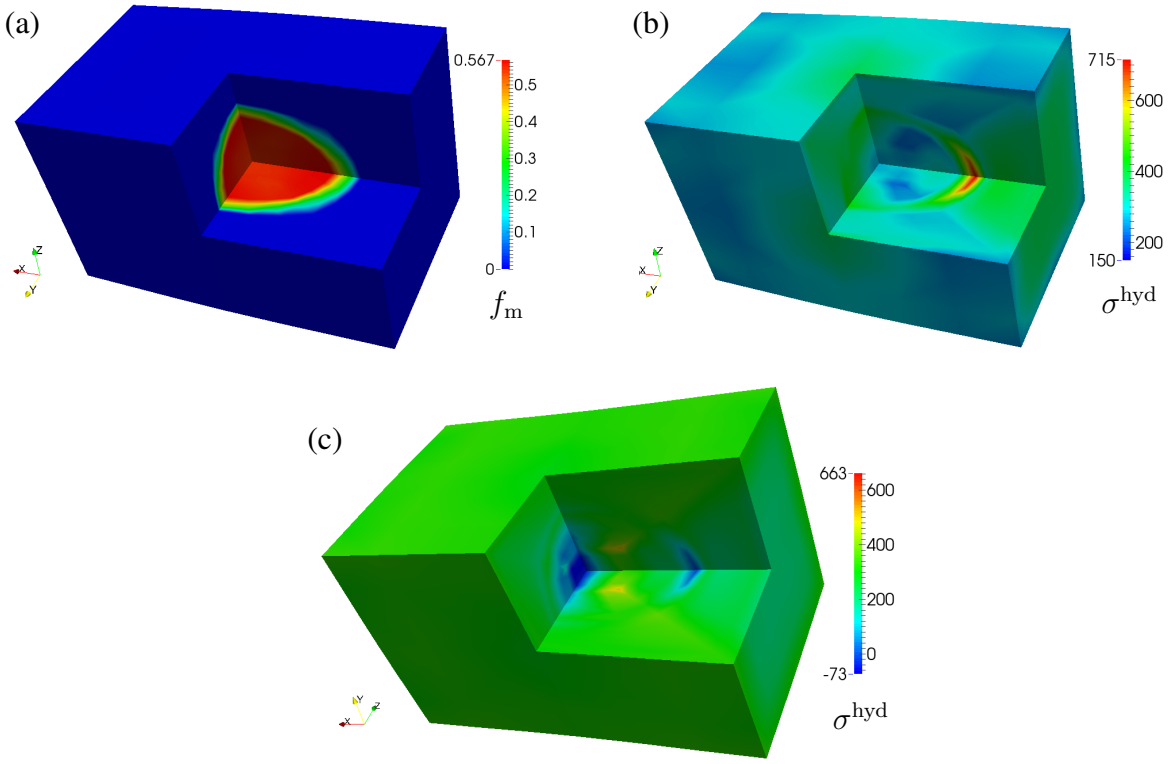


Figure 4: Distribution of martensite volume fraction (a) and hydrostatic Cauchy stress: (b) with and (c) without a→m transformation at the micro-scale under macroscopic uniaxial tension ($\bar{\varepsilon} = 0.34$)

loading.

Also on the micro-scale, significantly different stress distributions are obtained. Comparing fig. 4b and c, one observes higher hydrostatic stresses in the matrix close to the transforming inclusion, revealing a potential location of ductile damage initiation, which is absent in the microstructure with non-transforming retained austenite. Here the highest hydrostatic stress is observed in the inclusion.

5 CONCLUSIONS

In this contribution a simple material model for the austenite to martensite phase transformation in high-alloyed TRIP-steels is adopted for the direct micro-macro simulation of multiphase, low-alloyed TRIP-steels. Minor modifications of the model are proposed to improve its implementation and performance. A parameter study is conducted to illustrate the influence of material parameters, such as the initial transformation barrier, transformation hardening modulus and a shape parameter of the transformation surface on the phase transformation kinetics under different homogeneous loading conditions. Comparing the simulated evolution of retained austenite to experiments, it is found that the non-monotonic stress-state dependence can be reproduced by the model. For a reasonable choice of the material parameters of the single phases, a direct micro-macro simulation of a low-alloyed TRIP-steel is carried out employing a simplified, periodic microstructure. It is observed that austenite to martensite phase transformation enhances the ductility through a dynamic stress and strain redistribution between the constituents.

REFERENCES

[1] P.J. Jacques, *Phase transformation in transformation induced plasticity (TRIP)-assisted multiphase steels, in: Phase Transformations in Steels, Vol. 2*. Woodhead Publishing, 2012.

[2] H.K.D.H. Bhadeshia, TRIP-assisted Steels?, *ISIJ International*, **42**, 1059-1060, 2002.

[3] P.J. Jacques, Transformation-induced plasticity for high strength formable steels, *Current Opinion in Solid State and Materials Science*, **8**, 259–265, 2004.

[4] O. Muránsky, P. Šittner, J. Zrník, E.C. Oliver, In situ neutron diffraction investigation of the collaborative deformation-transformation mechanism in TRIP-assisted steels at room and elevated temperatures, *Acta Materialia*, **56**, 3367–3379, 2008.

[5] R. Petrov, L. Kestens, A. Wasilkowska, Y. Houbaert, Microstructure and texture of a lightly deformed TRIP-assisted steel characterized by means of the EBSD technique, *Materials Science and Engineering: A*, **447**, 285–297, 2007.

[6] I. de Diego-Calderón, M.J. Santofimia, J.M. Molina-Aldareguia, M.A. Monclús, I. Sabirov, Deformation behavior of a high strength multiphase steel at macro- and micro-scales, *Materials Science and Engineering: A*, **611**, 201–211, 2014.

[7] M. Radu and J. Valy and A.F. Gourgues and F. Le Strat and A. Pineau, Continuous magnetic method for quantitative monitoring of martensitic transformation in steels containing metastable austenite, *Scripta Materialia*, **52**, 525–530, 2005.

[8] P.J. Jacques, Q. Furnémont, F. Lani, T. Pardoen, F. Delannay, Multiscale mechanics of TRIP-assisted multiphase steels: I. Characterization and mechanical testing, *Acta Materialia*, **55**, 3681–3693, 2007.

[9] A.F. Mark, X. Wang, E. Essadiqi, J.D. Embury, J.D. Boyd, Development and characterisation of model TRIP steel microstructures, *Materials Science and Engineering: A*, **576**, 108–117, 2013.

[10] G.N. Haidemenopoulos, N. Aravas, I. Bellas, Kinetics of strain-induced transformation of dispersed austenite in low-alloy TRIP steels, *Materials Science and Engineering: A*, **615**, 416–423, 2014.

[11] R. Mahnken, A. Schneidt, T. Antretter, Macro modelling and homogenization for transformation induced plasticity of a low-alloy steel, *International Journal of Plasticity*, **25**, 183–204, 2009.

[12] R. Mahnken, M. Wolff, A. Schneidt, M. Böhm, Multi-phase transformations at large strains Thermodynamic framework and simulation, *International Journal of Plasticity*, **39**, 1–26, 2012.

[13] M. Mukherjee, T. Bhattacharyya, S.B. Singh, Models for Austenite to Martensite Transformation in TRIP-Aided Steels: A Comparative Study, *Materials & Manufacturing Processes*, **25**, 206–210, 2010.

[14] W.J. Dan, W.G. Zhang, S.H. Li, Z.Q. Lin, Finite element simulation on strain-induced martensitic transformation effects in TRIP steel sheet forming, *Computational Materials Science*, **39**, 593–599, 2007.

[15] L. Delannay, P. Jacques, T. Pardoen, Modelling of the plastic flow of TRIP-aided multiphase steel based on an incremental mean-field approach, *International Journal of Solids and Structures*, **45**, 1825–1843, 2008.

[16] R.F. Kubler, M. Berveiller, P. Buessler, Semi phenomenological modelling of the behavior of TRIP steels, *International Journal of Plasticity*, **27**, 299–327, 2011.

[17] J.-Y. Liu, H. Lu, J.-M. Chen, J.-F. Jullien, T. Wu, Simulation of mechanical behavior of multiphase TRIP steel taking account of transformation-induced plasticity, *Computational Materials Science*, **43**, 646–654, 2008.

[18] I. Papatriantafillou, M. Agoras, N. Aravas, G. Haidemenopoulos, Constitutive modeling and finite element methods for TRIP steels, *Computational Methods in Applied Mechanics and Engineering*, **195**, 5094–5114, 2006.

[19] I.O. Papatriantafillou, *TRIP steels: constitutive modeling and computational issues*. PhD-thesis, 2005.

[20] T.K. Shan, S.H. Li, W.G. Zhang, Z.G. Xu, Prediction of martensitic transformation and deformation behavior in the TRIP steel sheet forming, *Materials & Design*, **29**, 1810–1816, 2008.

[21] K.S. Choi, W.N. Liu, X. Sun, M.A. Khaleel, Microstructure-based constitutive modeling of TRIP steel: Prediction of ductility and failure modes under different loading conditions, *Acta Materialia*, **57**, 2592–2604, 2009.

[22] R. Sierra, J.A. Nemes, Investigation of the mechanical behaviour of multi-phase TRIP steels using finite element methods, *International Journal of Mechanical Sciences*, **50**, 649–665, 2008.

[23] D.D. Tjahjanto, S. Turteltaub, A.S.J. Suiker, S. van der Zwaag, Transformation-induced plasticity in multiphase steels subjected to thermomechanical loading, *Philosophical Magazine*, **88**, 3369–3387, 2008.

[24] R. Hill, On constitutive macro-variables for heterogeneous solids at finite strain, *Proceedings of the Royal Society A*, **326**, 131–147, 1972.

[25] H. Hallberg, P. Håkansson, M. Ristinmaa, A constitutive model for the formation of martensite in austenitic steels under large strain plasticity, *International Journal of Plasticity*, **23**, 1213–1239, 2007.

[26] M. Budnitzki, M. Kuna, A thermomechanical constitutive model for phase transformations in silicon under pressure and contact loading conditions, *International Journal of Solids and Structures*, **49**, 1316–1324, 2012.

[27] J. Besson, G. Cailletaud, J.-L. Chaboche, S. Forest, *Non-Linear Mechanics of Materials*. Springer, 2010.

[28] F. Auricchio and R.L. Taylor, A return-map algorithm for general associative isotropic elasto-plastic materials in large deformation regimes, *International Journal of Plasticity*, **15**, 1999.

[29] J.C. Simo, T.J.R. Hughes, *Computational Inelasticity*. Springer, 1998.

[30] D. Balzani, L. Scheunemann, D. Brands, J. Schröder, Construction of two- and three-dimensional statistically similar RVEs for coupled micro-macro simulations, *Computational Mechanics*, **54**, 2014.

[31] D. Brands, D. Balzani, L. Scheunemann, J. Schröder, H. Richter, D. Raabe, Computational modeling of dual-phase steels based on representative three-dimensional microstructures obtained from EBSD data, *Archive of Applied Mechanics*, **86**, 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

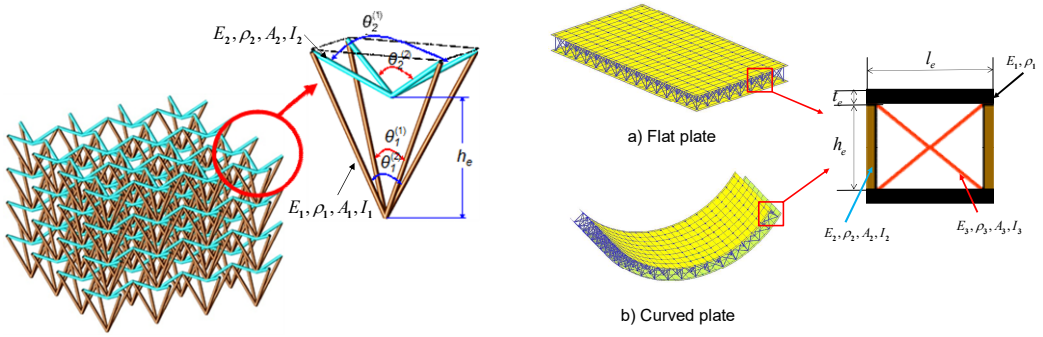
HOMOGENIZATION METHOD FOR DESIGNING NOVEL ARCHITECTURED CELLULAR MATERIALS

Zheng-Dong Ma[1]

[1] The University of Michigan
Ann Arbor, Michigan, USA
e-mail: mazd@umich.edu

Keywords: Homogenization Method, Discrete Homogenization Method, Cellular Materials, Architectured Materials, Coordinate Reduction, Computational Method.

Abstract. *Two types of novel architectured cellular materials have been developed [1,2]. To estimate the effective material properties and to optimally design such materials, a suitable homogenization method was needed. An existing asymptotic homogenization process has provided an effective means for the multiscale modeling of continuum solids; however, that method was not easily adapted to the aforementioned materials, which are naturally discrete systems. First, we developed a strain-based homogenization method equivalent to the existing asymptotic homogenization method, but it was developed based on an engineering approach rather than on a mathematical approach such as the one used in asymptotic homogenization. The new approach separates the strain field into a homogenized strain field and a strain variation field in the local cellular domain superposed on the homogenized strain field. The Principle of Virtual Displacements for the relationship between the strain variation field and the homogenized strain field is then used to condense the strain variation field onto the homogenized strain field, and the homogenization process becomes a coordinate reduction process comparable to the Guyan Reduction used in structural dynamics analyses. The characteristic modes and the stress recovery process are also discussed. The new method is then extended to a stress-based homogenization process based on the Principle of Virtual Forces, and it is further applied to address the discrete systems of the aforementioned architectured cellular materials.*



3D NPR material [1] b) BTR material [2]

Figure 1: Novel architectured cellular materials.

1 INTRODUCTION

Multiscale modeling and homogenization have become rich areas of research. Fruitful results have been published with a wide range of applications, as listed in recent literature review papers [3-5]. Homogenization, especially asymptotic homogenization [6-8], provides a powerful mathematical tool for bridging different scale modeling problem, and solving micro-macro, local-global, nano-macro, etc., multiscale modeling problems. The applicable areas include so-called heterogeneous materials, cellular materials, granular materials, fiber-reinforced polymers, etc. The homogenization method has also been utilized in topology optimization processes [9-11].

Two types of useful architectured cellular materials have been developed in the literature [1,2]. One type is three-dimensional Negative Poisson's Ratio (NPR) materials [1] (shown in Fig. 1a), and the other is so-called Biomimetic Tendon-Reinforced (BTR) materials [2] (shown in Fig. 1b). Both are originally obtained through the topology optimization process developed in [9]. These new materials are similar to traditional lattice materials, but they can be made of multiple raw materials and cannot be referred to as "micro" because of their relatively larger sizes. To estimate the effective material properties and to optimally design such materials, a suitable homogenization method was needed.

The advantages in the mathematical process of asymptotic homogenization should be appreciated; however, they are often overlooked in favor of the "mechanics logics" inside the homogenization modeling process. Existing asymptotic homogenization methods have provided effective means for the multiscale modeling of continuum solids; however, there is a need for extending them to more general applications such as handling the discrete systems discussed in this paper.

This paper provides a view of the existing homogenization process that is based on engineering and mechanics rather than on mathematics. Furthermore, most discoveries in this paper actually exist elsewhere in various previous publications [12-16]; however, it is useful to provide a more systematic explanation for mechanics-based homogenization. Another important point is that theoretically the mathematical asymptotic homogenization process requires the micro-cell to be small or infinitely small to assume convergence of the process, but this is not necessary for mechanics-based homogenization. In fact, mechanics-based homogenization can be considered a coordinate reduction process like the *Guyan Reduction* [17], which is widely used in solving structural dynamics problems. Therefore, relatively larger size cells can be treated in various applications, including in architectured cellular materials.

First, continuum solids are considered, and a mechanics-based homogenization process is developed based on the *Principle of Virtual Displacements*. Although the resultant formulations for the homogenized material stiffness matrix are the same as those obtained from the asymptotic homogenization method, mechanics-based homogenization provides a better understanding in terms of the mechanics behind the homogenization process. The new approach separates the strain field into a homogenized strain field along with a strain variation field in the local cellular domain superposed on the homogenized strain field. The *Principle of Virtual Displacements* describing the relationship between the strain variation field and the homogenized strain field is then used to condense the strain variation field to the homogenized strain field. Hence, the homogenization process is treated as a coordinate reduction process like the *Guyan Reduction*.

New formulas are also obtained to calculate the effective mass density and body forces with improved approximations, which can be further extended to consider other body forces such as thermal and magnetic forces. The characteristic stain modes and corresponding characteristic displacement modes are discussed. A recovery process is further considered for recovering the local strain and stress in the cellular domain after the global analysis.

The new method can have two variants: one is strain-based homogenization, and the other is stress-

based homogenization. The stress-based homogenization process is obtained based on the *Principle of Virtual Forces*. Similar to strain-based homogenization, strain-based homogenization separates the stress field into a homogenized stress field and a stress variation field superposed on the homogenized stress field. The principle of virtual forces is then used to condense the stress variation field into the homogenized stress field. The stress-based homogenization process can be extended easily to address discrete systems. The new method is then utilized to obtain the effective material properties of the architectured cellular materials discussed in this paper.

2 THE ASYMPTOTIC HOMOGENIZATION METHOD

An asymptotic homogenization method [3-5] for an elastostatic continuum is derived using the mathematical approach described below. First, supposing that the structure in a domain Ω_ε has a Y-periodic microstructure, the state equation can be written as:

$$\int_{\Omega_\varepsilon} E^\varepsilon_{ijkl} \frac{\partial u_i^\varepsilon}{\partial y_j} \frac{\partial v_k}{\partial y_l} d\Omega + \int_{\Omega_\varepsilon} f_i v_i d\Omega + \int_{\Gamma_t} t_i v_i d\Gamma = 0 \quad \text{for } \forall v \in V . \tag{1}$$

The periodic microstructure is assumed to be near to an arbitrary point x of a given linearly elastic structure. The periodicity is represented by a parameter ε, which is very small, and the elastic tensor is given in the form of

$$E^\varepsilon_{ijkl} = E_{ijkl}\left(x, \frac{x}{\varepsilon}\right), \tag{2}$$

where $y \to E_{ijkl}(x,y)$ is Y-periodic, x is the macroscopic variation of the material parameters, and $y = x/\varepsilon$ gives the microscopic, periodic variations. Now, supposing that the structure is subjected to a macroscopic body force and a macroscopic surface traction, the resulting displacement field $u^\varepsilon(x)$ can then be expanded as

$$u^\varepsilon(x) = u_0(x) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon^2 u_2\left(x, \frac{x}{\varepsilon}\right) + \cdots, \tag{3}$$

where the leading term $u_0(x)$ is a macroscopic deformation field that is independent of the microscopic variable. Substituting Eqs. (2) and (3) into Eq. (1) and comparing the terms in the same order of ε, one can obtain the effective rigidity tensor

$$E^H_{ijkl}(x) = \frac{1}{|Y|}\int_Y \left[E_{ijkl}(x,y) - E_{ijpq}(x,y)\frac{\partial \chi_p^{kl}}{\partial y_q} \right] dy . \tag{4}$$

Here, χ^{kl} is a microscopic displacement field that is given as the Y-periodic solution of the cell problem in the weak form

$$\int_Y E_{ijpq} \frac{\partial \chi_p^{kl}}{\partial y_q} \frac{\partial v_i(y)}{\partial y_j} dy = \int_Y E_{ijkl} \frac{\partial v_i(y)}{\partial y_j} dy . \tag{5}$$

The homogenized equation of Eq. (1) becomes

$$\int_\Omega E^H_{ijkl} \frac{\partial u_i}{\partial y_j} \frac{\partial v_k}{\partial y_l} d\Omega + \int_\Omega f_i^H v_i d\Omega + \int_{\Gamma_t} t_i v_i d\Gamma = 0 \quad \text{for } \forall v \in V , \tag{6}$$

where f_i^H is an average of the force f_i on the cellular domain.

$$f_i^H = \frac{1}{|Y|}\int_Y f_i \, dy \tag{7}$$

Note that the above approach is obtained based on the assumption of infinitely small cellular cells; however, it does not explain why, in many cases, one cannot obtain better results by considering higher order terms of ε, which indicates that the Taylor expansion in Eq. (3) may not generally converge. The approach also has a theoretical limitation in the extension to larger cells. It lacks explicit instructions for how to determine the boundary conditions in a micro-cell analysis, with the exception of mentioning the use of the periodic boundary conditions. In fact, the periodic boundary conditions may not be able to remove the rigid body motion of the micro-cell, and different boundary conditions may be required to solve different subcases in the homogenization process, which will be discussed in Sec. 4.2.

3 HOMOGENIZATION AS A COORDINATE REDUCTION PROCESS

3.1 Problem Definition

The elastostatics problem in Eq. (1) is converted into a form of Principle of Virtual Displacements and extended to an elastic-dynamics problem. Ω_ε ($\Omega_\varepsilon \subset R^3$) is assumed to be the structural domain with a cellular structure that is in periodic or other forms, as long as the proper connections among the adjacent cells can be defined; Γ_t is the traction boundary of Ω_ε. The equilibrium of these types of cellular materials can be formulated by the *Principle of Virtual Displacements*:

$$\int_{\Omega_\varepsilon} \delta\boldsymbol{\varepsilon}^T \mathbf{D}^\varepsilon \boldsymbol{\varepsilon} d\Omega + \int_{\Omega_\varepsilon} \delta\mathbf{u}^T \rho^\varepsilon \ddot{\mathbf{u}} d\Omega = \int_{\Omega_\varepsilon} \delta\mathbf{u}^T \mathbf{f}^\varepsilon d\Omega + \int_{\Gamma_t} \delta\mathbf{u}^T \mathbf{t} d\Gamma_t , \tag{8}$$

where $\boldsymbol{\varepsilon}$ denotes the *engineering strain* vector ($\boldsymbol{\varepsilon}^T = \left\{\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{yz}, \gamma_{zx}, \gamma_{xy}\right\}^T$), $\mathbf{u}$ denotes the displacement vector, $\ddot{\mathbf{u}} = \dfrac{\partial^2 \mathbf{u}}{\partial t^2}$ denotes the acceleration vector, $\mathbf{f}^\varepsilon$ denotes the body force vector, and $\mathbf{t}$ is the boundary traction vector; $\delta\boldsymbol{\varepsilon}$ denotes the virtual stain vector, $\delta\mathbf{u}$ denotes the virtual displacement vector, $V = \left\{\delta\mathbf{u} \in \mathrm{H}^1(\Omega_\varepsilon) \,|\, \delta\mathbf{u}\big|_{\Gamma_d} = 0\right\}$ denotes the space of kinematically admissible displacement fields, and Γ_d is the displacement boundary. $\mathrm{H}^1(\Omega_s)$ is the Sobolev space in which the strain energy of the structure is finite. $\mathbf{D}^\varepsilon$ denotes the elasticity matrix before the homogenization, and ρ^ε is the density before the homogenization.

For simplicity, in the following, we assume a three-dimensional solids problem, even though the formulation obtained can be applied to two-dimensional solids problems and problems such as beams, plates, or shells. We also assume a linear elasticity problem with a constitute law $\boldsymbol{\sigma} = \mathbf{D}\boldsymbol{\varepsilon}$, where $\boldsymbol{\sigma}$ is a Cauchy stress vector ($\boldsymbol{\sigma}^T = \left\{\sigma_x, \sigma_y, \sigma_z, \tau_{yz}, \tau_{zx}, \tau_{xy}\right\}^T$), and $\mathbf{D} = \left[D_{ij}\right]_{6\times 6}$ is the material stiffness matrix, even though the formulation obtained can be extended easily to nonlinear problems.

Note, the *Principle of Virtual Displacements* requires the following continuity conditions to be satisfied:

$$\boldsymbol{\varepsilon} = \mathbf{E}(\nabla)\mathbf{u}, \;\; \delta\boldsymbol{\varepsilon} = \mathbf{E}(\nabla)\delta\mathbf{u} \tag{9}$$

where

$$\mathbf{E}(\nabla)=\begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 & 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ 0 & \frac{\partial}{\partial y} & 0 & \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} \\ 0 & 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \end{bmatrix}^{\mathbf{T}} \tag{10}$$

3.2 Homogenized Process as a Coordinate Reduction

The asymptotic homogenization method described in Sec. 2 is derived from a mathematical treatment. Now, we employ an approach based on an engineering approach and the mechanics of the problem. We consider mechanics-based homogenization as a coordinate reduction process that is analogous to the *Guyan Reduction* described in Appendix A.

Instead of separating the nodal coordinates into internal coordinates and boundary coordinates as in the *Guyan Reduction*, we separate the strain field $\boldsymbol{\varepsilon}$ into a homogenized strain field $\boldsymbol{\varepsilon}_h$ that is constant over the cellular domain and continuous over the homogenized structural domain, and a strain variation field $\boldsymbol{\varepsilon}_v = \boldsymbol{\varepsilon}_v(y)$ ($y \in Y$) superposed on $\boldsymbol{\varepsilon}_h$, which is defined in cellular Y domain and varies in Y, to obtain

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_h + \boldsymbol{\varepsilon}_v . \tag{11}$$

The displacement fields corresponding to $\boldsymbol{\varepsilon}_h$ and $\boldsymbol{\varepsilon}_v$ are denoted as $\mathbf{u}_h$ and $\mathbf{u}_v$, and we get

$$\mathbf{u} = \mathbf{u}_h + \mathbf{u}_v , \tag{12}$$

as well as the following relationships:

$$\boldsymbol{\varepsilon}_h = \mathbf{E}(\nabla)\mathbf{u}_h \quad (in\ \Omega) \quad \textbf{and} \quad \boldsymbol{\varepsilon}_v = \mathbf{E}(\nabla)\mathbf{u}_v \quad (in\ Y) . \tag{13}$$

Substituting Eq. (11) into Eq. (8) produces

$$\int_{\Omega_\varepsilon} \left(\delta\boldsymbol{\varepsilon}_h + \delta\boldsymbol{\varepsilon}_v\right)^T \mathbf{D}^\varepsilon \left(\boldsymbol{\varepsilon}_h + \boldsymbol{\varepsilon}_v\right) d\Omega + \int_{\Omega_\varepsilon} \delta\mathbf{u}^T \rho^\varepsilon \ddot{\mathbf{u}} d\Omega = \int_{\Omega_\varepsilon} \delta\mathbf{u}^T \mathbf{f}^\varepsilon d\Omega + \int_{\Gamma_t} \delta\mathbf{u}^T \mathbf{t} d\Gamma_t . \tag{14}$$

Equation (14) can be separated as

$$\begin{cases} \displaystyle\int_{\Omega_\varepsilon} \delta\boldsymbol{\varepsilon}_h^T \mathbf{D}^\varepsilon \left(\boldsymbol{\varepsilon}_h + \boldsymbol{\varepsilon}_v\right) d\Omega + \int_{\Omega_\varepsilon} \delta\mathbf{u}^T \rho^\varepsilon \ddot{\mathbf{u}} d\Omega = \int_{\Omega_\varepsilon} \delta\mathbf{u}^T \mathbf{f}^\varepsilon d\Omega + \int_{\Gamma_t} \delta\mathbf{u}^T \mathbf{t} d\Gamma_t \\ \displaystyle\int_{Y} \delta\boldsymbol{\varepsilon}_v^T \mathbf{D}^\varepsilon \left(\boldsymbol{\varepsilon}_h + \boldsymbol{\varepsilon}_v\right) dy = 0 \qquad (\forall Y \subset \Omega_\varepsilon) \end{cases} . \tag{15}$$

Notably, if both equations in Eq. (15) are satisfied while satisfying the continuity conditions in Eq. (13) and the connectivity among adjacent cells, then the original equation, Eq. (14) (i.e., Eq. (8)), will be satisfied. Here, theoretically, we do not have to assume that the cellular Y domain is small.

To solve the second equation in Eq. (15), we represent the strain variation $\boldsymbol{\varepsilon}_v = \boldsymbol{\varepsilon}_v(y)$ with a mode

superposition technique; whereas, in Eq. (16), $\boldsymbol{\varphi}(y)$ is a matrix constructed of d modes (d=6 for the three-dimensional solids, d=3 for the two-dimensional solids, and d can be other numbers for, e.g., beams, plates, and shells problems) with $\boldsymbol{\varepsilon}_h$ as the modal coordinates, namely

$$\boldsymbol{\varepsilon}_v(y) = \boldsymbol{\varphi}(y)\boldsymbol{\varepsilon}_h \quad (y \in Y). \tag{16}$$

Note, Eq. (16) is an equivalent transformation; therefore, there is no error induced as long as the modes in $\boldsymbol{\varphi}$ are linearly independent. Substituting Eq. (16) into the second row equation in Eq. (15) results in

$$\left[\int_Y \delta\boldsymbol{\varepsilon}_v^T \mathbf{D}^\varepsilon \left(\mathbf{I}+\boldsymbol{\varphi}\right) dy\right]\boldsymbol{\varepsilon}_h = 0 \tag{17}$$

or

$$\int_Y \delta\boldsymbol{\varepsilon}_v^T \mathbf{D}^\varepsilon \left(\mathbf{I}+\boldsymbol{\varphi}\right) dy = 0\,. \tag{18}$$

Solving Eq. (18) with the continuity conditions in Eq. (13) and the properly defined boundary conditions (periodic boundary conditions for now), we obtain $\boldsymbol{\varphi}(y)$ ($y \in Y$), and $\boldsymbol{\varepsilon}_v$ now becomes a function of $\boldsymbol{\varepsilon}_h$. We now assume that the displacement field corresponding to $\boldsymbol{\varepsilon}_v$ can be obtained as

$$\mathbf{u}_v = \boldsymbol{\theta}(\mathbf{y})\mathbf{u}_h \quad (y \in Y), \tag{19}$$

where $\boldsymbol{\theta}(\mathbf{y})$ satisfies

$$\boldsymbol{\varphi}(y) = \mathbf{E}(\nabla)\boldsymbol{\theta}(y) \quad (y \in Y). \tag{20}$$

Then, substituting Eqs. (16) and (19) into the first row equation in Eq. (15) results in

$$\int_\Omega \delta\boldsymbol{\varepsilon}_h^T \mathbf{D}^H \boldsymbol{\varepsilon}_h d\Omega + \int_\Omega \delta\mathbf{u}_h^T \rho^H \ddot{\mathbf{u}}_h d\Omega = \int_\Omega \delta\mathbf{u}_h^T \mathbf{f}^H d\Omega + \int_{\Gamma_t} \delta\mathbf{u}_h^T \mathbf{t}^H d\Gamma_t \tag{21}$$

where, Ω stands for the homogenized domain of Ω_ε, and

$$\mathbf{D}^H = \frac{1}{|Y|}\int_Y \mathbf{D}^\varepsilon(\mathbf{I}+\boldsymbol{\varphi})dy \tag{22}$$

$$\boldsymbol{\rho}^H = \frac{1}{|Y|}\int_Y \left(\mathbf{I}+\boldsymbol{\theta}(\mathbf{y})\right)^T \rho^\varepsilon \left(\mathbf{I}+\boldsymbol{\theta}(\mathbf{y})\right) dy \tag{23}$$

$$\mathbf{f}^H = \frac{1}{|Y|}\int_Y \left(\mathbf{I}+\boldsymbol{\theta}(\mathbf{y})\right)^T \mathbf{f}_v^\varepsilon dy \tag{24}$$

$$\mathbf{t}^H = \frac{1}{|L|}\int_{\Gamma_{tY}} \left(\mathbf{I}+\boldsymbol{\theta}(\mathbf{y})\right)^T \mathbf{t} ds \tag{25}$$

where, $\Gamma_{tY} = \Gamma_t \bigcap Y$.

Notably, the homogenized material stiffness matrix in Eq. (22) is exactly the same as in Eq. (4) obtained from the asymptotic homogenization method; however, the effective mass density in Eq. (23), the effective internal force in Eq. (24), and the effective traction in Eq. (25) are different from the tra-

ditional asymptotic homogenization process, which may provide better approximations for these quantities.

3.3 Solution using a Finite Element Method

By following a standard finite element approach and assuming an assembly of nodal displacement vectors, $\boldsymbol{\chi}=\left[\boldsymbol{\chi}^{(1)},\boldsymbol{\chi}^{(2)},...,\boldsymbol{\chi}^{(d)}\right]$ and a shape function $\mathbf{N}=\mathbf{N}(y)$ in Y, $\boldsymbol{\theta}(y)$ in Eq. (19) can be represented as

$$\boldsymbol{\theta}=\mathbf{N}(y)\boldsymbol{\chi}, \tag{26}$$

and we can have

$$\boldsymbol{\varphi}=\mathbf{B}_Y\boldsymbol{\chi} \quad \text{and} \quad \delta\boldsymbol{\varepsilon}_1=\mathbf{B}_Y\delta\boldsymbol{\xi}, \tag{27}$$

where

$$\mathbf{B}_Y=\mathbf{E}\left(\nabla\right)\mathbf{N}(y). \tag{28}$$

Substituting Eq. (27) into Eq. (18) gives

$$\int_Y \delta\boldsymbol{\xi}^T\left(\mathbf{B}_Y^T\mathbf{D}^\varepsilon\mathbf{B}_Y\right)\boldsymbol{\chi}dy+\int_Y\delta\boldsymbol{\xi}^T\mathbf{B}_Y^T\mathbf{D}^\varepsilon dy=0 \ . \tag{29}$$

Equation (29) results in the following finite element equations:

$$\mathbf{K}_Y\boldsymbol{\chi}^{(i)}=\mathbf{F}_Y^{(i)} \ \left(i=1,2,\cdots,d\right), \tag{30}$$

where

$$\mathbf{K}_Y=\int_Y\left(\mathbf{B}_Y^T\mathbf{D}^\varepsilon\mathbf{B}_Y\right)dy, \tag{31}$$

and $\mathbf{F}_Y^{(i)}$ is the *i*-th column in $\mathbf{F}_Y$ and

$$\mathbf{F}_Y=-\int_Y\mathbf{B}_Y^T\mathbf{D}^\varepsilon dy. \tag{32}$$

Solving Eq. (30) for $\boldsymbol{\chi}=\left[\boldsymbol{\chi}^{(1)},\boldsymbol{\chi}^{(2)},...,\boldsymbol{\chi}^{(d)}\right]$ gives

$$\boldsymbol{\chi}=\mathbf{K}_Y^{-1}\mathbf{F}_Y=-\left(\int_Y\mathbf{B}_Y^T\mathbf{D}^\varepsilon\mathbf{B}_Y dy\right)^{-1}\left(\int_Y\mathbf{B}_Y^T\mathbf{D}^\varepsilon dy\right), \tag{33}$$

and then the homogenized material stiffness matrix can be obtained as

$$\mathbf{D}^H=\frac{1}{|Y|}\int_Y\mathbf{D}^\varepsilon\left(\mathbf{I}+\mathbf{B}_Y\boldsymbol{\chi}\right)dy. \tag{34}$$

Note, it is interesting to compare Eq. (33) to Eq. (A-5) in the *Guyan Reduction* to see the similarities, as well as the similarity of the homogenization process with the *Guyan Reduction*.

3.4 Characteristic Strain Modes and Deformation Modes

Notably, $\boldsymbol{\varphi}$ in Eq. (16) is a matrix constructed of *d* vectors (e.g., *d*=6 for three-dimensional solids),

namely $\boldsymbol{\varphi}=\left[\boldsymbol{\varphi}^{(1)},\boldsymbol{\varphi}^{(2)},\cdots,\boldsymbol{\varphi}^{(d)}\right]$, where each $\boldsymbol{\varphi}^{(i)}$ is labeled a *characteristic strain mode* of the cellular structure corresponding to a forced unique strain field applied over the cellular Y domain, where $i=1,\ 2,\ \cdots,\ d$. In fact, Eq. (16) can be rewritten as

$$\boldsymbol{\varepsilon}_v = \sum_{i=1}^{d} \boldsymbol{\varphi}^{(i)} \varepsilon_h^{(i)}, \tag{35}$$

where $\varepsilon_h^{(1)}=\varepsilon_x,\ \varepsilon_h^{(2)}=\varepsilon_y,\ \cdots,\ \varepsilon_h^{(d)}=\gamma_{xy}$, and $\boldsymbol{\varepsilon}_h$ can be rewritten using a strain coordinates system, $\mathbf{e}^{(i)}\ (i=1,2,\cdots,d)$, as

$$\boldsymbol{\varepsilon}_h = \sum_{i=1}^{d} \mathbf{e}^{(i)} \varepsilon_h^{(i)}. \tag{36}$$

Here, for $d=6$,

$$\mathbf{e}^{(1)}=\begin{Bmatrix}1\\0\\0\\0\\0\\0\end{Bmatrix},\ \mathbf{e}^{(2)}=\begin{Bmatrix}0\\1\\0\\0\\0\\0\end{Bmatrix},\ \cdots,\ \mathbf{e}^{(6)}=\begin{Bmatrix}0\\0\\0\\0\\0\\1\end{Bmatrix}. \tag{37}$$

By substituting Eqs. (35) and (36) into the second row equation in Eq. (16), the equation for each *characteristic strain mode* $\boldsymbol{\varphi}^{(i)}$ can be obtained:

$$\int_Y \delta\boldsymbol{\varepsilon}_v^T \mathbf{D}^{\varepsilon} \boldsymbol{\varphi}^{(i)} dy + \int_Y \delta\boldsymbol{\varepsilon}_v^T \mathbf{D}^{\varepsilon} \mathbf{e}^{(i)} dy = 0 \quad (i=1,\ 2,\ \cdots,\ d). \tag{38}$$

Equation (38) describes the physical meaning of the *characteristic strain mode* $\boldsymbol{\varphi}^{(i)}$, which is the response to an applied uniform unit strain field $\mathbf{e}^{(i)}$ in the cellular domain. As shown in the second term of Eq. (38), each uniform stain $\mathbf{e}^{(i)}$ results in a stress field, described as a pre-stress applied in the cellular domain, such that:

$$\boldsymbol{\sigma}_0^{(i)} = \mathbf{D}^{\varepsilon} \mathbf{e}^{(i)} \quad (i=1,\ 2,\ \cdots,\ d). \tag{39}$$

Therefore, Eq. (38) can also be rewritten as

$$\int_Y \delta\boldsymbol{\varepsilon}_v^T \mathbf{D}^{\varepsilon} \boldsymbol{\varphi}^{(i)} dy + \int_Y \delta\boldsymbol{\varepsilon}_v^T \boldsymbol{\sigma}_0^{(i)} dy = 0 \quad (i=1,\ 2,\ \cdots,\ d). \tag{40}$$

The *characteristic deformation modes* of the cell can be obtained by solving Eq. (30) with the following loading condition:

$$\mathbf{F}_Y^{(i)} = -\int_Y \mathbf{B}_Y^T \mathbf{D}^{\varepsilon} \mathbf{e}^{(i)} dy \quad (i=1,\ 2,\ \cdots,\ d). \tag{41}$$

Solving Eq. (30) with Eq. (41) for $\boldsymbol{\chi}^{(i)}$ results in

$$\boldsymbol{\chi}^{(i)} = -\left(\int_Y \mathbf{B}^T \mathbf{D}^{\varepsilon} \mathbf{B} dy\right)^{-1} \left(\int_Y \mathbf{B}^T \mathbf{D}^{\varepsilon} \mathbf{e}^{(i)} dy\right) \quad (i = 1,\ 2,\ \cdots,\ d). \tag{42}$$

As a special example, for a two-dimensional solid problem, we have $\boldsymbol{\varepsilon}^T = \left\{\varepsilon_x, \varepsilon_y, \gamma_{xy}\right\}^T$, $\mathbf{u}^T = \left\{u_x, u_y\right\}^T$, $\mathbf{f}^T = \left\{f_x, f_y\right\}^T$, $\mathbf{t}^T = \left\{t_x, t_y\right\}^T$, and $\boldsymbol{\varepsilon}_h$ can be written as

$$\boldsymbol{\varepsilon}_h = \sum_{i=1}^{3} \mathbf{e}^{(i)} \varepsilon_h^{(i)}, \tag{43}$$

where

$$\mathbf{e}^{(1)} = \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix},\ \mathbf{e}^{(2)} = \begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix},\ \mathbf{e}^{(3)} = \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix}. \tag{44}$$

Figure 2 illustrates the uniform unit strain fields in Eq. (44) applied in the two-dimensional cellular domain.

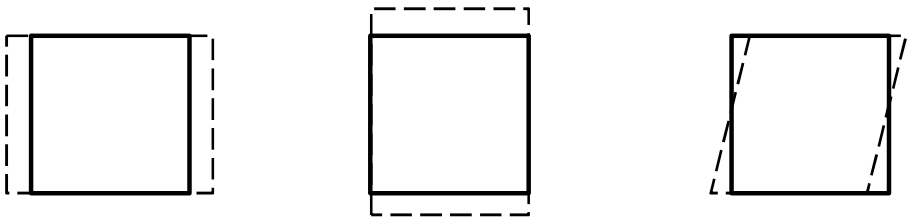


Figure 2: The unit uniform strain field $\mathbf{e}^{(i)}$ $(i = 1,2,3)$ applied in the 2-dimesional cellular domain.

Note, in Eq. (43), it is assumed that $\varepsilon_h^{(1)} = \varepsilon_x$, $\varepsilon_h^{(2)} = \varepsilon_y$, $\varepsilon_h^{(3)} = \gamma_{xy}$.

3.5 Strain and Stress Recoveries

When the global analysis is conducted and the homogenized strain $\boldsymbol{\varepsilon}_h$ is obtained, it is easy to recover the local strain in the cellular domain as

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_h + \boldsymbol{\varepsilon}_v = (\mathbf{I} + \boldsymbol{\varphi})\boldsymbol{\varepsilon}_h, \tag{45}$$

and therefore the local stress can be written as

$$\boldsymbol{\sigma} = \mathbf{D}^{\varepsilon}\boldsymbol{\varepsilon} = \mathbf{D}^{\varepsilon}(\mathbf{I} + \boldsymbol{\varphi})\boldsymbol{\varepsilon}_h = (\mathbf{I} + \boldsymbol{\varphi})\boldsymbol{\sigma}_h. \tag{46}$$

Note that these local strain and stress equations can be used to determine the failure and failure modes of cellular structures, or to predict the plastic deformation in the cellular domain. Furthermore,

the failure modes can be analyzed to determine the critical loads and to improve the local architecture design of the cellular material.

4 EXTENSION TO THE STRESS-BASED HOMOGENIZATION

4.1 Continuum Systems

The elastostatics problem in Eq. (8) can also be stated using the *Principle of Virtual Forces* as

$$\int_{\Omega_\varepsilon} \delta\boldsymbol{\sigma}^T \mathbf{A}^\varepsilon \boldsymbol{\sigma} d\Omega + \int_{\Gamma_u} \delta\mathbf{p}^T \bar{\mathbf{u}} d\Gamma = 0, \tag{47}$$

where $\boldsymbol{\sigma}$ ($\boldsymbol{\sigma}^T = \left\{\sigma_x, \sigma_y, \sigma_z, \tau_{yz}, \tau_{zx}, \tau_{xy}\right\}^T$ for 3-dimensional solids) stands for the stress vector satisfying

$$\mathbf{E}\left(\nabla\right)^T \boldsymbol{\sigma} = \mathbf{f} \ \text{ in } \Omega_\varepsilon \quad \text{and} \quad \mathbf{E}\left(\eta\right)^T \boldsymbol{\sigma} = \mathbf{t} \ \text{ on } \Gamma_u \ , \tag{48}$$

where $\mathbf{A}^\varepsilon$ is the material flexibility matrix; in general, we obtain

$$\mathbf{A}^\varepsilon = \left[\mathbf{D}^\varepsilon\right]^{-1}. \tag{49}$$

It is also assumed that $\delta\boldsymbol{\sigma}$ satisfies

$$\mathbf{E}\left(\nabla\right)^T \delta\boldsymbol{\sigma} = 0 \ \text{ in } \Omega_\varepsilon \quad \text{and} \quad \mathbf{E}\left(\nu\right)^T \delta\boldsymbol{\sigma} = \delta\mathbf{p} \ \text{ on } \Gamma_u \ . \tag{50}$$

Similar to the process described in Sec. 3.2, we assume

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_h + \boldsymbol{\sigma}_v, \tag{51}$$

where $\boldsymbol{\sigma}_h$ stands for the homogenized stress field, $\boldsymbol{\sigma}_v$ is the stress variation from the homogenized stress field in the cellular domain; also, we assume

$$\mathbf{p} = \mathbf{p}_h + \mathbf{p}_v, \tag{52}$$

where $\mathbf{p}_h$ stands for the boundary force corresponding to the homogenized stress field $\boldsymbol{\sigma}_h$, $\mathbf{p}_v$ stands for the boundary force corresponding to the stress variation $\boldsymbol{\sigma}_v$. Eq. (47) can be rewritten as

$$\begin{cases} \int_{\Omega_\varepsilon} \delta\boldsymbol{\sigma}_h^T \mathbf{A}^\varepsilon \left(\boldsymbol{\sigma}_h + \boldsymbol{\sigma}_v\right) d\Omega + \int_{\Gamma_u} \delta\left(\mathbf{p}_h + \mathbf{p}_v\right)^T \bar{\mathbf{u}} d\Gamma = 0 \\ \int_Y \delta\boldsymbol{\sigma}_v^T \mathbf{A}^\varepsilon \left(\boldsymbol{\sigma}_h + \boldsymbol{\sigma}_v\right) dy = 0 \qquad \textbf{for } \forall Y \in \Omega \end{cases}. \tag{53}$$

Similar to the process in Sec. 3.2, we assume

$$\boldsymbol{\sigma}_h = \sum_{i=1}^{d} \mathbf{e}_1^{(i)} \sigma_h^{(i)} \ \text{ and } \ \boldsymbol{\sigma}_v = \sum_{i=1}^{d} \mathbf{z}_1^{(i)} \sigma_h^{(i)}. \tag{54}$$

Substituting Eq. (54) into the second row equation in Eq. (53) results in

$$\int_Y \delta\boldsymbol{\sigma}_v^T \mathbf{A}^\varepsilon \mathbf{z}^{(i)} dy + \int_Y \delta\boldsymbol{\sigma}_v^T \mathbf{A}^\varepsilon \mathbf{e}_1^{(i)} dy = 0 \ \ (i = 1,\ 2,\ \cdots,\ d). \tag{55}$$

$\mathbf{z}_1^{(i)}$ $(i = 1, 2, \cdots, d)$ can be obtained by solving Eq. (55) with the properly defined (periodic) boundary conditions, and then the first row equation in Eq. (53) becomes

$$\int_{\Omega} \delta \boldsymbol{\sigma}_h^T \mathbf{A}^H \boldsymbol{\sigma}_h d\Omega + \int_{\Gamma_u} \delta \mathbf{p}_h^T \bar{\mathbf{u}}^H d\Gamma = 0\,, \tag{56}$$

where

$$\mathbf{A}^H = \frac{1}{|Y|}\int_Y \mathbf{A}^{\varepsilon}\left(\mathbf{I} + \mathbf{z}_1\right) dy \text{ and} \tag{57}$$

$$\bar{\mathbf{u}}^H = \frac{1}{|L|}\int_{\Gamma_{uY}} \left(\mathbf{I} + \boldsymbol{\omega}\right)^T \bar{\mathbf{u}} ds\,, \tag{58}$$

such that

$$\mathbf{E}\left(\nu\right)^T \mathbf{z}_1 = \boldsymbol{\omega} \;\; \textbf{on}\; \Gamma_{uY} \;\; and \;\; \Gamma_{uY} = \Gamma_u \bigcap Y\,. \tag{59}$$

Note, it is not necessary for the stress-based homogenization process to produce the same effective material properties as the strain-based homogenization process. In fact, based on the *Principle of Virtual Displacements* and *Principle of Virtual Forces*, the strain-based homogenization process may provide an upper boundary for the effective material properties; the stress-based homogenization process provides a lower bound for the effective material properties.

As a special example, for a two-dimensional solids problem, we have $\boldsymbol{\sigma}^T = \left\{\sigma_x, \sigma_y, \tau_{xy}\right\}^T$, and $\boldsymbol{\sigma}_h$ becomes

$$\boldsymbol{\sigma}_h = \sum_{i=1}^{3} \mathbf{e}^{(i)} \sigma_h^{(i)}\,. \tag{60}$$

Figure 3 illustrates the uniform unit stress fields in Eq. (61) applied in the two-dimensional cellular domain.

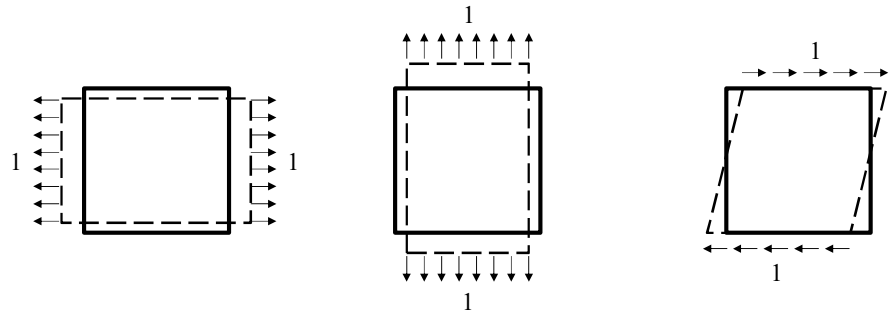


Figure 3: A unit uniform stress field $\mathbf{e}^{(i)}$ $(i = 1,2,3)$ applied on the 2-dimesional cellular domain.

Note that in Eq. (60), $\sigma_h^{(1)} = \sigma_x$, $\sigma_h^{(2)} = \sigma_y$, $\sigma_h^{(3)} = \tau_{xy}$ is assumed.

4.2 Extension to the Discrete Systems

With the above understanding, the extension to discrete systems can easily be performed by first, converting the applied unit stress field into concentrated forces and applying them to the proper nodes of the cellular structure; second, one must calculate the displacements at the properly selected nodes to approximate the strain in the cellular domain and use the obtained information to approximate the effective material properties. As an example, Fig. 4 uses a two-dimensional version of the NPR material shown in Fig. 1a to illustrate the loading conditions and boundary conditions for the homogenization of a 2D NPR cell structure.

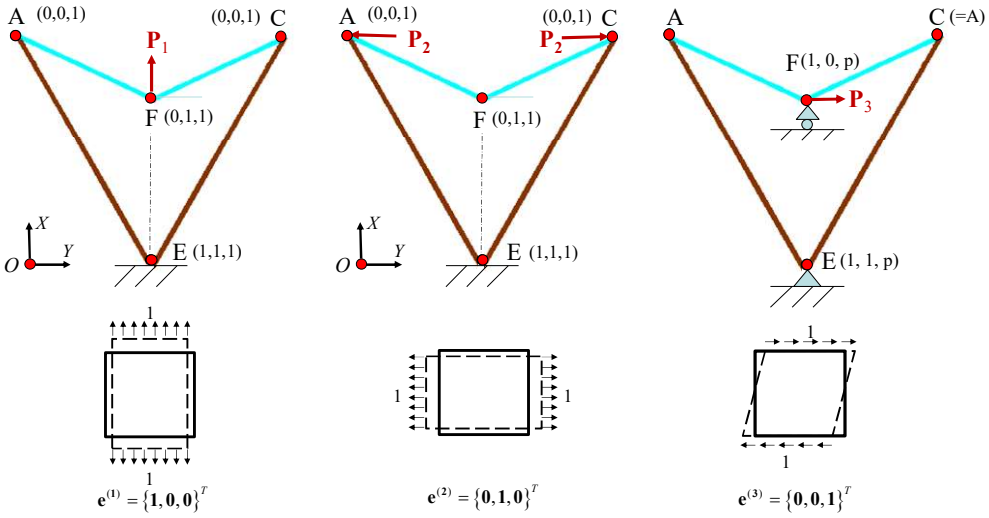


Figure 4: The loading and boundary conditions for the homogenization of a 2D NPR cell structure.

In Fig. 4, the vertical direction is set as the x-direction, and the horizontal direction is as the y-direction. As shown in Figure 4, there are three analysis cases in total corresponding to the three characteristic modes in the homogenization problem. Each loading case corresponds to a unit stress field, discussed in Sec. 4.1, with an amplitude equal to the facing area multiplied by the unit stress, which is 1. The boundary conditions are determined by the periodicity consideration as well as the symmetry of the cell structure and the connectivity to the adjacent cells.

As shown in the left figure in Fig. 4, a concentrated force P_1 is applied at Point F, which represents the tension stress on the top of the cell. Here, boundary condition E (1, 1, 1) indicates that node E is fixed along all of the directions, including the x- and y-directions as well as the deformation angle. Additionally, A (0, 0, 1) indicates that only the deformation angle is fixed at node A, and this is the same for node C.

As shown in the middle figure in Fig. 4, two concentrated forces with equal amplitudes P_2 are applied at Points A and C in opposite directions, representing the tension stress on the two sides of the cell. Here, the boundary conditions are the same as in the left figure.

The right figure in Fig. 4 illustrates a shear force applied on the cellular cell, for which C=A indicates the periodicity boundary conditions needing to be applied at nodes A and C, and F (1, 1, p) and E (1, 1, p) indicate the periodicity boundary condition for the deformation angle at nodes F and E.

Note that the boundary conditions must be considered case by case for the typical cell configuration, structural symmetry, and connectivity with the adjacent cells in addition to the periodic arrangement of the cells in the global domain.

5 EXAMPLES

5.1 Three-dimensional NPR Materials

Figure 5 illustrates the material cell model for the three-dimensional NPR materials invented in the literature [1]. NPR materials are also called auxetic, anti-rubber, and dilational materials. Since Lakes published his article in Science in 1987 [18], NPR materials have attracted increasingly more attention because of their unique behavior. Unlike conventional materials, NPR materials may shrink when compressed (expand when stretched) along the perpendicular direction. The unique features of the NPR materials include, but are not limited to, 1) stiffening under a load, 2) impact energy absorption, 3) function-graded and function-oriented design, and 4) the ability to be engineered to have other desired functionalities.

As shown in Fig. 5, the invented NPR structure comprises a pyramid-shaped unit cell having four base points, A, B, C, and D, defining the corners of a square lying in a horizontal plane. Four stuffers of equal length or different lengths extend from a respective one of the base points to a point E spaced apart from the plane. Four tendons of equal length or different lengths, but less than the lengths of the stuffers, extend from one of the base points to a point F between point E and the plane. There are, in general, five parameters that determine the cell geometry, i.e., $\theta_1^{(1)}$, $\theta_1^{(2)}$, $\theta_2^{(1)}$, $\theta_2^{(2)}$, h_e. For simplicity, in this paper, we assume $\theta_1^{(1)} = \theta_1^{(2)} = \theta_1$ and $\theta_2^{(1)} = \theta_2^{(2)} = \theta_2$, such that the geometrical parameters are reduced to three, θ_1, θ_2, h_e. The design parameters of the NPR cell also include the material properties of the stuffers and the tendons, as well as shapes of the strips and the cross-section shapes of the stuffers and tendons. In this paper, we assume that all of the strips are straight and have constant cross-sectional shapes that can be represented by E_1, A_1, I_1, ρ_1 for the stuffers and E_2, A_2, I_2, ρ_2 for the tendons. Note, in these examples, we assume both the stuffers and the tendons are made of steel.

In the three-dimensional configurations, a collection of unit cells is arranged as tiles in the same horizontal plane with the base points of each cell connected to the base points of the adjoining cells, thereby forming a horizontal layer. A collection of horizontal layers is then stacked with each point E of the cells in one horizontal layer being connected to a point F of the cells in an adjacent layer. The above facts can be used to determine the connectivity among the adjacent cells and then used to determine the boundary conditions for the cell analyses in the homogenization process.

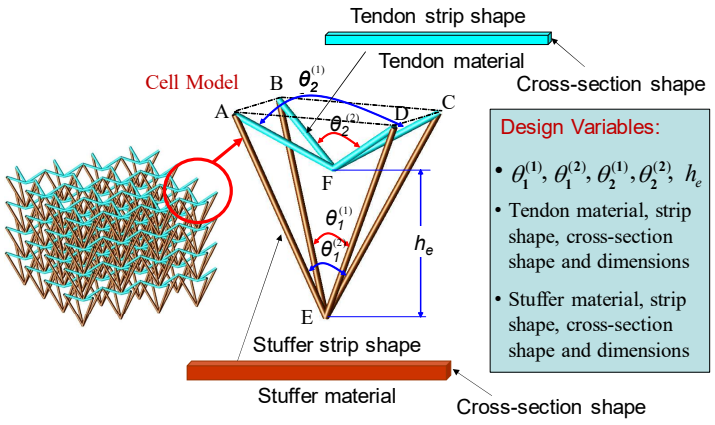


Figure 5: Design variables in a 3D NPR cell.

Figure 4 has illustrated three analytical cases with the loading and boundary conditions for the homogenization of a 2D NPR cell structure. In the following analyses, we consider the homogenization problem of a 3D NPR material shown in Fig. 5. Due to the symmetry of the cell structure, we can consider only a quarter of the cell structure, which results in a two-dimensional analysis problem, as shown in Fig. 6; meanwhile, Fig. 6 illustrates the problem setup for the first mode under a unit stress $\mathbf{e}^{(1)}=\{1,0,0,0,0,0\}^T$. Here, in Fig. 6, the concentrated force P_1 should be a quarter of the total force (unit stress multiplied by the top area of the cell) applied on the top of the cell. The boundary conditions are determined by considering the periodic conditions as well as the connectivity among the adjacent cells, as discussed in Sec. 4.2.

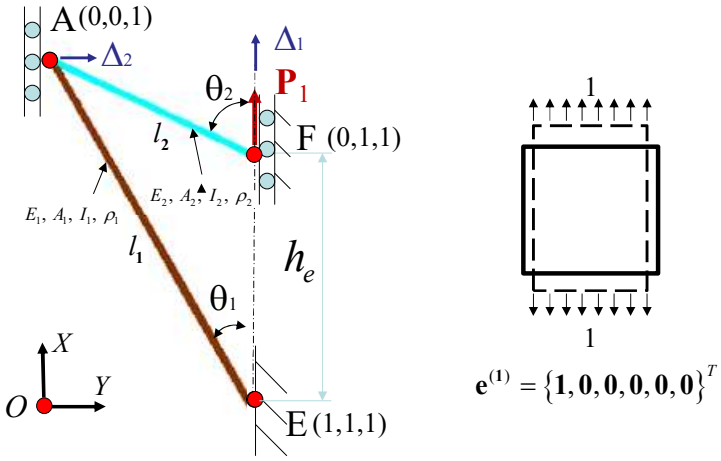


Figure 6: Analysis model for characteristic mode 1.

The analytical problem shown in Fig. 6 can be analytically solved using the assumption of the Euler–Bernoulli beam for the bending along with the tension of all of the members. As a result, we obtain the analytical solution for Δ_1 and Δ_2 as functions of P_1 as follows:

$$\begin{cases} \Delta_1=\dfrac{\left(\mathbf{g}_{11}^1+\mathbf{g}_{11}^2\right)\left(\mathbf{g}_{22}^1+\mathbf{g}_{22}^2\right)-\left(\mathbf{g}_{12}^1+\mathbf{g}_{12}^2\right)^2}{\mathbf{g}_{11}^1+\mathbf{g}_{11}^2}P_1 \\ \Delta_2=\dfrac{\mathbf{g}_{11}^2\mathbf{g}_{12}^1-\mathbf{g}_{12}^2\mathbf{g}_{11}^1}{\mathbf{g}_{11}^1+\mathbf{g}_{11}^2}P_1 \end{cases}, \tag{61}$$

where

$$\begin{cases} g_{11}^1=a_1\sin^2\theta_1+b_1\cos^2\theta_1 \\ g_{12}^1=\left(a_1-b_1\right)\sin\theta_1\cos\theta_1 \\ g_{22}^1=a_1\cos^2\theta_1+b_1\sin^2\theta_1 \end{cases}, \quad \begin{cases} g_{11}^2=a_2\sin^2\theta_2+b_2\cos^2\theta_2 \\ g_{12}^2=\left(a_2-b_2\right)\sin\theta_2\cos\theta_2 \\ g_{22}^2=a_2\cos^2\theta_2+b_2\sin^2\theta_2 \end{cases} \tag{62}$$

$$a_1=\frac{l_1}{E_1A_1},\ b_1=\frac{l_1^3}{12E_1I_1},\ a_2=\frac{l_2}{E_2A_2},\ b_2=\frac{l_2^3}{12E_2I_2}. \tag{63}$$

In the above formulation, there are only 3 independent design variables among θ_1, θ_2, l_1, and l_2, and the following relationship exists:

$$l_1 \sin\theta_1 = l_2 \sin\theta_2 . \tag{64}$$

The mass density can be obtained as

$$\rho^H = 4\left(\rho_1 l_1 A_1 + \rho_2 l_2 A_2\right)/V_e , \tag{65}$$

where V_e is the total volume of the cell.

Based on the above analysis, the corresponding strains along X and Y directions are obtained. Here, $\varepsilon_X = \Delta_1 / h_e$ and $\varepsilon_Y = \Delta_2/(l_1 \sin\theta_1)$. Figure 7 illustrates the effective Young's modulus (GPa) as a function of two design variables θ_1 and θ_2. Here, the horizontal axis is θ_1, and the vertical axis is the difference angle between θ_2 and θ_1. For this typical setup, the material becomes stiffer when θ_1 becomes small and the difference between θ_2 and θ_1 becomes small; the effective Young's modulus (in black) can vary from 1 GPa to 50 GPa in the parameter range considered. Figure 7 also shows the density contour (in red) as a percentage with respect to the design variables. It is seen that the density also increases when θ_1 becomes small and the angle difference becomes small; meanwhile, from the map, the optimal designs can be obtained for the given densities. For example, a 50 GPa design can be obtained at a material density of 25% when θ_1 is approximately 8 degrees and θ_2 is 53 degrees. Figure 8 illustrates the effective Poisson's ratio (as a percentage) obtained from the analysis. It is seen that the Poisson's ratio reaches the most negative number (less than -0.6) when θ_1 is within 10-15 degrees, θ_2 is within 20-30 degrees, and the material density is near 25%. Figure 9 further illustrates the critical pressure of the cellular material, indicating that a failure analyses can also be conducted using the model developed. Notably, the analysis results depend on the other parameters used for the cell element, and studying the actual material properties is not the purpose of this paper.

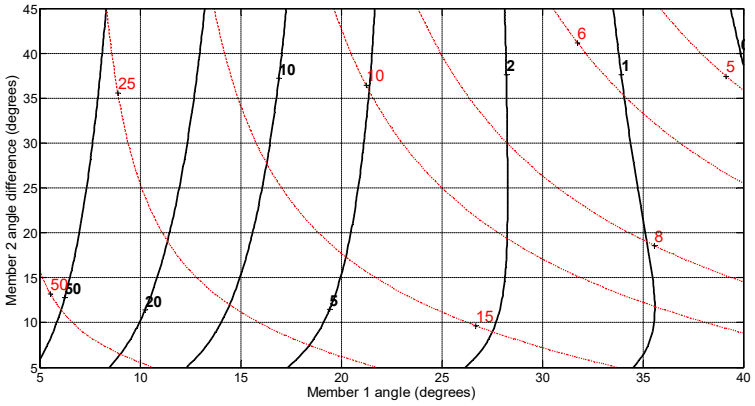


Figure 7: The effective Young's modulus (in black and GPa) for mode 1.

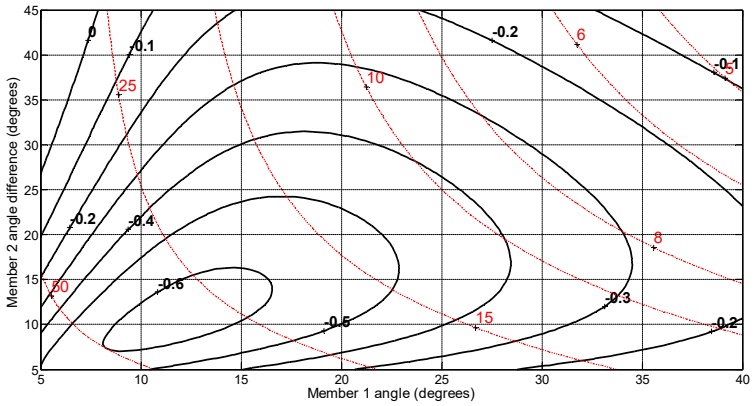


Figure 8: The effective Poisson's ratio (in black and 100%) for mode 1.

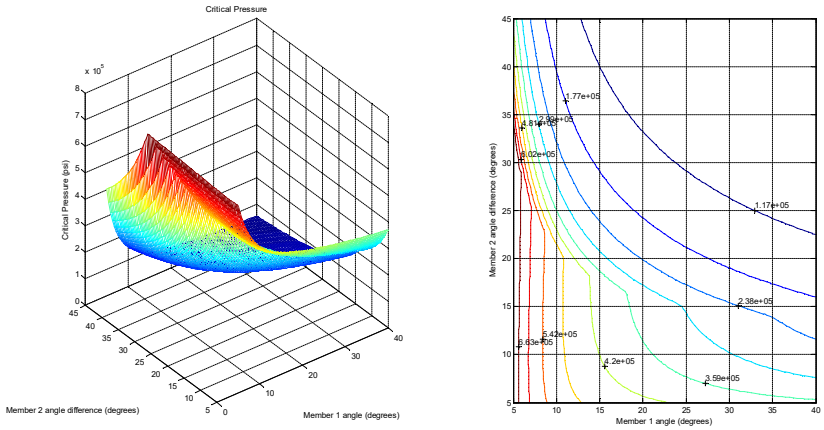


Figure 9: The critical pressure of the cellular structure for mode 1.

Figure 10 illustrates the problem setup for the second mode under a unit stress $\mathbf{e}^{(2)} = \{0,1,0,0,0,0\}^T$. Due to the symmetry of the cell structure, we can also consider only a quarter of the cell structure and a two-dimensional analysis problem, as show in Fig. 10. Here, the concentrated force P_2 should be the total force due to the unit stress applied on the side of the cell.

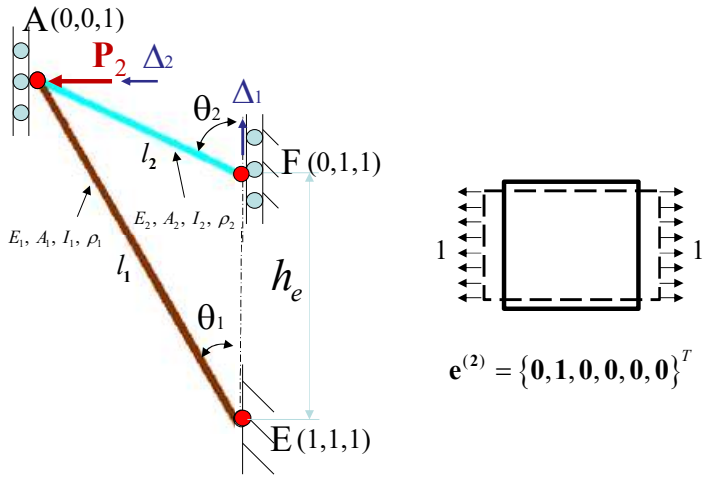


Figure 10: The analysis model for characteristic mode 2.

The analytical problem shown in Fig. 10 can be analytically solved with the same assumption as in the first case. As a result, for mode 2, we have

$$\begin{cases} \Delta_1 = \dfrac{\mathbf{g}_{11}^2\mathbf{g}_{12}^1 - \mathbf{g}_{12}^2\mathbf{g}_{11}^1}{\mathbf{g}_{11}^1 + \mathbf{g}_{11}^2} P_2 \\ \Delta_2 = \dfrac{\mathbf{g}_{11}^1\mathbf{g}_{12}^2}{\mathbf{g}_{11}^1 + \mathbf{g}_{11}^2} P_2 \end{cases}. \tag{66}$$

Figure 11 illustrates the effective Young's modulus (GPa) (in black) as a function of the two design variables θ_1 and θ_2. For this typical setup, the material becomes stiffer when θ_1 becomes larger, and the difference between θ_2 and θ_1 becomes larger; the effective Young's modulus can vary from 0.5 GPa to 20 GPa in the parameter range considered. Figure 12 illustrates the effective Poisson's ratio (100%) obtained from the analysis. The Poisson's ratio reaches the most negative number (less than -2.0) when θ_1 approaches 40 degrees and θ_2 is within 70-80 degrees. Figure 13 further illustrates the critical pressure on the side of the cellular structure indicating that side failure analyses can also be conducted using the model developed.

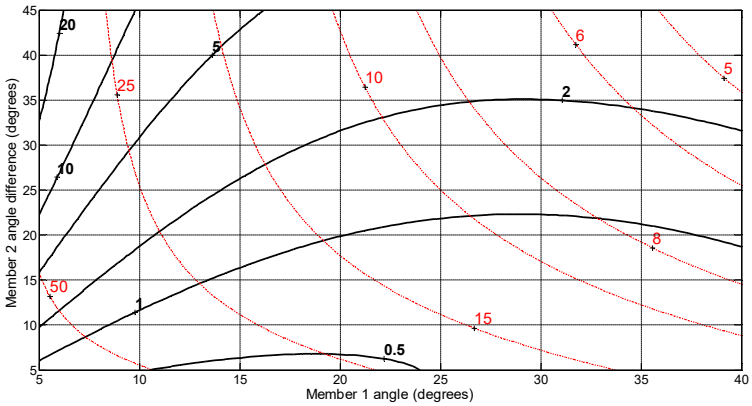


Figure 11: The effective Young's modulus (in black and GPa) for mode 2.

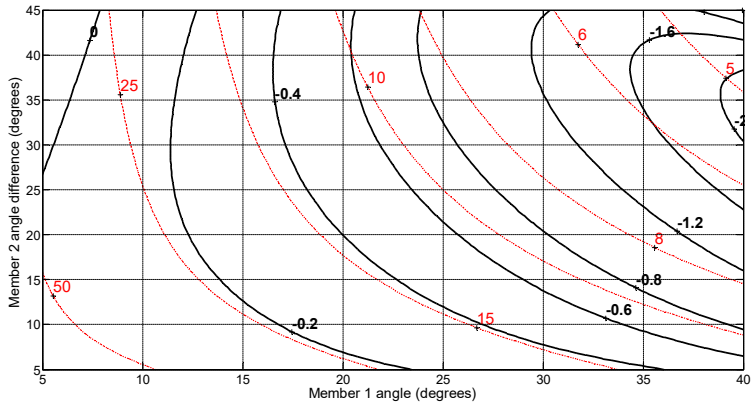


Figure 12: The effective Poisson's ratio (in black and 100%) for mode 2.

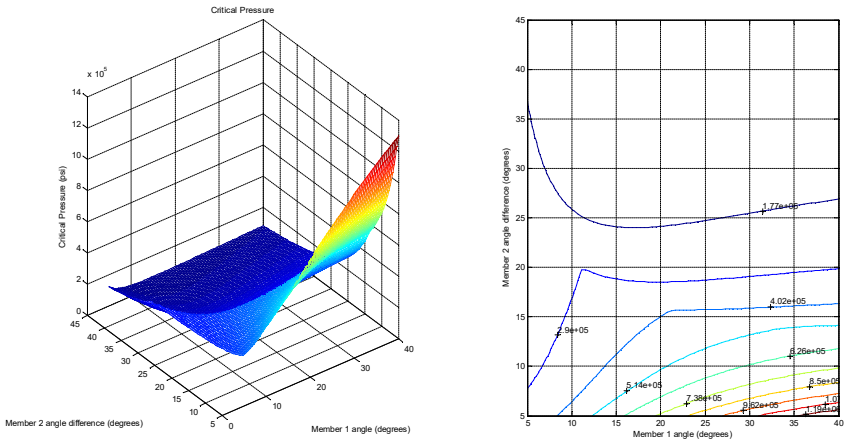


Figure 13: The critical pressure of the cellular structure for mode 2.

Note that the shear mode shown on the right in Fig. 4 and other shear modes in the three-dimensional problem can also be considered; in fact, two of the shear modes can be obtained using the

analysis model shown in Fig. 4, which is omitted in this paper.

5.2 BTR Materials

Figure 14 illustrates the parameterization of the BTR materials proposed in the literature [2]. As shown in Fig. 14, BTR has three major components: 1) cover sheets on the top and bottom, which are usually made of thin composites (such as fiber-reinforced polymer) or metallic sheets; 2) stuffers, which are usually made of stiff materials such as metallic columns, ceramics, or high stiffness composites; and 3) tendons, which are usually made of high strength tension materials, such as metal wires or high strength fibers. In summary, BTR materials can be made of various raw materials depending on the applications. One of major features of the BTR materials is their lightweight characteristics and high material efficiencies, especially, for the out-of-plane bending stiffness as compared to the existing composite materials. As shown in Fig. 14, the major design parameters for the geometry of BTR materials are h_2, l_1, and t_1. Other design parameters include the material properties of sheets, stuffers, and tendons represented by E_i and ρ_i $(i=1,2,3)$, the cross-sectional area of the stuffers represented by A_2, and the cross-sectional area of the tendons, A_3.

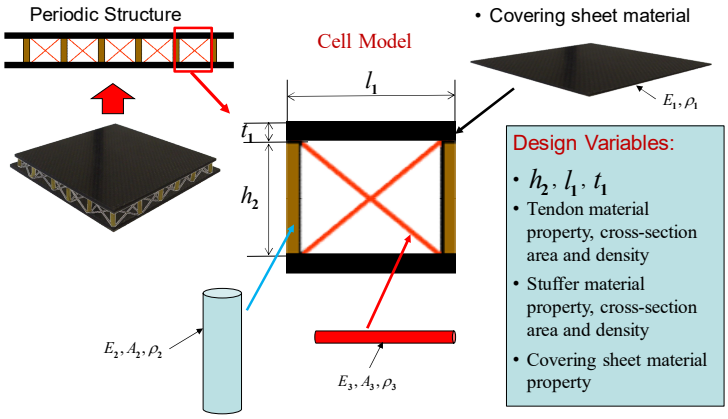


Figure 14: The design variables in a BTR cell.

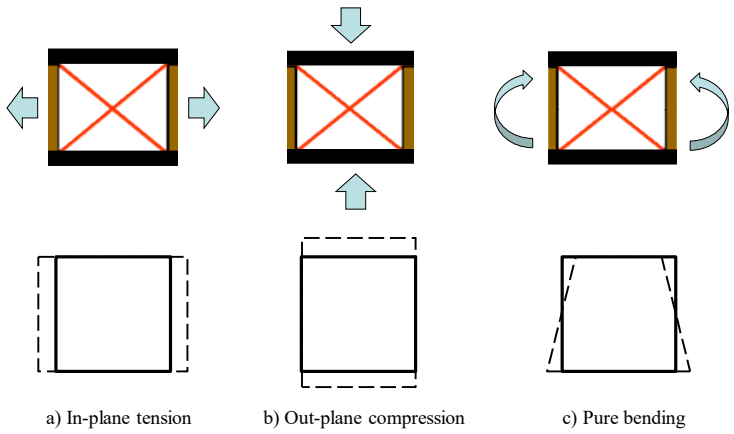


Figure 15： The strain modes of the BTR cell.

To estimate the effective material properties of BTR materials, a strain-based homogenization process is used. Fig. 15 illustrates the three major strain modes considered in this paper. Note that the two-dimensional figures shown in Fig. 15 actually represent three-dimensional structures. Additionally, note that because a major advantage of BTR materials is the high out-of-plane bending stiffness and out-of-plane bending strength, the effective bending stiffness is considered (Fig. 15c) instead of shear modulus.

From Fig. 15a, the homogenized in-plane tension modulus $E^H_{in-plane}$ can be obtained as

$$E^H_{in-plane} = \frac{2t_1}{h_2 + 2t_1} E_1 + \frac{2l_1^2 A_3}{\left(h_2 + 2t_1\right)\left(l_1^2 + h_2^2\right)^{\frac{3}{2}}} E_3 . \tag{67}$$

From Fig. 15b, the homogenized out-of-plane compression modulus $E^H_{out-plane}$ can be obtained as

$$E^H_{out-plane} = \frac{1}{l_1^2}\left(\frac{\left(h_2 + 2t_1\right) E_1 E_2}{h_2 E_1 + 2t_1 E_2} A_2 + \frac{4\left(h_2 + 2t_1\right) h_2^3 A_3}{l_1\left(l_1^2 + h_2^2\right)^{\frac{3}{2}}} E_3 \right). \tag{68}$$

From Fig. 15c, the homogenized (unit) bending stiffness $\left(EI\right)^H_{bending}$ can be obtained as

$$\left(EI\right)^H_{bending} = \frac{1}{12} E_1 \left[\left(h_2 + 2t_1\right)^3 - \left(h_2\right)^3 \right]. \tag{69}$$

And the effective mass density ρ^H can be obtained as

$$\rho^H = \frac{2t_1}{h_2 + 2t_1}\rho_1 + \frac{A_2}{l_1^2}\rho_2 + \frac{4\sqrt{l_1^2 + h_2^2}\, A_3}{l_1^2 h_2}\rho_3 . \tag{70}$$

Figures 16-18 illustrate example results obtained from the above formulations for a typical design case. Here, only two design variables, the stuffer height h_2 and the sheet thickness t_1, are varied for the parametric studies. Note that all of the results in Figs. 16-18 are normalized with the initial values $t_1 = 1\,\text{mm}$ and $h_2 = 15\,\text{mm}$.

Figure 16 shows the normalized bending stiffness with the normalized area density. The bending stiffness increases by increasing either design parameter, but an increase in the stuffer height is more effective than increasing the sheet thickness. This map is useful for selecting a proper design. For example, by selecting $t_1 = 1.5\,\text{mm}$ and $h_2 = 21.6\,\text{mm}$, we can increase the bending stiffness 3 times while only increasing the area density by 1.4 times.

Figure 17 shows the normalized in-plane modulus with the normalized area density. The in-plane modulus is insensitive to the design changes in this case. Figure 18 further shows the normalized out-plane modulus with the normalized area density. In this case, the out-of-plane modulus decreases when the stuffer height is increased, and it is less sensitive to design changes in the sheet thickness. Note that the analysis results also depend on the other parameters used for the cell element and studying the actual material properties is not the purpose of this paper.

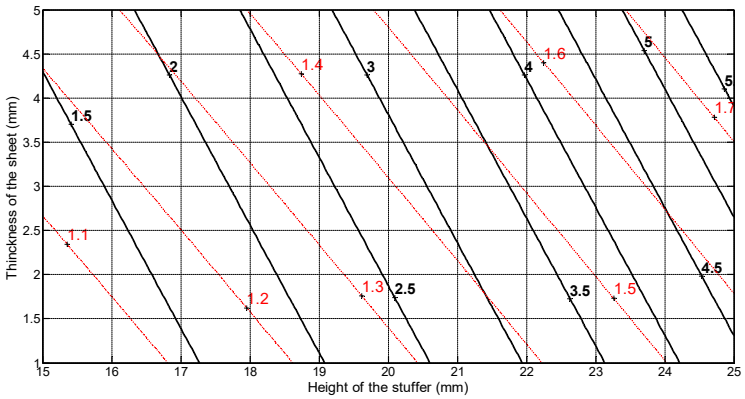


Figure 16：The normalized bending stiffness – normalized area density map.

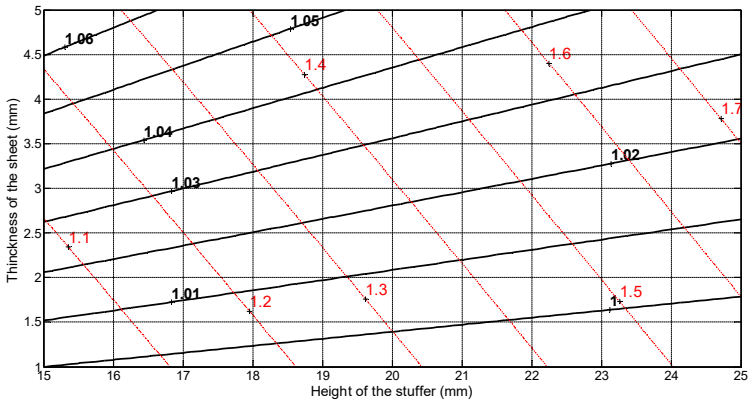


Figure 17：The normalized in-plane modulus – normalized area density map.

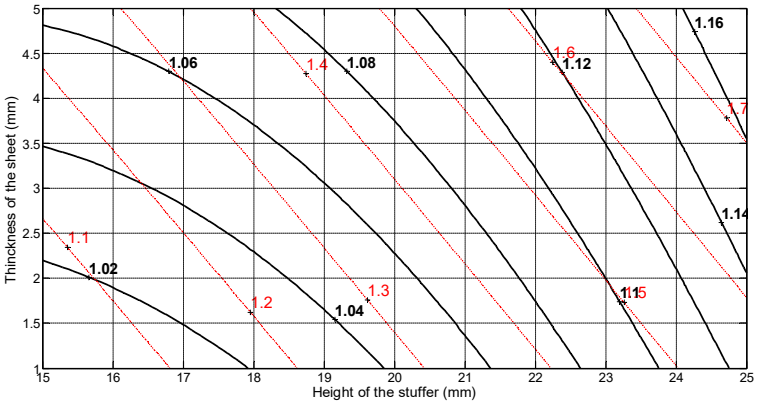


Figure 18：The normalized out-plane modulus – normalized area density map.

6 CONCLUSIONS

A mechanics-base homogenization approach is presented, resulting in two versions of the method: strain-based homogenization and stress-based homogenization. The strain-based homogenization process is obtained based on the *Principle of Virtual Displacements*; meanwhile, the stress-based homogenization process is obtained based on the *Principle of Virtual Forces*. Strain-based homogenization separates the strain field into a homogenized strain field and a strain variation field superposed on the homogenized strain field. The stress-based separates the stress field into a homogenized stress field and a stress variation field. The *Principle of Virtual Displacements* (*Principle of Virtual Forces*) for the relationship between the strain (stress) variation field and the homogenized strain (stress) field is then used to condense the strain (stress) variation field to the homogenized strain (stress) field, and the homogenization processes become coordinate reduction similar to the Guyan Reduction.

The new derivation in this paper provides improved engineering insight and enhanced physical understanding for dealing with boundary conditions, internal forces, and other issues in homogenization modeling processes. The stress-based homogenization process can be extended easily to handle discrete systems, providing a bridge between discrete and continuum systems. The mode analysis and the recovery process presented in this paper are useful tools for failure mode prediction, failure mode management, and design optimization of architectured materials.

The new method provides a way to further improve the accuracy and efficiency of multi-scale analysis problems, especially in nonlinear and structural dynamics analyses, and it more easily extended to handle various mechanical simulation and design problems. The newly derived method has been utilized to obtain effective material properties of the architectured cellular materials discussed in this paper.

REFERENCES

[1] Z.-D. Ma, Three-dimensional auxetic structures and applications thereof, *US Patent* 7,910,193 B2, 2011.

[2] Z.-D. Ma, Lightweight, rigid composite structures, *US Patent* 7,563,497 B2, 2009.

[3] E. Weinan, B. Engquist, X. Li, W. Ren, E. Vanden-Eijnden, The ~~heterogeneous~~ multiscale methods: a review. *Communications in Computational Physics*, ISSN 1815-2406, 2 (3), 367-450.

[4] P. Kanouté, D. P. Boso , J. L. Chaboche, B. A. Schrefler, Multiscale methods for composites: a review, *Archives of Computational Methods in Engineering*, 16 (1), 31-75, 2009.

[5] V. P. Nguyen, M. Stroeven, L. J. Sluys, Multiscale continuous and discontinuous modeling of heterogeneous materials: a review on recent developments, *Journal of Multiscale Modelling*, 3 (4), 1–42, 2011.

[6] E. Sanchez-Palencia Non-homogenous media and vibration theory. *Lect Notes Phys*, 127, 1980.

[7] A. Benssousan, JL, Lions, G. Papanicoulau, Asymptotic analysis for periodic structures. *Amesterdam*, North Holland, 1978.

[8] D. Cioranescu, JSJ. Paulin, Homogenization in open sets with holes. *J Math Aanal Appl,* 71, 590-607, 1979.

[9] MP Bendsùe, N. Kikuchi, Generating optimal topologies in structural design using homogenization method. *Comp Meth Appl Mech Engng*, 71, 197-224, 1988.

[10] Z.-D. Ma, N. Kikuchi, H.-C. Cheng, Topological design for vibrating structures. *Comp Meth Appl Mech Engng*, 121, 259-280, 1995.

[11] K. Matsui, K. Terada, “Continuous approximation of material distribution for topology optimi-

zation. *Int. J. Numer. Meth. Engng*, 59, 1925-1944, 2004.

[12] F. Dos Reis, J.F. Ganghoffer, Discrete homogenization of architectured materials: implementation of the method in a simulation tool for the systematic prediction of their effective elastic properties, *TECHNISCHE MECHANIK*, 30, 1 (3), 85–109, 2010.

[13] S. Arabnejad, D. Pasini, Mechanical properties of lattice materials via asymptotic homogenization and comparison with alternative homogenization methods. *International Journal of Mechanical Sciences*, 77, 249–262, 2013.

[14] K. Terada, N. Kikuchi, A class of general algorithms for multi-scale analyses of heterogeneous media, Comput. Methods Appl. Mech. Engrg. 190, 40-41, 2001.

[15] A. Abdulle, Y. Bai, Reduced-order modelling numerical homogenization. *ANMC, Mathematics Section*, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, 2014.

[16] Y. Cong, S. Nezamabadi, H. Zahrouni, J. Yvonnet, Multiscale computational homogenization of heterogeneous shells at small strains with extensions to finite displacements and buckling. *Int. J. Numer. Meth. Engng*, 104 (4), 235-259, 2015.

[17] R.J. Guyan, Reduction of stiffness and mass matrices. *AIAA Journal*, 3 (2), 380, 1964.

[18] R. S. Lakes, Advances in negative Poisson's ratio materials. *Advanced Materials*, 5, 293-296, 1993.

APPENDIX A: GUYAN REDUCTION

Guyan Reduction [17] is well known and widely used in the analyses of finite element-based structure dynamics problems to reduce the size of the problem to be solved. It is assumed that the finite element equation of a substructure can be written as:

$$\mathbf{m}\ddot{\mathbf{u}} + \mathbf{k}\mathbf{u} = \mathbf{f} \tag{A-1}$$

where $\mathbf{u}$ stands for the nodal coordinate vector of the finite element model; $\mathbf{m}$, $\mathbf{c}$, $\mathbf{k}$, and $\mathbf{f}$ stand for the mass matrix, damping matrix, stiffness matrix, and nodal force vector of the structure, respectively.

The Guyan Reduction process divides the nodal coordinates $\mathbf{u}$ as

$$\mathbf{u} = \begin{Bmatrix} \mathbf{u}_I \\ \mathbf{u}_J \end{Bmatrix}, \tag{A-2}$$

where $\mathbf{u}_J$ contains the remaining coordinates (in which force is applied), and $\mathbf{u}_I$ contains the coordinates to be condensed (in which no force is applied). One can rewrite the corresponding static problem of Eq. (A-1) as

$$\begin{cases} \mathbf{k}_{II}\mathbf{u}_I + \mathbf{k}_{IJ}\mathbf{u}_J = \mathbf{0} \\ \mathbf{k}_{JI}\mathbf{u}_I + \mathbf{k}_{JJ}\mathbf{u}_J = \mathbf{f}_J \end{cases}. \tag{A-3}$$

Because there is no force applied on $\mathbf{u}_I$, $\mathbf{u}_I$ can be consider a function depending on $\mathbf{u}_J$; namely, we can find a matrix $\mathbf{X}$ that satisfies

$$\mathbf{u}_I = \mathbf{X}\mathbf{u}_J . \tag{A-4}$$

Substituting Eq. (A-4) into the first row equation in Eq. (A-3) gives

$$\mathbf{X} = -\mathbf{k}_{II}^{-1}\mathbf{k}_{IJ} . \tag{A-5}$$

Substituting Eqs. (A-4) and (A-5) into the second row equation in Eq. (A-3) gives

$$\mathbf{k}^*\mathbf{u}_J = \mathbf{f}_J , \tag{A-6}$$

where, $\mathbf{k}^*$ in Eq. (A-7) is called the reduced stiffness matrix (or condensed stiffness matrix, effective stiffness matrix),

$$\mathbf{k}^* = \mathbf{k}_{JJ} - \mathbf{k}_{JI}\mathbf{k}_{II}^{-1}\mathbf{k}_{IJ} . \tag{A-7}$$

Equations (A-4) and (A-5) can also be used to condense the mass matrix in Eq. (A-1), and we have

$$\mathbf{m}^* = \mathbf{m}_{JJ} + \mathbf{X}^T\mathbf{m}_{IJ} + \mathbf{m}_{JI}\mathbf{X} + \mathbf{X}^T\mathbf{m}_{II}\mathbf{X} , \tag{A-8}$$

where $\mathbf{m}^*$ is called the reduced mass matrix.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A HOMOGENIZATION TECHNIQUE FOR ELASTO-PLASTIC COMPOSITES

F. Covezzi[1], S. de Miranda[1], S. Marfia[2], and E. Sacco[2]

[1]DICAM, University of Bologna
Viale del Risorgimento, Bologna, Italy
e-mail: {federica.covezzi3,stefano.demiranda}@unibo.it

[2] DICeM, University of Cassino and Southern Lazio
Via G. di Biasio 43, Cassino, Italy
e-mail: {marfia,sacco}@unicas.it

Keywords: Composite materials, homogenization, nonuniform TFA, complementary formulation, hardening.

Abstract. *The present study proposes a homogenization technique for the estimation of the overall behavior of composite materials characterized by a nonlinear behavior. A mixed nonuniform TFA procedure is proposed in order to study the mechanical response of periodic composites characterized by linear isotropic hardening plasticity. As the case of periodic composites is considered, the homogenization is performed on a repetitive unit cell that plays the equivalent role of representative volume element for random media. A numerical application is performed to test the effectiveness of the proposed technique. In particular, the homogenization results are compared with the ones carried out by micromechanical nonlinear finite element analyses.*

1 INTRODUCTION

The popularity and the use of composite materials have recently increased in the realm of structural engineering, due to their high performances deriving from their superior properties in terms of light weight and resistance. For this reason, effective tools able to predict the mechanical behavior of these media at different scales are required in the design of composite materials. A significant contribution is given by homogenization techniques, which allow to determine the overall behavior of a composite medium starting from its microscopic properties and finally to derive an equivalent homogenized material for numerical computations.

Among the homogenization techniques, an interesting approach for the investigation of the nonlinear response of composites is the Transformation Field Analysis (TFA), originally proposed in [1]. This technique takes into account the nonlinear behavior of the constituents, introducing an approximation of the inelastic strain. In the literature, several TFA schemes have been developed, differing mainly in the assumption concerning the distribution of the inelastic strain and in the computation of the internal variable evolution [2, 3, 4, 5, 6].

Recently, a new mixed nonuniform TFA (MxTFA) scheme has been presented by the authors [7], with the aim to improve the description of the inelastic strain field and the computation of the evolutive problem. In this paper, an extension of the MxTFA to the analysis of periodic composites characterized by linear isotropic hardening plasticity is presented. In particular, a mixed variational approach, originally proposed in [8] and recently implemented in [9, 10], involving the weak form of compatibility and plastic admissibility equations is adopted to derive the evolution laws of the internal variables in the TFA framework. In fact, the literature shows how stress recovery techniques are able to produce very good results [11, 12].

According to the MxTFA, the RVE is subdivided into subsets and in each subset a uniform plastic multiplier and equilibrated, linearly varying stresses are assumed. As a result, the inelastic strain distribution depends on the assumed distributions of stresses and plastic multiplier, that are the independent variables of the problem, and whose parameters are simultaneously evaluated by enforcing the compatibility condition and the elasto-plastic consistency condition.

The effectiveness of the proposed procedure is shown through a numerical example in the framework of 2D plane stress hardening plasticity, where the results are compared with the ones carried out by micromechanical nonlinear finite element analyses.

In the following, the Voigt notation is adopted; the strain and stress tensors are reported as the vectors with 3 components $\boldsymbol{\varepsilon} = \{\varepsilon_{11}\ \varepsilon_{22}\ \gamma_{12}\}^T$ and $\boldsymbol{\sigma} = \{\sigma_{11}\ \sigma_{22}\ \sigma_{12}\}^T$, respectively.

2 MxFTA HOMOGENIZATION WITH HARDENING PLASTICITY

Let a composite material be considered assuming a periodic microstructure, so that the micromechanical analysis and the homogenization of the material can be performed on a unit cell (UC) .

The micromechanical and homogenization problem for the UC denoted by Ω consists in computing the periodic displacement field $\widetilde{\mathbf{u}}$, the strain field $\boldsymbol{\varepsilon}$, the periodic strain field $\widetilde{\boldsymbol{\varepsilon}}$, the inelastic strain field $\boldsymbol{\pi} = \{\pi_{11}\ \pi_{22}\ \pi_{12}\}^T$, the stress field $\boldsymbol{\sigma}$ and its average value $\bar{\boldsymbol{\sigma}}$, for a prescribed value of the average strain $\bar{\boldsymbol{\varepsilon}}$ assigned in Ω. To this end the compatibility, the constitutive, the equilibrium with zero body forces equations, with periodic suitable boundary conditions as well as the strain and the stress average equations have to be satisfied.

The total strain at the point $\mathbf{x}$ of the UC depends on the contributions deriving from the average strain $\bar{\boldsymbol{\varepsilon}}$ and the inelastic strain $\boldsymbol{\pi}$. The periodic strain field $\widetilde{\boldsymbol{\varepsilon}}$ can be therefore split into:

$$\widetilde{\boldsymbol{\varepsilon}} = \widetilde{\boldsymbol{\varepsilon}}^{\varepsilon} + \widetilde{\boldsymbol{\varepsilon}}^{\pi} \qquad \text{with} \quad \widetilde{\boldsymbol{\varepsilon}}^{\varepsilon} = \mathbf{L}^{\varepsilon}\bar{\boldsymbol{\varepsilon}}\,,\ \widetilde{\boldsymbol{\varepsilon}}^{\pi} = \mathbf{L}^{\pi}\boldsymbol{\pi} \tag{1}$$

with $\widetilde{\boldsymbol{\varepsilon}}^{\varepsilon}$ and $\widetilde{\boldsymbol{\varepsilon}}^{\pi}$ the perturbation strains due to the average strain $\bar{\boldsymbol{\varepsilon}}$ and to the presence of the inelastic field $\boldsymbol{\pi}$, respectively; $\mathbf{L}^{\varepsilon}$ and $\mathbf{L}^{\pi}$ are localization matrices depending on the point $\mathbf{x}$.

Assuming an associative evolution law, the evolution of the inelastic strain is governed by:

$$\dot{\boldsymbol{\pi}} = \dot{\gamma}\,\mathbf{N}\left(\boldsymbol{\sigma}\right) \qquad \text{with} \quad \mathbf{N}\left(\boldsymbol{\sigma}\right) = \frac{\partial f}{\partial \boldsymbol{\sigma}}\,, \tag{2}$$

where γ is the plastic multiplier and f is the yield function, satisfying the Kuhn-Tucker and consistency conditions:

$$f\left(\boldsymbol{\sigma}\right) \leq 0\,, \qquad \dot{\gamma} \geq 0\,, \qquad f\left(\boldsymbol{\sigma}\right)\dot{\gamma} = 0\,, \qquad \dot{f}\left(\boldsymbol{\sigma}\right)\dot{\gamma} = 0\,. \tag{3}$$

The compatibility condition and the Kuhn-Tucker conditions (3) can be rewritten in the weak form as:

$$\int_{\Omega} \delta\boldsymbol{\sigma}^{T}\mathbf{C}^{-1}\boldsymbol{\sigma}\,dV - \int_{\Omega} \delta\boldsymbol{\sigma}^{T}\left(\bar{\boldsymbol{\varepsilon}} + \mathbf{D}\,\widetilde{\mathbf{u}} - \boldsymbol{\pi}\right)\,dV = 0 \quad \forall\,\delta\boldsymbol{\sigma}:\ \mathbf{D}^{T}\delta\boldsymbol{\sigma} = \mathbf{0}, \tag{4}$$

$$\int_{\Omega} \delta\gamma\, f\left(\boldsymbol{\sigma}\right)\,dV \leq 0\,, \quad \dot{\gamma} \geq 0\,, \quad \int_{\Omega} \delta\gamma\, f\left(\boldsymbol{\sigma}\right)\dot{\gamma}\,dV = 0 \qquad \forall\,\delta\gamma \geq 0\,, \tag{5}$$

where $\mathbf{D}$ is the compatibility matrix operator, $\mathbf{D}^{T}$ the equilibrium operator, $\mathbf{C}$ the material constitutive matrix and V the volume of Ω.

Considering plane stress elasto-plasticity with a von Mises yield criterion in the framework of linear isotropic hardening, the limit function is:

$$f\left(\boldsymbol{\sigma}\right) = \frac{3}{2}\,\boldsymbol{\sigma}^{T}\mathbf{M}\,\boldsymbol{\sigma} - \left(\sigma_{y} + H\,\bar{\varepsilon}_{p}\right)^{2}\,, \tag{6}$$

where

$$\mathbf{M} = \frac{1}{3}\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 6 \end{bmatrix},$$

σ_{y} is the flow stress, H is the hardening parameter, and $\bar{\varepsilon}_{p}$ is the accumulated plastic strain, whose evolution is defined as:

$$\dot{\bar{\varepsilon}}_{p} = \sqrt{\frac{2}{3}}\|\dot{\boldsymbol{\pi}}\|. \tag{7}$$

Let the UC be divided into n subsets Ω^{j} with $j = 1, .., n$, each of them characterized by a volume V^{j}, such that:

$$\Omega = \bigcup_{j=1}^{n} \Omega^{j}\,, \qquad V = \sum_{j=1}^{n} V^{j}\,. \tag{8}$$

A representation form is assumed for both the stress and plastic multiplier fields. In particular, it is assumed that on the $j-$th subset Ω^{j} :

$$\boldsymbol{\sigma}^{j} = \mathbf{P}^{j}\,\widehat{\boldsymbol{\sigma}}^{j}\,, \tag{9}$$

$$\dot{\gamma}^j = \sum_{i=1}^{\rho} r_i\left(\mathbf{x}\right) \dot{\widehat{\gamma}}_i^j = \left(\mathbf{r}^j\right)^T \dot{\widehat{\boldsymbol{\gamma}}}^j \,, \tag{10}$$

where $\mathbf{P}^j$ is a $3 \times p$ matrix collecting a set of p self-equilibrated stress modes, $\widehat{\boldsymbol{\sigma}}^j$ the vector of the unknown p stress parameters, $\mathbf{r}^j$ and $\dot{\widehat{\boldsymbol{\gamma}}}^j$ are the two vectors collecting the ρ approximation functions and coefficients of a linear combination, respectively. In particular, it is assumed a linear approximation for the stress field in the subset Ω^j; consequently, it results $p = 7$ and

$$\mathbf{P}^j = \begin{bmatrix} 1 & 0 & 0 & x_2 & 0 & x_1 & 0 \\ 0 & 1 & 0 & 0 & x_1 & 0 & x_2 \\ 0 & 0 & 1 & 0 & 0 & -x_2 & -x_1 \end{bmatrix} ; \tag{11}$$

moreover, it is set $\rho = 1$, so that the plastic multiplier is approximated in the typical subset Ω^j by a uniform function, i.e. $\dot{\gamma}^j = \dot{\widehat{\gamma}}^j$.

Taking into account the equations (9) and (10), the flow rule equation (2) in the subset Ω^j takes the form:

$$\dot{\boldsymbol{\pi}}^j = \left(\mathbf{r}^j\right)^T \dot{\widehat{\boldsymbol{\gamma}}}^j \hat{\mathbf{N}}\left(\widehat{\boldsymbol{\sigma}}^j\right) \qquad \text{with} \quad \hat{\mathbf{N}}\left(\widehat{\boldsymbol{\sigma}}^j\right) = \mathbf{N}\left(\mathbf{P}^j\, \widehat{\boldsymbol{\sigma}}^j\right) \,. \tag{12}$$

From Equation (12) it can be highlighted that the representation form for the inelastic strain rate depends on the approximations introduced for the stress and plastic multiplier.

Substituting the introduced approximations in equations (6) and (12) for each subset Ω^j the discretized form of the yield function and of the evolution law are recovered. Finally, the evolution problem is solved adopting a step-by-step backward Euler algorithm. At each time step, the solution of the evolution problem is performed by establishing a predictor-corrector strategy. Details on the numerical model can be found in [7].

3 NUMERICAL TEST

In this section, the performance of the proposed homogenization procedure is verified on the repetitive UC representative of a realistic application, subjected to periodic boundary conditions. The microstructure of the UC is characterized by a plastic matrix embedding elastic circular inclusions with volume fraction equal to 36%. The UC geometry is shown in Fig. 1a; due to symmetry only one quarter of the UC is modeled. The following geometrical parameters are assumed: $a = 0.5$ mm and $r = 0.34$ mm. The material parameters are collected in Table 1.

The localization matrices $\mathbf{L}^{\varepsilon,j}$ and $\mathbf{L}^{\pi^i,j}$ are evaluated performing linear elastic finite element analyses of the UC subjected to each component, assumed equal to one, of $\bar{\boldsymbol{\varepsilon}}$ and of $\boldsymbol{\pi}^i$, respectively, determining the perturbation strain at each Gauss point of the finite element mesh. The MxTFA homogenization analysis is performed discretizing the UC with 7 subsets (6 subsets for the matrix and 1 subset for the inclusion), as shown Fig. 1b. Thus, the number of linear elastic pre-analyses is equal to the 3 components of $\bar{\boldsymbol{\varepsilon}}$ plus the number of the subsets times the 7 components of $\widehat{\boldsymbol{\sigma}}^j$ ($3 + 6 \times 7 = 45$); the number of internal variables results $6 \times 7 = 42$. For the finite element model employed to perform the linear elastic pre-analyses and to compute the micromechanical reference solution, the quarter of UC is discretized using 2D plane stress four-node finite elements. In particular, 304 finite elements are used, 176 for the matrix and 128 for the inclusion. The internal variables of the nonlinear micromechanical analysis are equal to $176 \times 4 \times 3 = 2112$, significantly higher than the ones of the MxTFA.

The results obtained with the homogenization procedure are compared with nonlinear micromechanical analyses (FEA), carried out using finite elements characterized by elastic and plastic constitutive models.

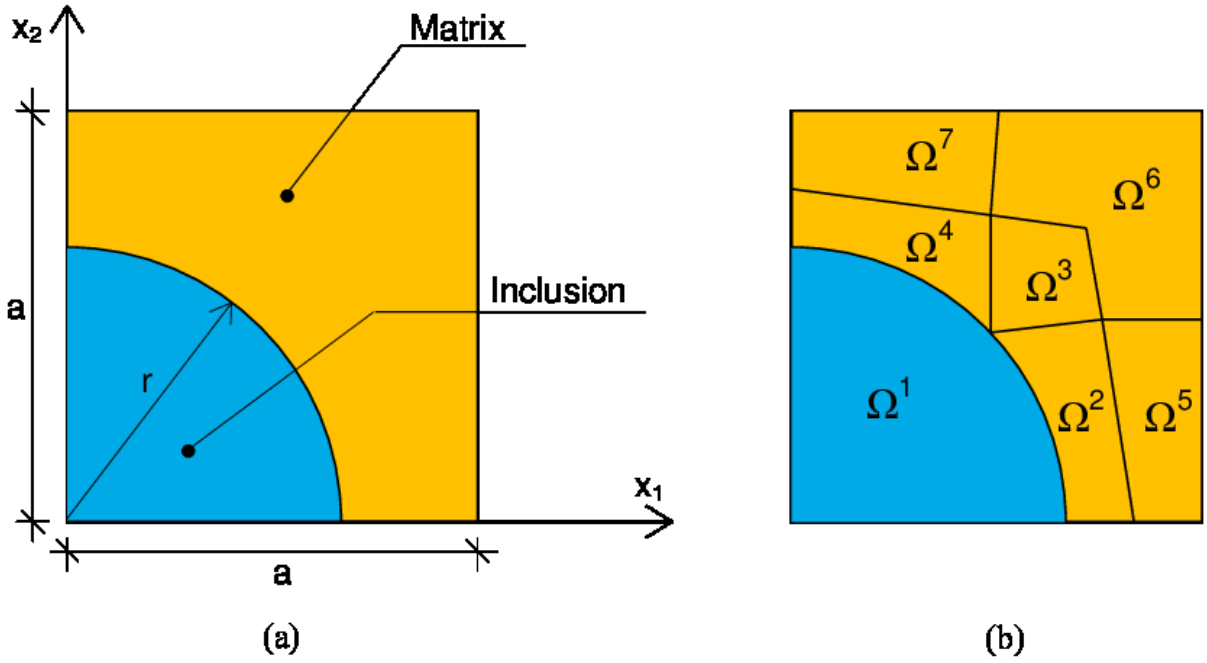


Figure 1: Geometry of the UC (a) and subset discretization (b).

Constituent	E [MPa]	ν	σ_y [MPa]	H [MPa]
Inclusion (elastic material)	410000	0.19	-	-
Matrix (plastic material)	75000	0.33	426	2894

Table 1: Material properties of the constituents.

In this test the UC is subjected to uniaxial loading along the x_1-axis, with the average strain $\bar{\varepsilon}_{11}$ monotonically increased until the final value $\bar{\varepsilon}_{11} = 0.03$. Fig. 2 shows the constitutive response in both cases of elasto perfectly plastic and linear hardening (LH) of the equivalent homogenized material in terms of $\bar{\sigma}_{11}$ and $\bar{\varepsilon}_{11}$, together with the micromechanical reference solution. As it can be noted, the MxTFA is able to accurately reproduce the overall behavior of the UC. Moreover, the final value of the average stress $\bar{\sigma}_{11}$ obtained with the MxTFA is compared with the reference one deriving from the micromechanical analysis in Table 2; as it can be noted, the error is less than 12%.

4 CONCLUSIONS

A new mixed nonuniform TFA for the analysis of periodic composites characterized by linear isotropic hardening is proposed. The inelastic field is recovered from the equilibrated stresses and plastic multiplier parameters, that are the independent variables of the problem. The evolution equations involving the internal variables are rationally derived from a variational principle based on the complementary formulation.

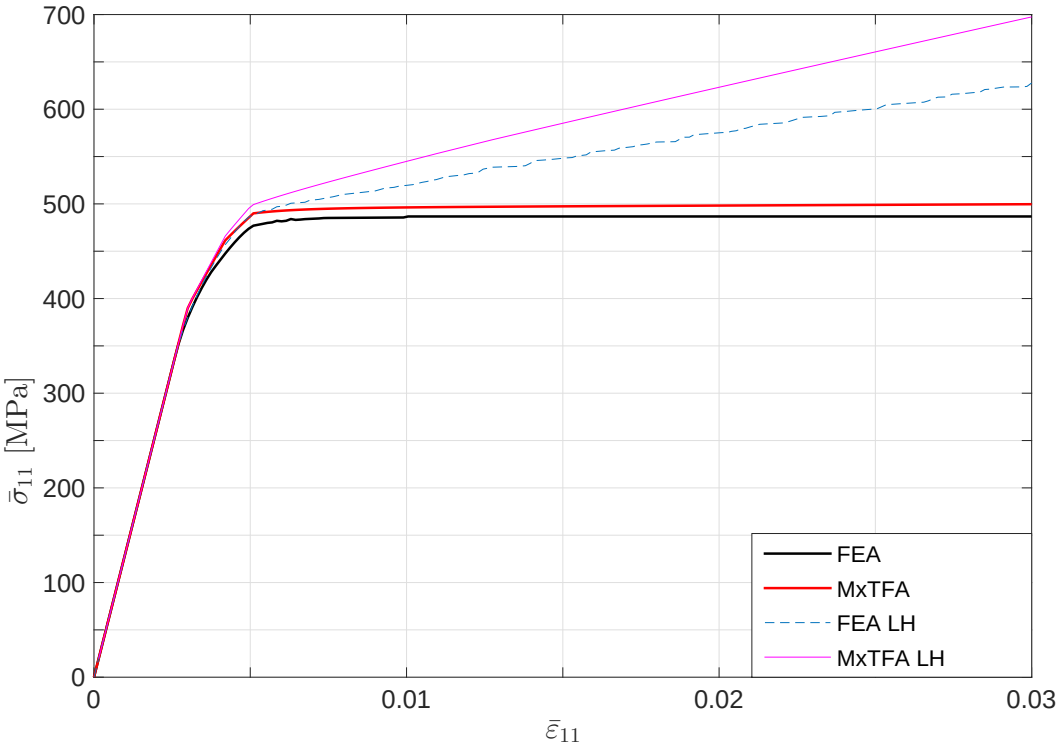


Figure 2: Mechanical response of the UC subjected to uniaxial loading.

Type of analysis	$\bar{\sigma}_{11}$(MPa)	Error (%)
FEA	486.71	-
MxTFA	499.66	2.66
FEA LH	627.73	-
MxTFA LH	697.53	11.11

Table 2: Uniaxial loading: final value of the average stress.

The numerical example shows the ability of the proposed MxTFA technique in reproducing the micromechanical nonlinear response for periodic composite characterized by Mises linear isotropic hardening behavior.

The presented approach is much less expensive than the micromechanical finite element analyses from a computational point of view. In fact, in the proposed approach the elasto-plastic problem is solved at each time-step iteration at subset level, adopting a variational formulation resulting in a system of equation involving a number of internal variables significantly lower than in the nonlinear micromechanical analysis.

Because of the computation efficiency, the proposed MxTFA approach results very promising for being implemented at the Gauss point level for the development of an effective multiscale procedure.

Acknowledgments The financial supports of PRIN 2010-11 project “Advanced mechanical modeling of new materials and technologies for the solution of 2020 European challenges”

CUP n. F11J12000210001 and of the University of Cassino and of the Southern Lazio are gratefully acknowledged.

REFERENCES

[1] G. Dvorak, Transformation field analysis of inelastic composite materials, Proceedings of the Royal Society of London A 437 (1992) 311–327.

[2] J. Michel, P. Suquet, Nonuniform transformation field analysis, International Journal of Solids and Structures 40 (2003) 6937–6955.

[3] S. Marfia, E. Sacco, Micromechanics and homogenization of sma-wire-reinforced materials, Journal of Applied Mechanics 72 (2005) 259–268.

[4] F. Fritzen, T. Böhlke, Three-dimensional finite element implementation of the nonuniform transformation field analysis, International Journal for Numerical Methods in Engineering 278 (2010) 186–217.

[5] S. Marfia, E. Sacco, Multiscale damage contact-friction model for periodic masonry walls, Computer Methods in Applied Mechanics and Engineering 205-208 (2012) 189–203.

[6] V. Sepe, S. Marfia, E. Sacco, A nonuniform TFA homogenization technique based on piecewise interpolation functions of the inelastic field, International Journal of Solids and Structures 50 (2013) 725–742.

[7] F. Covezzi, S. de Miranda, S. Marfia, E. Sacco, Complementary formulation of the tfa for the elasto-plastic analysis of composites, Composite Structures 141 (2016) In press.

[8] J. Simo, J. Kennedy, R. L. Taylor, Complementary mixed finite element formulations for elastoplasticity, Computer Methods in Applied Mechanics and Engineering 74 (1989) 177–206.

[9] A. Bilotta, L. Leonetti, G. Garcea, Three field finite elements for the elastoplastic analysis of 2d continua, Finite Elements in Analysis and Design 47 (2011) 1119–1130.

[10] F. Daghia, S. de Miranda, F. Ubertini, Patch based recovery in finite element elastoplastic analysis, Computational Mechanics 52 (2013) 827–836.

[11] G. Castellazzi, S. de Miranda, F. Ubertini, Adaptivity based on the recovery by compatibility in patches, Finite Elements in Analysis and Design 46 (2010) 379–390.

[12] G. Castellazzi, S. de Miranda, F. Ubertini, Patch based stress recovery for plate structures, Computational Mechanics 47 (2011) 379–394.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THEORETICAL PREDICTION OF PLASTIC INSTABILITY FOR ASYMMETRICAL MATERIALS

Shenghua Wu[1], **Nannan Song**[2], **and F.M.Andrade Pires**[2]

[1] Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias 4200-465 Porto, Portugal
e-mail: {shenghua,fpires}@fe.up.pt

[2] INEGI, University of Porto
Rua Dr. Roberto Frias 4200-465 Porto, Portugal
snannan@inegi.up.pt

Keywords: Forming Limit Diagram (FLD), Plastic instability, Necking, Yield function.

Abstract. *The prediction of formability is one of the most important tasks in sheet metal process simulation. The forming limit diagram (FLD) is an important tool that is used to understand and characterize the formability of sheet metals. This method meets both manufacturer and user's requirements and is widely used in factory and research laboratories* [1, 2]. *However, experimental measurements and determination of the FLD is costly and time consuming. Therefore analytical and numerical predictions based on the theory of plastic instability allow determining the critical value under different loading paths and has already attracted significant attention for formability evaluation.*

In this contribution, a mathematical formulation is derived by using Swift and Hill analytical theory. The formulation is numerically implemented in a user developed FLD MATLAB script to predict diffuse and localized necking. This work also presents and discusses the influence of different yield surface and asymmetrical effect on the formability under different loading strain path.

1 INTRODUCTION

In sheet metal forming, necking and wrinkling are commonly observed failure. In order to describe the occurrence of necking and present the formability, a most realistic and general method was introduced at the beginning of 1960s by Keeler [1] and Goodwin [2] who proposed the concept of Forming Limit Diagram, which involved a wide range of forming limit strains of a homogeneous material at different strain-states (from uniaxial tension, to plane strain tension and biaxial tension). The forming limit strain was constructed with experiments on various proportional loading paths by measuring the distortions of small grids pre-marked on the sheet surface [3-5]. However, due to the time-consuming, costly and tedious experimental work involved in FLD measurement, a theoretical prediction of the FLD has become popular and essentially important in sheet metal forming and the related industries.

Since the 1950s, based on different failure criteria, a number of analytical/theoretical models for predicting FLDs have been developed [6]. According to the concept the limit strains can be calculated on the basis of certain plastic instability criteria. One of the first criteria made in order to evaluate diffuse instability and evaluate formability of sheet material was proposed by Swift in 1952 [7]. He estimates diffuse necking to occur for an isotropic material when the major principal strain reaches limit values. This swift criterion is limited to isotropic material. To overcome this limitation Moore and Wallace [8] expanded the criterion to take into account anisotropic behavior using Hills's 48 yield criterion. Hill in 1952 [9] found that the forming limits predicted by the diffuse plastic instability criterion underestimate the forming capacity for the left hand side of forming limit diagram. To overcome this underprediction, Hill proposed a localized plastic instability criterion, which states that a local instability can occur only after the formation of a diffuse instability, the necking direction is coincident with the direction of zero-elongation, and gave the final limited strain formulation for isotropic materials. This study led to the well-known zero extension assumption, i.e. that the localization band develops along the zero extension direction in a sheet metal. This analysis predicted that localized necking would not occur in a uniform sheet, subject to positive biaxial stretching for which no zero extension direction exists. Therefore, Hill's criterion is only applicable to the LHS of the FLD. J. Majak *et al* [10] and Dudzinski et al [11] proposed a simple algorithm for anisotropic yield criteria, however, this algorithm cannot capture the shape evolution of yield surface and strength differential (SD) effect for asymmetrical material.

The purpose of this contribution is to extend the algorithm proposed by Majak and develop new theories to capture the anisotropic and asymmetrical mechanical behavior simultaneously. This algorithm is used for the limited strain determination in connection with the Swifts instability condition for diffuse necking and the Hills instability condition for localized necking.

2 THEORETICAL FRAMEWORK

2.1 Yield function

To describe both the asymmetry between tension and compression and the incompressible plastic anisotropy observed in HCP metal sheets, Cazacu and Barlat [12] introduced a general and rigorous method which is based on the theory of representation of tensor functions. A 4th order linear transformation is operated on the stress deviator $\boldsymbol{S}$ to obtain the transformed stress $\boldsymbol{\Sigma}$, which can be defined as

$$\boldsymbol{\Sigma} = \boldsymbol{CS} \tag{1}$$

where $\boldsymbol{C}$ is a constant 4th order tensor, which includes 9 independent anisotropy coefficients, Let (x,y,z) be the reference frame associated with orthotropy. In the case of a sheet, x, y and

z represent the rolling, transverse, and the normal directions. Relative to the orthotropy axes (x,y,z), the tensor $\boldsymbol{C}$ is represent by

$$C=\begin{bmatrix} C_{11} & C_{11} & C_{11} & 0 & 0 & 0 \\ C_{11} & C_{11} & C_{11} & 0 & 0 & 0 \\ C_{11} & C_{11} & C_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{11} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{11} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{11} \end{bmatrix} \tag{2}$$

Thus, the final orthotropic yield function can be given by

$$\bar{\sigma}=\left[\left(\left|\Sigma_1\right|-k\Sigma_1\right)^a+\left(\left|\Sigma_2\right|-k\Sigma_2\right)^a+\left(\left|\Sigma_3\right|-k\Sigma_3\right)^a\right]^{1/a} \tag{3}$$

where Σ_1, Σ_2, Σ_3 are the principal values of $\boldsymbol{\Sigma}$. The only restrictions imposed on the tensor $\boldsymbol{C}$ are: (i) to satisfy the major and minor symmetries and (ii) to be invariant with respect to the orthotropy group. Thus, for 3-D stress conditions the orthotropic criterion involves 9 independent anisotropy coefficients; it reduces to the isotropic criterion when $\boldsymbol{C}$ equal to the 4th order identify tensor. It is worth noting that although the transformed tensor is not deviatoric, the orthotropic criterion is insensitive to hydrostatic pressure and thus the condition of plastic incompressibility is satisfied. For $k\in[-1,1]$ and any integer $a\geq 1$, the anisotropic yield function is convex in the variables Σ_1, Σ_2, Σ_3 (principal transformed stresses).

For the sheet metal forming, the material is considered to behave as an orthotropic membrane under the plane stress conditions, thus the stress and strain component in the third direction are vanished, as

$$\sigma_{33}=0,\sigma_{13}=0,\sigma_{23}=0 \tag{4}$$

Thus the transformed stress can be rewritten as

$$\Sigma=\begin{bmatrix} \phi_1\sigma_{11}+\varphi_1\sigma_{22} \\ \phi_2\sigma_{11}+\varphi_2\sigma_{22} \\ \phi_3\sigma_{11}+\varphi_3\sigma_{22} \\ C_{44}\sigma_{12} \end{bmatrix} \tag{5}$$

In this situation, the principle stress can be rewritten as

$$\begin{aligned} \Sigma_1&=\frac{(\phi_1+\phi_2)\sigma_{11}+(\varphi_1+\varphi_2)\sigma_{22}}{2}+\sqrt{\left[\frac{(\phi_1-\phi_2)\sigma_{11}+(\varphi_1-\varphi_2)\sigma_{22}}{2}\right]^2+C_{44}^2\sigma_{12}^2} \\ \Sigma_2&=\frac{(\phi_1+\phi_2)\sigma_{11}+(\varphi_1+\varphi_2)\sigma_{22}}{2}-\sqrt{\left[\frac{(\phi_1-\phi_2)\sigma_{11}+(\varphi_1-\varphi_2)\sigma_{22}}{2}\right]^2+C_{44}^2\sigma_{12}^2} \\ \Sigma_3&=\phi_3\sigma_{11}+\varphi_3\sigma_{22} \end{aligned} \tag{6}$$

where

$$\phi_1 = \frac{1}{3}\left(2C_{11} - C_{12} - C_{13}\right)$$

$$\phi_2 = \frac{1}{3}\left(2C_{12} - C_{22} - C_{23}\right)$$

$$\phi_3 = \frac{1}{3}\left(2C_{13} - C_{23} - C_{33}\right)$$

$$\varphi_1 = \frac{1}{3}\left(-C_{11} + 2C_{12} - C_{13}\right)$$

$$\varphi_2 = \frac{1}{3}\left(-C_{12} + 2C_{22} - C_{23}\right)$$

$$\varphi_3 = \frac{1}{3}\left(-C_{13} + 2C_{23} - C_{33}\right)$$

Since the effective stress $\bar{\sigma}$ is the first order homogeneous function in stresses, from the work equivalence principle it follows that the law of evolution for the effective plastic strain (associated with $\bar{\sigma}$) reduces to $d\bar{\varepsilon} = d\gamma$.

2.2 Plastic instability formulation

The first instability criterion was proposed by Considère in 1885 [13], who analyses the formation of the necking in tensile specimens. The geometrical construction is given by the relationship

$$F = \sigma A \tag{7}$$

where F, σ and A are the force, stress and cross-sectional area respectively, and by the total differential

$$dF = \sigma dA + d\sigma A \tag{8}$$

Under tensile conditions, the force-displacement curves presents a local maximum at the beginning of the neck formation, $dF = 0$, then equations yields

Swift used Considère's criterion to determinate the limit strains in biaxial tension. The sheet element was analyzed for two perpendicular directions. Therefore the diffuse necking should be occur when the force differentials at these two perpendicular directions are equals to zero, as

$$\begin{aligned} dF_1 &= 0 \\ dF_2 &= 0 \end{aligned} \tag{9}$$

After a mathematical manipulation by taking the plastic incompressibility into account, the diffuse constraints can be written as

$$\begin{aligned} d\sigma_1 &= \sigma_1 d\varepsilon_1 \\ d\sigma_2 &= \sigma_2 d\varepsilon_2 \end{aligned} \tag{10}$$

Combining with proportional loading constraints, the Swift's diffuse necking condition can be presented in tensor notations as

$$d\sigma_i = \sigma_1 A^{Swift} d\varepsilon_j \tag{11}$$

where the Swift's instability tensor can be calculated by

$$A_{ij}^{Swift}(\alpha) = \begin{bmatrix} 1 & 0 \\ \alpha & 0 \end{bmatrix}.$$

where α represents the stress ratio ($\alpha = \sigma_2 / \sigma_1$).

For analysis of the negative minor strain region of FLD, the localized necking develops along one direction, which is inclined with respect to the loading direction. Hill's localized necking condition is employed. According to Hill's instability condition the localized necking (through thickness neck) occurs when the rate of strain hardening is equal to the rate of geometric softening.

$$dT = d\sigma_1 t + \sigma_1 dt = 0 \tag{12}$$

where t represents the thickness of the sheet.

Taking the strain though thickness place of in-plane strain, the constraints of localized necking are written as following

$$\begin{aligned} d\sigma_1 &= \sigma_1 \left(d\varepsilon_1 + d\varepsilon_2 \right) \\ d\sigma_2 &= \alpha d\sigma_1 \end{aligned} \tag{13}$$

Similarly, we can written it as a compact form, the Hill's localized necking condition reads

$$d\sigma_i = \sigma_1 A_{ij}^{Hill} d\varepsilon_j \tag{14}$$

where A_{ij}^{Hill} is the Hill's instability tensor, which can be calculated by

$$A_{ij}^{Hill}\left(\alpha\right) = \begin{bmatrix} 1 & 1 \\ \alpha & \alpha \end{bmatrix} \tag{15}$$

After comparison with Swift instability tensor, it is clear seen that two instability condition has similar formulation, so here we can written it as a generic way by

$$d\sigma_i = \sigma_1 A^{instability} d\varepsilon_j \tag{16}$$

where $A^{instability}$ is taken equal with A_{ij}^{Hill} and A^{Swift}, respectively.

The mechanical response of the sheet metal will be described by a rigid-plastic model. Hence for the total strains and total strain increments are equal to the corresponding plastic strains and plastic strain increments, respectively. Given the strain history from the previous step, the effective strain increment at current step, strain path β (stress path α), stress, strain, strain increment in region (a) can be calculated. The main ingredient of the constitutive model is the yield function:

$$\Phi = \bar{\sigma}\left(\sigma_{11}, \sigma_{22}\right) - \sigma_Y\left(\bar{\varepsilon}\right) = 0 \tag{17}$$

where σ_Y represent equivalent stress and is calculated from the hardening law.

If the normality flow rule is assumed to hold for plastic flow of materials, according to the consistency condition $d\Phi = 0$, the differential yield function can be written as

$$d\Phi = \frac{\partial \bar{\sigma}}{\partial \bar{\varepsilon}} d\bar{\varepsilon} + \frac{\partial \bar{\sigma}}{\partial \sigma_i} d\sigma_i - \frac{\partial \sigma_Y}{\partial \bar{\varepsilon}} d\bar{\varepsilon} = 0 \tag{18}$$

By applying classical plasticity theory the instability criterion is derived in above subsection, the plastic anisotropy, associated plastic flow law and hardening parameters, the final plastic instability function should be written as

$$\frac{1}{\bar{\sigma}} \frac{\partial \bar{\sigma}}{\partial \bar{\varepsilon}} + \frac{1}{f} \frac{\partial \bar{\sigma}}{\partial \sigma_i} A_{ij}^{instability} \frac{\partial \bar{\sigma}}{\partial \sigma_j} - \frac{1}{\sigma_Y} \frac{\partial \sigma_Y}{\partial \bar{\varepsilon}} = 0 \tag{19}$$

where the function f is given by ratio $f = \bar{\sigma} / \sigma_1$.

Due to the equivalent stress is a homogeneous function of degree one with respect to the stress components, the first item of the above equation is only dependent on the equivalent plastic strain. After solving this equation by using iterative methods, the limit strains $\bar{\varepsilon}^*$ can be obtained.

3 RESULTS AND DISCUSSION

The material used in this contribution was high-purity (99%) titanium. The initial texture of the as received material has a basal texture, with the majority of the grains having their c-axis at 30^0 to the normal to the plane of the plate. It is considered that the materials'hardening is isotropic and governed by the equivalent plastic strain according to a power-law:

$$\sigma_Y = K\left(\varepsilon_0 + \bar{\varepsilon}\right)^n \tag{20}$$

where K, ε_0 and n are materials parameters.

All the material parameters can be seen in Table 1 and Table 2.

$\sigma_Y = K\left(\varepsilon_0 + \bar{\varepsilon}\right)^n$		
K	ε_0	n
413MPa	0.6445	1.0

Table 1: Materials parameters of pure titanium.

$\bar{\varepsilon}$	k	C_{22}	C_{33}	C_{12}	C_{13}	C_{23}	C_{44}
0.02	-0.304	0.971	1.316	0.022	0.189	0.152	0.972
0.05	-0.313	0.989	1.243	0.089	0.193	0.173	0.909
0.1	-0.363	0.992	1.046	0.016	0.075	0.053	0.983
0.15	-0.419	0.996	0.915	-0.015	0.021	0.000	1.016
0.2	-0.472	0.998	0.849	-0.048	-0.012	-0.034	1.050
0.25	-0.518	0.998	0.815	-0.089	-0.041	-0.068	1.092
0.3	-0.554	0.998	0.797	-0.130	-0.068	-0.099	1.134
0.35	-0.635	1.000	0.772	-0.178	-0.097	-0.135	1.183

Table 2: Coefficients of Cazacu06 for pure titanium [14].

The yield surface evolution at different deformation level was shown in Figure1. From there, it is clear seen that the shape of yield surface changed when the material suffer different deformation. It leads to the identified anisotropic parameters in Cazacu06 model at different equivalent plastic deformations have different value, and equivalent stress value is not only dependent on the stress, but also the equivalent plastic strain. Therefore, it is necessary to take this effect into account when we use the analytical method to predict formability.

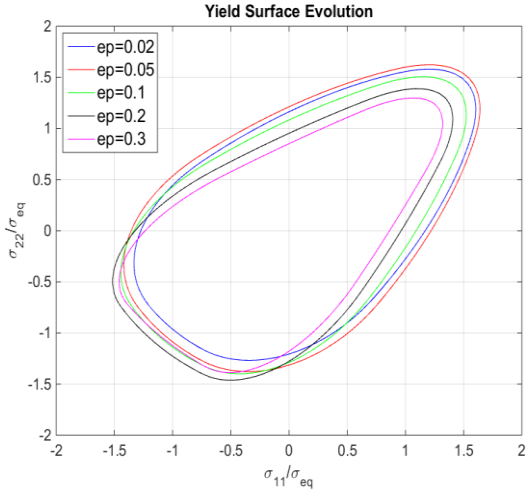


Figure 1: Yield surface evolution of pure titanium.

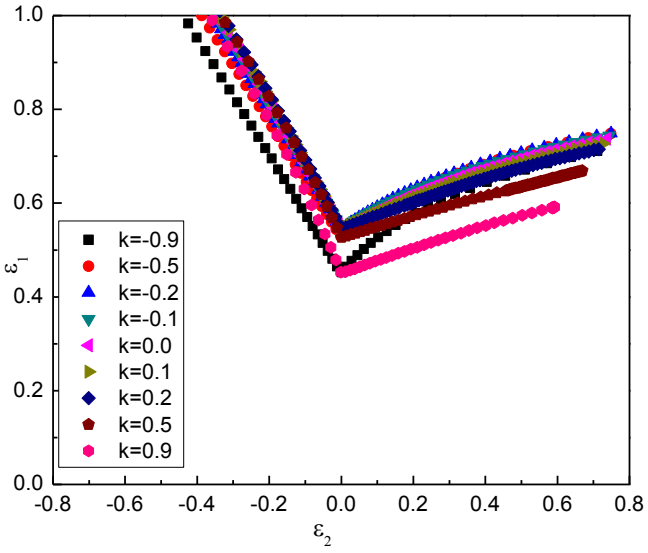


Figure 2: FLD for different *k* value.

To determine a complete FLD, Swift's and Hill's theories are used calculate the forming limit strains on the left and the right side, respectively, of the FLD. Before start to study yield surface evolution, we want to demonstrate the influence of material parameter *k* in Cazacu06 yield function on the FLD, which can be seen in Figure 2. By comparison of FLD at different *k* values, it can be seen that the coefficient k has great effect on the forming limited strain at the right hand side of FLD.

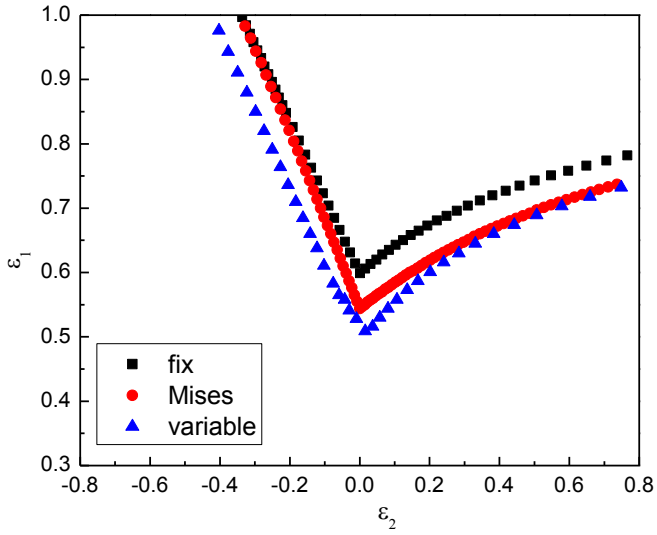


Figure 3: FLD for different yield surfaces.

Figure 3 displays the predicted FLD for von Mises yield criterion, Cazacu06 with fixed anisotropic coefficients (it means that the shape of the yield surface doesn't changes when the materials surfer plastic deformation), Cazacu06 with different anisotropic coefficients for dif-

ferent deformation level. It is clear seen that the Cazacu06 with fixed coefficient overestimated the formability of the materials, particularly for the right hand side of FLD.

4 CONCLUSIONS

In this paper, an extended Swift and Hill's plastic instability by considering the yield surface evolution was proposed and implemented into a unified instability analytical algorithm. The shape of the yield surface has a great influence on the forming limit strains. By considering the yield surface evolution, it can reduce the overestimation of limited strains.

REFERENCES

[1] S.P. Keeler, W.A Backofen, Plastic instability and fracture in sheet stretched over grid punches, *American Society of Metals Transactions* Quarterly 56 25-48, 1964.

[2] G.M. Goodwin, Application of strain analysis to sheet metal forming in the press shop, *Society of Automotive Engineers*, 1968.

[3] H.J. Kleemola, and J.O.Kumpulainen, Factors influencing the forming limit diagram: Part I – The experimental determination of the forming limits of sheet steel, *Journal of Materials Processing Technology*, 3, 289-302, 1980.

[4] H.J. Kleemola, and J.O.Kumpulainen, Factors influencing the forming limit diagram: Part II – Influence of sheet thickness, *Journal of Mechanical Working Technology*, 3(3-4), 303-311, 1980.

[5] ASTM Standard E2218-02, Standard Test Method for Determining Forming Limit Curves, 2002.

[6] D. Banabic, H.J. Bünge, K. Pöhlandt, A.E. Tekkaya, Formability of Metallic Materials I, Plastic Anisotropy, Formability Testing and Forming Limits, Springer Verlag, Heidelberg, pages 173-213, 2000.

[7] H.W. Swift, Plastic instability under plane stress, *Journal of the Mechanics and Physics of Solids*, 1,1:1-18,1952.

[8] G.G. Moore and J. F. Wallance, The effect of anisotropy on instability in sheet metal forming, *Journal of the Institute of Metals*, 9:33-38, 1965.

[9] R.Hill, On discontinuous plastic states, with special reference to localized necking in thin sheets, *Journal of the Mechanics and Physics of Solids*, 1:19-30, 1952.

[10] J. Majak, M.Pohlak, R.Kuttner, A simple algorithm for formability analysis, *Journal of Achievements in Materials and Manufacturing engineering*, v22 57-60, 2007.

[11] D. Dudzinski, A. Molinari, Perturbation analysis of thermoviscoplastic instabilities in biaxial loading, *International Journal of Solids and Structures*, 27(5): 601–628, 1991.

[12] O. Cazacu, B. Plunkett, Orthotropic yield criterion for hexagonal closed packed metals, International Journal of Plasticity. 22(7), 1171-1194, 2006.

[13] A. Considère, Annales des Ponts et Chaussées, 9: 574-775, 1885.

[14] B. R. Baudard, O. Cazacu, P. Flater, N. Chandola, J. L. Alves, Unusual plastic deformation and damage features in Titanium: Experimental test and constitutive modeling, *Journal of the Mechanics and Physics of Solids*, V 88, 100-122, 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MULTISCALE SIMULATION OF THREE-DIMENSIONAL UNSTEADY GAS FLOWS IN MICROCHANNELS OF TECHNICAL SYSTEMS

Viktoriia O. Podryga[1], Yury N. Karamzin[1], Tatiana A. Kudryashova[1], Sergey V. Polyakov[1]

[1] Keldysh Institute of Applied Mathematics (RAS)
Miusskaya square, 4, 125047, Moscow, Russia
e-mail: pvictoria@list.ru

Keywords: Gas Dynamics, Molecular Dynamics, Microchannels, Parallel Calculations, Nitrogen Flows.

Abstract. *The work is devoted to modeling the gas flows in microchannels of technical systems in conditions of multiscale computational domain. As an example, the problem of nitrogen flow in nickel microchannel for three-dimensional geometry is considered. General attention is paid to the calculation of gaseous medium macroparameters considering the molecular processes that occur in the gas flow and on the walls of the microchannel. The difference in the scales of the computational domain (the length of the channel, cross section of the channel, the free path of the molecules, the thickness of the boundary layer) and near-surface interaction of the gas with the metal lead to the necessity taking into account the relief and the properties of the microchannel at the molecular level. As a result, the mathematical model of the research flow can not be fully formulated within the framework of the macroscopic approach. For decision of the problem multiscale approach combining the solution of a quasigasdynamic (QGD) equations and correction of gasdynamic parameters by molecular dynamics method (MD), is used. QGD system of equations is solved by method of finite volumes. The MD system of equations is applied within each control volume and is solved by Verlet scheme. In MD calculations particles interactions are described by the potentials determining the basic properties of the components of the considered system. Parallel implementation of the approach is based on method of splitting into physical processes and separation of areas. Algorithms are focused on the use of computer systems with central and hybrid architectures. Calculations showed that the overall numerical algorithm is resistant to the use of data for the flow correction obtained by the MD calculations. With the help of MD methods basic coefficient relations for QGD system were obtained, the transitions from MD to QGD and back were checked, three-dimensional calculation of the nitrogen flow in the nickel microchannel was produced. The results confirmed the efficiency of the developed approach.*

1 INTRODUCTION

Present paper is devoted to the description and the parallel program realization of the multiscale approach intended for modeling of nonlinear processes of gas mixture transportation via the technical systems microchannels. The physics for functioning of similar systems is usually described by whole hierarchy of mathematical models descending down to the atomic level. As a result of such association the possibility appears to make an exact predicting the properties of the modeled objects and processes. However at the same time the level of computing complexity of specific tasks significantly increases that becomes surmountable only by using powerful clusters and supercomputers.

As an example the supersonic cold gasdynamic spraying (SCGS) the nanoparticles on a substrate [1] is considered. This problem is relevant for many directions of nanotechnology, including the production of new materials in electronics and medicine. The main technological process in the SCGS installations is called a nanoimprinting, it is accelerating nanoparticles by a supersonic flow and delivering them to the substrate warmed up to the necessary temperature where sticking occurs. Among the various aspects of the problem let's distinguish a transport question that is related to the calculations of gas flow parameters in all parts of the SCGS installation. The purpose of such calculation is to determine the optimum flow conditions.

Features of a transport problem are connected with a spraying installation design. In the simplified look the SCGS installation consists of a Dewar bottle up to several tens liters in lower part of which there is the mobile work desk with a set of replaceable substrates. Over a table there is a subsystem supplying gas and nanoparticles, this subsystem is attached to the top wall of the reservoir. The free space in the reservoir is filled with either technical vacuum, or strongly rarefied gas. The subsystem giving the nanoparticles consists of ballons with pure gas mixture and gas mixture with nanoparticles, of micronozzles matrix and of microchannels on which nanoparticles are fed to a near area of a substrate. Thus a pure gas mixture is used to manage the process, and the gas mixture with nanoparticles is used to give them a start pulse.

From the mathematical point of view the transport problem consists in calculating the gas and nanoparticles motions in all parts of installation. It is necessary to consider that the simulated environment near solid surfaces is not continuous, and the equation of state for gas mixture with nanoparticles is unknown. Modeling of a transport problem in full leads to take into account the real geometry of the SCGS installation and all physical processes proceeding in it in different scales on time and space. It demands too big computing expenses.

Within this work the following combined approach is offered. It combines calculations of the gaseous medium flow in specific microchannels and calculations of interacting the streams with a substrate, and also movements of nanoparticles in a gas flow and their interaction with a substrate at contact with the deposition place. At that simulation is carried out on two scale levels – the basic level having the characteristic sizes from several tens microns to tens of millimeters, and additional level having the characteristic sizes about a micron or less. At the basic level (in macroscale) gas and nanoparticles flows in all parts of installation are calculated, and also the external factors connected with management of spraying process are taken into account. At the additional level (in microscale) the interactions are calculated for: 1) gas molecules among themselves (forming the equation of state for mixture and realizing the mixing of components), 2) gas molecules and solid surfaces atoms (describing phenomena in boundary layers), 3) gas molecules and nanoparticles atoms (describing formation of boundary layers on a surface of nanoparticles), 4) gas molecules and substrate atoms (considering processes of heat exchange with a substrate), 5) atoms of nanoparticles with the substrate atoms (processes of repulsion, sticking, implementation of nanoparticles into substrate).

The specified physical processes are complex and multistage. Therefore at this stage of researches the simplified description of a problem is used. It assumes splitting into physical processes. The main simulation loop represents alternation in time of calculations at the macro- and microlevels. In the represented embodiment of research the calculation of nitrogen - hydrogen mixture flow after leaving the nozzle and exiting to the half-open microchannel and further in free space is considered (Fig. 1).

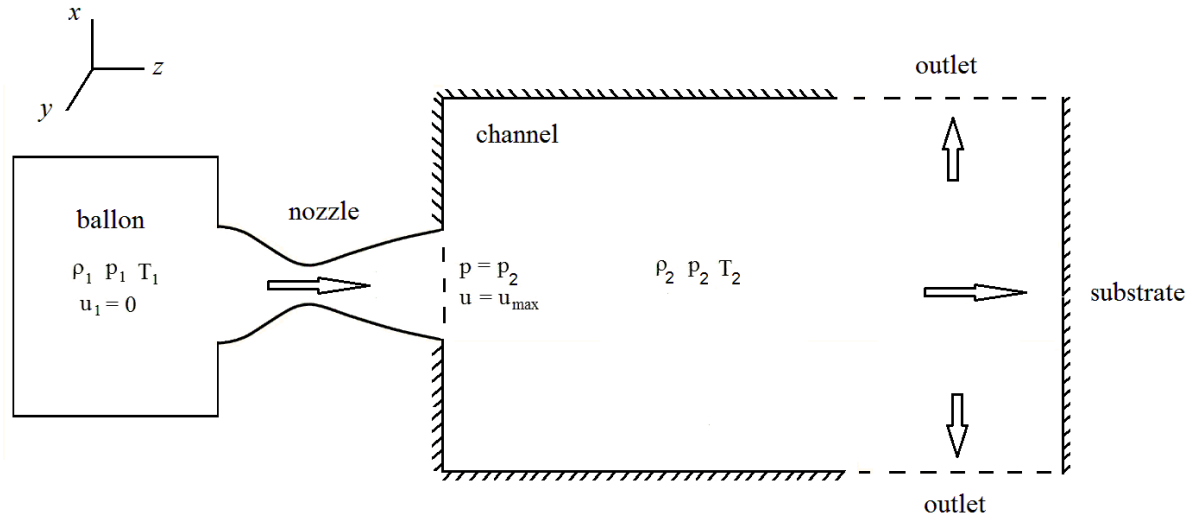


Figure 1: The geometry of a model problem for convenience presented in the form of a two-dimensional section. From a tank through a micronozzle in microchannel the gas mixture is supplied. Gas mixture is used to manage the nanoparticles trajectories for delivering them in the right places of a substrate.

Some part of research has been done in the previous works [2-6]. In particular, in [2-5] methods of thermodynamic equilibrium calculations in metal-metal, gas-gas and gas-metal systems were developed. In work [6] association micro- and macromodels of the gas environment in uniform object of research was proposed and partially approved. The main objective of this research is development of a full three-dimensional multiscale numerical method and parallel algorithm for its realization, suitable for calculating of gas mixture flow in the microchannel and near a substrate taking into account the boundary effects resolved by MD [7]. Description of the model and the algorithm details is given in the following sections.

2 MATEMATICAL MODELS

2.1 Macroscopic model

The macroscopic model of the flow in all parts of installation is based on the quasigasdynamic (QGD) equations [8]. The choice of QGD equations is associated with two factors. Firstly, QGD equations well proved in calculating supersonic flows of the rarefied viscous and heat-conducting gases. Secondly, they are well suited to calculating the flows in microsystems as the natural parameter of a nondimensionalization in them is the mean free path. Thirdly, QGD system of the equations is applicable in the wide range of Knudsen numbers that gives the chance to model complex system of microchannels with different diameters, using the same mathematical model.

In case of gas mixture the QGD system of equations is written for each gas separately and has an identical form [8-9]. These equations in case of binary mixture in invariant concerning system of coordinates with the constraint and the equations of state have the form:

$$\frac{\partial \rho_l}{\partial t} + div\, \mathbf{W}_l^{(\rho)} = 0,$$
$$\mathbf{W}_l^{(\rho)} = \rho_l \mathbf{u}_l - \rho_l \mathbf{w}_l, \quad \mathbf{w}_l = \tau \left[\left(\mathbf{u}_l, \nabla\right)\mathbf{u}_l + \frac{1}{\rho_l}\nabla p_l \right], \quad l = a, b; \tag{1}$$

$$\frac{\partial}{\partial t}\rho_l u_{l,k} + div\, \mathbf{W}_l^{(\rho u_k)} = S_l^{(\rho u_k)},$$
$$\mathbf{W}_l^{(\rho u_k)} = \rho_l \mathbf{u}_l u_{l,k} + \mathbf{e}_k \left(p_l + \frac{2}{3}\mu_l\, div\, \mathbf{u}_l \right) - \tag{2}$$
$$- \mu_l \left(\nabla u_{l,k} + \left(\nabla, \mathbf{e}_k\right)\mathbf{u}_l \right) - \left(\rho_l w_{l,k} \mathbf{u}_l + \rho_l \mathbf{w}_l u_{l,k} \right),$$
$$S_l^{(\rho u_k)} = \nu_{ll'} \rho_l \left(u_{l,k}^* - u_{l,k} \right), \quad l = a, b, \quad l' = b, a, \quad k = x, y, z;$$

$$\frac{\partial}{\partial t} E_l + div\, \mathbf{W}_l^{(E)} = S_l^{(E)},$$
$$\mathbf{W}_l^{(E)} = \left(\rho_l \mathbf{u}_l - \rho_l \mathbf{w}_l \right) H_l - \chi_l \nabla T_l + \left(\frac{2}{3}\mu_l\, div\, \mathbf{u}_l \right)\mathbf{u}_l - \tag{3}$$
$$- \sum_{k=x,y,z} \mu_l \left(\nabla u_{l,k} + \left(\nabla, \mathbf{e}_k\right)\mathbf{u}_l \right) + \left(\rho_l \mathbf{w}_l, \mathbf{u}_l \right)\mathbf{u}_l,$$
$$S_l^{(E)} = \nu_{ll'} \rho_l \left(E_l^* - E_l \right), \quad l = a, b, \quad l' = b, a,$$

$$E_l = \frac{1}{2}\rho_l \left|\mathbf{u}_l\right|^2 + \rho_l \varepsilon_l, \quad H_l = \frac{E_l + p_l}{\rho_l}, \tag{4}$$
$$p_l = Z_l \rho_l \Re_l T_l, \quad \varepsilon_l = c_{V,l} T_l, \quad l = a, b.$$

Here we assume that gas mixture consists of a and b gases with numerical densities (concentrations) n_a and n_b and mass densities $\rho_a = m_a n_a$ and $\rho_b = m_b n_b$, where m_a and m_b – masses of gas molecules a and b. Each gas is characterized by its temperature T_l and macroscopic velocity $\mathbf{u}_l$. Other parameters of mixture components: p_l – partial pressures, E_l – total energy densities, ε_l – specific internal energies, H_l – total enthalpies, $\mu_l = \mu_l\left(T_l\right)$ and $\chi_l = \chi_l\left(T_l\right)$ – coefficients of viscosity and heat conduction, $Z_l = Z_l\left(T_l\right)$ – compressibility coefficients, $c_{V,l} = c_{V,l}\left(T_l\right)$ – specific heat capacities at constant volume, $\Re_l = k_B / m_l$ – gas constants (k_B – Boltzmann constant). Vectors $\mathbf{W}_l^{(\rho)}$, $\mathbf{W}_l^{(\rho u_k)}$, $\mathbf{W}_l^{(E)}$ up to a sign are identical to the density flux of the corresponding components of the momentum density and energy density, $\rho_l \mathbf{w}_l$ – QGD correction vectors to the density flux, that are proportional to the Maxwell relaxation time τ for gas mixtures, $\mathbf{e}_k$ – unit vectors.

The exchange members $S_l^{(\rho u_k)}$ and $S_l^{(E)}$ take into account the redistribution of momentum and energy between the mixture components. They contain components of velocity $\mathbf{u}_l^*$ and energy E_l^*, calculated at the molecular level (see Section 2.2), as well as $\nu_{ll'}$ –frequency of mutual collisions between molecules of variety l with molecules of variety l', which are calculated via ν_l – frequency of mutual collisions between the molecules of one type:

$$\nu_{ab} = \nu_a \left(\frac{d_{ab}}{d_a} \right)^2 \sqrt{\frac{m_a + m_b}{2m_b}} \frac{n_b}{n_a}, \quad \nu_{ba} = \nu_{ab} \frac{n_a}{n_b} = \nu_b \left(\frac{d_{ab}}{d_b} \right)^2 \sqrt{\frac{m_a + m_b}{2m_a}} \frac{n_a}{n_b},$$
$$d_{ab} = \frac{d_a + d_b}{2}, \quad \nu_l = \frac{p_l}{\mu_l} \Omega_l, \quad l = a, b. \tag{5}$$

where d_l – molecules diameters of variety l, Ω_l – dimensionless parameters characterizing the collision molecules of type l [10].

Coefficients of viscosity μ_l, heat conduction χ_l and compressibility Z_l, and the specific heats $c_{V,l}$ and the mean free path λ_l for the mixture components are determined from a database of molecular calculations or by direct MD computations (see Section 2.2). The local sound velocities for the mixture components are calculated according to the formulas:

$$a_l = \sqrt{\gamma_l Z_l \Re_l T_l}, \quad \gamma_l = 1 + \frac{\Re_l}{c_{v,l}}, \quad l = a, b, \tag{6}$$

where $\gamma_l = \gamma_l(T_l)$ – adiabatic indexes.

The parameters of the gas mixture at macrolevel are defined as follows:

$$\rho = \rho_a + \rho_b, \quad \mathbf{u} = \frac{\rho_a \mathbf{u}_a + \rho_b \mathbf{u}_b}{\rho}, \quad p = p_a + p_b, \quad E = E_a + E_b, \quad n = n_a + n_b,$$
$$T = \frac{n_a Z_a T_a + n_b Z_b T_b}{nZ}, \quad m = \frac{m_a n_a + m_b n_b}{n}, \quad Z = Z_a + Z_b, \quad \Re = \frac{p}{\rho Z T} = \frac{k_B}{m}, \quad \tau = \frac{\mu}{p}. \tag{7}$$

To determine the viscosity of the mixture we use the Buddenberg-Wilke formula [11]:

$$\mu = \mu_a \left(1 + G_{ab} \frac{\rho_b}{\rho_a} \frac{M_a}{M_b} \right)^{-1} + \mu_b \left(1 + G_{ba} \frac{\rho_a}{\rho_b} \frac{M_b}{M_a} \right)^{-1},$$
$$G_{ab} = \left(1 + \sqrt{\frac{\mu_a}{\mu_b} \sqrt{\frac{M_b}{M_a}}} \right)^2 \left(2\sqrt{2\left(1 + \frac{M_a}{M_b} \right)} \right)^{-1}. \tag{8}$$

Here M_a and M_b – molar masses of gases a and b. Value G_{ba} is calculated similarly by cyclic substitution of the indexes. The average local sound velocity, and the mean free path, the Mach number and the Reynolds number for the mixture are equal to:

$$a = \frac{m_a a_a + m_b a_b}{m_a + m_b}, \quad \lambda = \frac{m_a \lambda_a + m_b \lambda_b}{m_a + m_b}, \quad Ma = \frac{|\mathbf{u}|}{a}, \quad \mathrm{Re} = \frac{\rho |\mathbf{u}| \lambda}{\mu}. \tag{9}$$

The system of equations (1)-(3) is closed initial and boundary conditions. The initial conditions correspond to the equilibrium gas environment in the absence of interaction with external factors:

$$\rho_l = \rho_{l,0}, \quad \mathbf{u}_l = 0, \quad p_l = p_{l,0}, \quad T_l = T_0, \quad l = a, b. \tag{10}$$

Here $\rho_{l,0}$, $p_{l,0}$ – the initial densities and pressures of gas components, T_0 – initial temperature.

The area of a nozzle and a tank with gas mixture were not included into calculation. It was assumed that in an initial time the partition between the nozzle and the area of the microchannel opened, and gas from area of a high pressure began to come to the area of low

pressure. The nozzle was considered as an ideal adiabatic Laval nozzle. Parameters of the gas environment at the exit from a nozzle (that is on an entrance to the medium) were calculated according to the known formulas:

$$\begin{aligned}
&(\mathbf{u}_l,\mathbf{n}) = u_{in,l}, \quad \rho_l = \rho_{in,l}, \quad T_l = T_{in,l}, \quad l = a,b;\\
&p_{in,l}\left(m_l Z_l\left(T_{in,l}\right)\Re_l T_{in,l}/p_{in,l}\right)^{\gamma_l\left(T_{in,l}\right)} = p_{bal,l}\left(m_l Z_l\left(T_{bal}\right)\Re_l T_{bal}/p_{bal,l}\right)^{\gamma_l\left(T_{bal,l}\right)},\\
&u_{in,l} = \sqrt{2\left(H_{bal,l} - H_{in,l}\right)}, \quad \rho_{in,l} = p_{in,l}/Z_l\left(T_{in,l}\right)\Re_l T_{in,l},\\
&H_{bal,l} = \left(c_{v,l}\left(T_{bal}\right) + Z_l\left(T_{bal}\right)\Re_l\right)T_{bal}, \quad H_{in,l} = \left(c_{v,l}\left(T_{in,l}\right) + Z_l\left(T_{in,l}\right)\Re_l\right)T_{in,l}.
\end{aligned} \tag{11}$$

Here the parameters with index *bal* correspond to the values of parameters in the tank, and the parameters with index *in* – the respective parameters on the entrance to the environment.

At the microchannel walls and on the substrate following boundary conditions are imposed:

$$\begin{aligned}
&\left(\mathbf{W}_l^{(\rho)},\mathbf{n}\right) = -\alpha_l\left(\rho_l - \rho_l^{(w)}\right), \quad \left(\mathbf{W}_l^{(\rho u_k)},\mathbf{n}\right) = -\beta_{l,k}\left(\rho_l u_{l,k} - \rho_{l,w} u_{l,k}^{(w)}\right), \quad k = x,y,z,\\
&\left(\mathbf{W}_l^{(E)},\mathbf{n}\right) = -\eta_l\left(E_l - E_l^{(w)}\right), \quad l = a,b.
\end{aligned} \tag{12}$$

Generally the gas molecules can get into a metal wall, or be reflected from it at some corner. Interaction with a wall can be elastic or inelastic, thermal or isothermal. It is convenient near a wall to enter the special microsystem consisting of gas molecules and metal atoms. In this microsystem gas components have densities of masses $\rho_l^{(w)}$, moments $\rho_l^{(w)}\mathbf{u}_l^{(w)}$ and total energy $E_l^{(w)}$. Boundary conditions on a wall can be written down in the form of third-type boundary conditions describing an exchange of mass, momentum and energy components between the gas mixture in the flow and near the walls. Coefficients α_l, $\beta_{l,k}$ and η_l, associated with accommodation coefficients [12], The coefficients associated with the coefficients accommodation are priori unknown but can be determined from either physical quantities tables (which is possible only to a limited range of temperatures and pressures), or can be calculated using molecular dynamics. The second approach seems the most justified as it is universal and will help to coordinate the interaction processes on micro- and macrolevels.

On the free surfaces a "soft" boundary conditions are given [8]:

$$\frac{\partial\rho_l}{\partial n} = 0, \quad \frac{\partial\left(\rho_l\mathbf{u}_l\right)}{\partial n} = 0, \quad \frac{\partial p_l}{\partial n} = 0, \quad l = a,b. \tag{13}$$

In the case of pumping the mixture out of the system in (13) the minimum pressure is given instead of the last condition: $p_l = p_{\min}$.

2.2 Microscopic model

The microscopic model can be used for various purposes. First, it can help to clarify the equation of state (both on pressure, and on energy) and kinetic coefficients (viscosity, heat conduction, diffusion, etc.) used in QGD equations. Secondly, the microscopic model can be used to calculate the exchange members in the equations for momentum and energy. Thirdly, the microscopic model is especially demanded in calculation of the wall interactions of gas molecules with walls atoms. In this work all listed above situations are considered.

In a case when at the macrolevel the flow of the binary gas mixture without nanoparticles is calculated, at the microlevel it is also possible to be limited to consideration the molecules of two types *a* and *b*. However it is only valid far from the walls of the microchannel. Near

the walls it is necessary to add to consideration the particles of type c, making a material of walls. In the presence of nanoparticles it is necessary to add to consideration the particles of other types.

In this work we consider the case of pure binary mixture and one type of metal particles. Then at the microlevel the gas-metal system is represented as a set of particles which behavior is described by Newton's equations [7]:

$$m_{l,i}\frac{d\mathbf{v}_{l,i}}{dt}=\mathbf{F}_{l,i},\quad \mathbf{v}_{l,i}=\frac{d\mathbf{r}_{l,i}}{dt},\quad i=1,...,N_l,\quad l=a,b,c, \tag{14}$$

where i – particle number, $l=a,b,c$ – particle type (a – molecules of first gas, b – molecules of second gas, c – metal atoms), N_l –total particles number of type l, $m_{l,i}$ – particle mass of type l with number i, $\mathbf{r}_{l,i}=\left(r_{x,l,i},r_{y,l,i},r_{z,l,i}\right)$ and $\mathbf{v}_{l,i}=\left(v_{x,l,i},v_{y,l,i},v_{z,l,i}\right)$ – position vector and velocity vector of the i-th particle of type l, $\mathbf{F}_{l,i}=\left(F_{x,l,i},F_{y,l,i},F_{z,l,i}\right)$ – the total force acting on a this particle.

The forces are the sum of the component of i-th particle interaction with the surrounding particles and the component responsible for external action:

$$\mathbf{F}_{l,i}=-\frac{\partial U\left(\mathbf{r}_{l,1},...,\mathbf{r}_{l,N_l}\right)}{\partial\mathbf{r}_{l,i}}+\mathbf{F}_{l,i}^{ext},\quad i=1,...,N_l,\quad l=a,b,c, \tag{15}$$

where U – total potential energy, $\mathbf{F}_{l,i}^{ext}$ – force of interaction with the environment.

Potential energy of the system depends on particles coordinates and describes the interaction between them. The choice of interaction potential is based on comparison of mechanical properties of potential model and real material. For the solution of an objective it is necessary to consider interactions gas-gas, metal-metal and gas-metal:

$$\begin{aligned}
&U=U_{aa}+U_{bb}+U_{cc}+U_{ab}+U_{ac}+U_{bc},\\
&U_{aa}=\sum_{i=1,j>i}^{N_a}\varphi_{aa}\left(\left|\mathbf{r}_{a,i}-\mathbf{r}_{a,j}\right|\right),\quad U_{bb}=\sum_{i=1,j>i}^{N_b}\varphi_{bb}\left(\left|\mathbf{r}_{b,i}-\mathbf{r}_{b,j}\right|\right),\\
&U_{cc}=\sum_{i=1}^{N_c}\left[\varphi_{1,cc}\left(\mathbf{r}_{c,i}\right)+\sum_{j>i}\varphi_{2,cc}\left(\left|\mathbf{r}_{c,i}-\mathbf{r}_{c,j}\right|\right)\right],\quad U_{ab}=\frac{1}{2}\sum_{i=1}^{N_a}\sum_{j=1}^{N_b}\varphi_{ab}\left(\left|\mathbf{r}_{a,i}-\mathbf{r}_{b,j}\right|\right),\\
&U_{ac}=\frac{1}{2}\sum_{i=1}^{N_a}\sum_{j=1}^{N_c}\varphi_{ac}\left(\left|\mathbf{r}_{a,i}-\mathbf{r}_{c,j}\right|\right),\quad U_{bc}=\frac{1}{2}\sum_{i=1}^{N_b}\sum_{j=1}^{N_c}\varphi_{bc}\left(\left|\mathbf{r}_{b,i}-\mathbf{r}_{c,j}\right|\right).
\end{aligned} \tag{16}$$

Each type of interaction is described by a corresponding potential $\varphi_{ll'}$. For hydrogen-nitrogen mixture it was used Mi's potential in the form of "n-6" [13] adapted in work [14] to calculations of hydrogen and nitrogen mixture depending on their volume ratio in the allocated microvolume. For interaction of nickel atoms among themselves it was used the form of EAM [15] potential which considers not only pair interactions $\varphi_{2,cc}$, but also the impact of the environment on a particular particle $\varphi_{1,cc}$. To account the gas-metal interactions a standard potential Lennard-Jones [16] and Morse potential [17] were used.

The calculation of the compressibility factor and heat capacity is described in detail in [4]. The coefficients of viscosity and heat conduction were calculated according to the formulas [18]:

$$\mu_l = \frac{1}{k_B T_l V}\int_0^\infty \left\langle \left(p_{\alpha\beta,l}(t_0+t)\cdot p_{\alpha\beta,l}(t_0)\right)\right\rangle dt,\ \ \alpha\beta = xy, yz, zx,\ \ l = a,b,$$
$$p_{\alpha\beta,l} = \sum_i m_{l,i}\cdot v_{\alpha,l,i}\cdot v_{\beta,l,i} + \sum_i\sum_{j>i}\left(r_{\alpha,l,i} - r_{\alpha,l,j}\right)\cdot\left(F_{\beta,l,i} - F_{\beta,l,j}\right),\ \ i,j = 1...N_l,$$
$$\chi_l = \frac{1}{k_B T_l^2 V}\int_0^\infty \left\langle q_{\alpha,l}(t_0+t)\cdot q_{\alpha,l}(t_0)\right\rangle dt,\ \ \alpha = x,y,z, \tag{17}$$
$$q_{\alpha,l} = \frac{d\delta\varepsilon_{\alpha,l}}{dt},\ \ \delta\varepsilon_{\alpha,l} = \sum_i r_{\alpha,l,i}\left(\varepsilon_{l,i} - \left\langle\varepsilon_{l,i}\right\rangle\right),\ \ \varepsilon_{l,i} = \frac{m_l\,\mathbf{v}_{l,i}^{\,2}}{2} + \frac{1}{2}\sum_{j\neq i}^{N}\varphi_{ll}\left(\left|\mathbf{r}_{l,i} - \mathbf{r}_{l,j}\right|\right),$$

where V – volume of the investigated system, $\varepsilon_{l,i}$ – instantaneous total energy of molecule i of type l, t_0 – initial time in calculating the coefficients.

The mean free path is calculated according to the formulas [19]:

$$\lambda_a = \frac{1}{n_a\pi d_a^2\sqrt{2} + n_b\pi d_{ab}^2\sqrt{1+\frac{m_a}{m_b}}},\quad \lambda_b = \frac{1}{n_b\pi d_b^2\sqrt{2} + n_a\pi d_{ab}^2\sqrt{1+\frac{m_b}{m_a}}}, \tag{18}$$

where σ_l – effective interaction radius of molecules of type l [14].

The initial conditions at the microlevel are defined by an equilibrium thermodynamic state of particles system. Calculations of gas environment usually start with normal conditions or at first it is reduced to thermodynamic equilibrium at a given temperature and pressure. Metal parts of a microsystem must have at the start the lattice structure according to metal at given temperature. They should be no excessive stress. For more information about calculating the equilibrium state in metal-metal, gas-gas and gas-metal microsystems see the works [2-5].

The boundary conditions at the molecular level are selected depending on the simulated situation. In particular, when calculations are aimed on determining the properties of the medium, a certain allocated three-dimensional its volume is considered, out if which the periodic continuation medium unlimited distances in all three spatial directions is assumed. In this case, periodic boundary conditions are used [7].

If the calculations are made of real geometry microsystem, the one or more directions have a finite size. In this case as boundary conditions can be used or mirror boundary conditions (when particles interact with the specular reflection and don't leave thereof abroad), or a free exit of particles on one side of the allocated volume and an entrance of a particles flow – on another. For metal microsystems it is also characterized total absence of boundary conditions which are complied in a view of the forces of a mutual attraction of atoms in a metal lattice.

In this work both periodic, and mirror boundary conditions, and also an entrance of a particles flow in environment and a free exit of particles from it were used. In addition separate parts of Microsystems were thermostatted [20, 21].

3 REALIZATION

Realization of the developed multiscale approach relied on the numerical algorithm based on splitting into physical processes and using grid approximations of QGD equations (1)-(3) and subgrid calculations on the Newtonian dynamics equations (14)-(16).

The algorithm is as follows. In an initial time on the chosen grid the equilibrium condition of a macrosystem is given (10). If it is necessary near the borders of computational area equilibrium condition of gas and metal microsystems is also given.

Further on each time step predictor calculation of gas components macroparameters on grid analogs of QGD equations (1)-(3) without exchange members is made at first (QGD program block). As a result of calculation new gas components macroparameters in each control volume of a spatial grid are defined.

Then the subgrid MD computation in order to calculate the exchange members and kinetic coefficients is performed (MD program block). It is carried out independently in each control volume (cell) of a grid and produced with significantly smaller time step connected with evolution of a molecular subsystem. Criterion for stopping the MD calculation is or to achieve characteristic evolution time of molecular system, or change (for 1-2%) one or several macroparameters of molecular system (averages momentum, kinetic or potential energy). If strong changes of macroparameters don't happen, then the calculation is performed to achieve some predetermined points in time proportional to the Maxwellization time of MD system. Also in MD block the return to macroscopic level is carried out and correction of momentum densities and total energy densities is calculated by means of the computed exchange members. Correction is performed in each point of a grid. MD calculation process is finalized by calculations of kinetic coefficients and the equation of state. Thus, communication of macro- and microlevels in grid spatial and temporal areas is carried out by alternately recalculation of macroparameters on QGD and MD equations.

In case of nanoparticles presence in the flow the third block of calculations (NP program block) is used. In NP block it is decided or the convection-diffusion equations for concentration of nanoparticles of each type (this situation is considered for a case of small nanoparticles), or Newton's equations for the ballistic transport of nanoparticles (in the case of large nanoparticles). In this paper, that part of calculations is not considered.

The described part of algorithm is completed with check the criterion of the calculations end and transition to the following step on time in case of non-performance of criterion.

3.1 Finite volume scheme for QGD equations

The grid numerical method having a finite-volume method in the basis is used for calculation the macroparameters on QGD equations [22-24]. For this in the computational domain D the spatial grid Ω_D with cells C_m ($m=1,...,M$) and time grid ω_t with variable step Δt was introduced. Grid Ω_D in general is a hybrid, that is, it includes several element types: tetrahedrons, pentahedrons, hexahedrons, octahedrons and heptahedrons.

All parameters of gas components (density, pressure, temperatures, velocity vector components, etc.) have been carried to the centers of mass of grid elements, that is to the centers of cells. Stream variables have been set in the centers of cells sides. Spatial approximations of the main members have been executed by the technique presented in work [25]. The computing scheme on time was obvious. The resulting grid equations at a predictor stage have an appearance:

$$\hat{\rho}_{l,m} = \rho_{l,m} + \frac{\Delta t}{|V_m|}\sum_{p=1}^{M_m} W^{(\rho)}_{l,n,m,p}\left|S_{m,p}\right|, \quad W^{(\rho)}_{l,n,m,p} = \left(\mathbf{W}^{(\rho)}_{l,m,p}, \mathbf{n}_{m,p}\right),$$

$$\hat{\rho}_{l,m}\hat{u}_{l,k,m} = \rho_{l,m}u_{l,k,m} + \frac{\Delta t}{|V_m|}\sum_{p=1}^{M_m} W^{(\rho u)}_{l,n,m,p,k}\left|S_{m,p}\right|, \quad W^{(\rho u)}_{l,n,m,p,k} = \left(\mathbf{W}^{(\rho u_k)}_{l,m,p}, \mathbf{n}_{m,p}\right), \quad k = x,y,z, \quad (19)$$

$$\hat{E}_{l,m} = E_{l,m} + \frac{\Delta t}{|V_m|}\sum_{p=1}^{M_m} W^{(E)}_{l,n,m,p}\left|S_{m,p}\right|, \quad W^{(E)}_{l,n,m,p} = \left(\mathbf{W}^{(E)}_{l,m,p}, \mathbf{n}_{m,p}\right), \quad m = 1,...,M, \quad l = a,b.$$

Here f_m, $\hat{f}_m$ – values of the corresponding functions on the lower and upper layers in time in the center of the control volume V_m (coinciding into effect of the selected approximation with grid cell C_m), $|V_m|$ – value of the control volume V_m, $|S_{m,p}|$ – faces squares $S_{m,p}$ of the control volume V_m, M_m – faces number of the control volume V_m, $W^{(\alpha)}_{n,m,p}$ – scalar multiplication of flux $\mathbf{W}^{(\alpha)}_{m,p}$ via face $S_{m,p}$ on the external normal $\mathbf{n}_{m,p}$ to this face ($\alpha = \rho, \rho u_x, \rho u_y, \rho u_z, E$). Fluxes approximations $\mathbf{W}^{(\alpha)}_{m,p}$ are similar to the proposals in [25], but do not use artificial viscosity.

Obtained predictor variables $\hat{\rho}_l$, $\hat{\rho}_l \hat{\mathbf{u}}_l$, $\hat{E}_l$ are transmitted in MD block. Modified parameters $\hat{\rho}_l \hat{\mathbf{u}}^*_l$ and $\hat{E}^*_l$ are calculated in the result of its work. According to it the final values of momentum $\hat{\rho}_l \hat{\mathbf{u}}^{**}_l$ and energies $\hat{E}^{**}_l$ densities are calculated:

$$\begin{aligned} &\hat{\rho}_{l,m}\hat{u}^{**}_{l,k,m} = \hat{\rho}_l \hat{u}^*_{l,k,m} + \Delta t \nu_{ll'} \hat{\rho}_{l,m}\left(\hat{u}_{l,k,m} - \hat{u}^*_{l,k,m}\right), \quad k = x, y, z, \\ &\hat{E}^{**}_{l,m} = \hat{E}^*_{l,m} + \Delta t \nu_{ll'} \hat{\rho}_{l,m}\left(\hat{E}_{l,m} - \hat{E}^*_{l,m}\right), \quad m = 1, ..., M, \quad l = a, b, \quad l' = b, a. \end{aligned} \tag{20}$$

By the received in MD block macroparameters all the kinetic coefficients (μ_l, χ_l), the mean free path λ_l and the parameters in the equations of state Z_l, $c_{V,l}$ are calculated.

3.2 Algorithm for solving Newton's equations

The equation system (15) - (17) is solved by the Velocity Verlet algorithm [26]:

$$\begin{aligned} &0) \quad \mathbf{F}^0_{l,i} = \mathbf{F}\left(\mathbf{r}^0_{a,1}, ..., \mathbf{r}^0_{a,N_a}; \mathbf{r}^0_{b,1}, ..., \mathbf{r}^0_{b,N_b}; \mathbf{r}^0_{c,1}, ..., \mathbf{r}^0_{c,N_c}\right), \quad i = 1, ..., N_l, \quad l = a, b, c; \\ &\forall n = 0, 1, 2, ...: \\ &1) \quad \mathbf{r}^{n+1}_{l,i} = \mathbf{r}^n_{l,i} + \mathbf{v}^n_{l,i}\Delta t + \frac{\mathbf{F}^n_{l,i}}{m_l}\frac{(\Delta t)^2}{2}, \quad i = 1, ..., N_l, \quad l = a, b, c; \\ &2) \quad \mathbf{F}^{n+1}_{l,i} = \mathbf{F}\left(\mathbf{r}^{n+1}_{a,1}, ..., \mathbf{r}^{n+1}_{a,N_a}; \mathbf{r}^{n+1}_{b,1}, ..., \mathbf{r}^{n+1}_{b,N_b}; \mathbf{r}^{n+1}_{c,1}, ..., \mathbf{r}^{n+1}_{c,N_c}\right), \quad i = 1, ..., N_l, \quad l = a, b, c; \\ &3) \quad \mathbf{v}^{n+1}_{l,i} = \mathbf{v}^n_{l,i} + \frac{\mathbf{F}^{n+1}_{l,i} + \mathbf{F}^n_{l,i}}{2 m_l}\Delta t, \quad i = 1, ..., N_l, \quad l = a, b, c. \end{aligned} \tag{21}$$

Here Δt – integration step, n – step number, $\mathbf{F}^n_{l,i}$ – force value on step n, $\mathbf{F}$ – procedure for calculating the forces based on formulas (15), (16).

Berendsen thermostat [20] is used to achieve the desired temperatures of gas and metal. Langevin thermostat [21] is used to achieve the desired temperature and momentum of gas.

3.3 Parallel realization

Parallel realization of algorithm assumes using the cluster (or a supercomputer) with the central or hybrid architecture having on each node several multicore central processing units (CPU), and also several vector or graphic processing units (VPU or GPU). Parallelization of algorithm is made on the principles of geometrical parallelism and separating the areas. The main gasdynamic calculation is performed on discrete QGD equations on the grid distributed between cluster nodes by using technique "domain decomposition" [27]. Inside the QGD node

the computations are distributed between CPU threads. Subgrid MD computations are assigned to the CPU or VPU, or GPU, in its presence. Distribution of QGD calculations between CPU threads is also made geometrically. Parallelization of MD computations is made by partition of entire particles set relating to one cell of a grid on groups of identical power (explicitly these calculations were represented in [5]). Each unit of CPU or VPU, or GPU threads processes one or several molecular groups relating to one or several grid nodes. Program implementation of algorithm is executed on the hybrid technology [28] using MPI [29], OpenMP [30] and CUDA [31].

4 NUMERICAL EXPERIMENTS

For approbation of the developed approach the following preliminary testing was held. As computational domain the microchannel of rectangular section with sizes 15x15x90 μm^3 and with rectangular hole for nozzle of the diffuser with sizes 3x3 μm^2 was selected (see Fig. 2). The channel ended with an output in the free space, however in it the flow wasn't calculated any more. Pure nitrogen was considered as gas. Nickel was considered as the walls material.

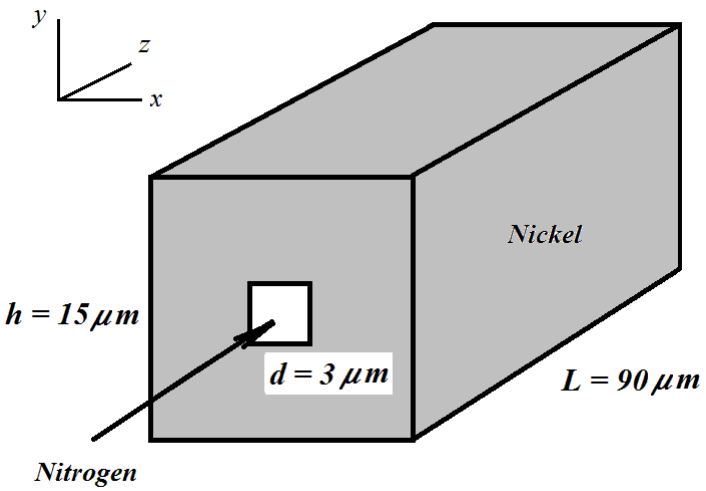


Figure 2: The model calculated area.

In a given computational domain uniform Cartesian grid, consisting of a cubic cells was constructed. The starting grid $\Omega^{(0)}$ had $M_V^{(0)} = 15 \times 15 \times 90 = 20250$ cells of linear size $a^{(0)} = 1\,\mu m$. Subsequent grid was obtained by grinding the cells of previous grid on 8 parts (by 2 times in each direction). As a result, grid $\Omega^{(k)}$ had $M_V^{(k)} = M_V^{(0)} 8^k$ cells. The volume of each cell in such grid is equal to $\left|V_m^{(k)}\right| = \left(a^{(k)}\right)^3\ \mu m^3$. The number of cells abutting to the side surface of calculated area excluding the input and output holes for grid $\Omega^{(k)}$ amounts to $M_S^{(k)} = 5400 \times 4^k + 216 \times 2^k$. Values $M_V^{(k)}$ and $M_S^{(k)}$ define the computing capacity of the developed algorithm. What they are bigger the more computing capacity of QGD block and less computing capacity of MD program block it is needed. We show this by example.

If the metal surface of the channel is described by a single repulsive potential without regard to its actual structure, we can restrict the following arguments. In each cell $C_m^{(0)}$ of starting grid $\Omega^{(0)}$ (having the volume $\left|V_m^{(0)}\right| = 1\,\mu m^3$) at normal conditions ($p = 101325\,Pa$,

$T = 273.15K$) there is a number of nitrogen molecules that is equal to $N_{part}^{(0)} = N_A \left| V_m^{(0)} \right| V_\mu^{-1} \approx 2.6867811 \times 10^7$ ($N_A \approx 6.022140857 \times 10^{23} \frac{1}{mol}$ – Avogadro's number, $V_\mu \approx 22.413962 \times 10^{15} \frac{\mu m^3}{mol}$ – mole volume gas at normal conditions). This amount is sufficient to obtain a representative statistics at MD calculations, however, too much for the actual calculations. Therefore, in the long-term calculations it is better to use a more detailed grid containing a larger number of cells, but with a smaller volume of single cell and comprising a smaller number of particles in it (see Table 1). The analysis shows that the optimal ratio of grid cells and particles in one cell of the grid is achieved for the grid $\Omega^{(3)}$. If the initial pressure in the channel is 0.001 atm, the starting number of gas molecules in the cell decreases to 52. This amount is insufficient for correct usage of local MD calculations within a single cell. Therefore, in areas of low pressure it is necessary to increase linear size of the cell the expense of neighboring cells. The particles number is calculated on the basis of densities values in the considered and the neighboring cells:

$$\bar{N}_{l,i,j,k} = \sum_{i',j',k'=-m,\ldots,+m} \frac{\rho_{l,i+i',j+j',k+k'}}{m_l} \left| V_{i+i',j+j',k+k'} \right|, \quad \left| \bar{V}_{i,j,k} \right| = \sum_{i',j',k'=-m,\ldots,+m} \left| V_{i+i',j+j',k+k'} \right|, \quad l = a,b, \quad (22)$$

where $m = 0,1,2,\ldots$ – the number of neighboring cells for each positive direction, which should be considered in the MD calculations.

k	$M_V^{(k)}$	$M_S^{(k)}$	$a^{(k)}$, μm	$\left\| V_m^{(k)} \right\|$, μm^3	$N_{part}^{(k)}$
0	20 250	5 616	1	1	26 867 811
1	162 000	22 032	0.5	0.125	3 358 476
2	1 296 000	87 264	0.25	0.015625	419 810
3	10 368 000	347 328	0.125	0.001953125	52 476
4	82 944 000	1 385 856	0.0625	0.000244140625	6 560

Table 1: Parameters of grids and amount of nitrogen particles in each cell.

If the metal surface is viewed at the atomic level, then in MD block computing it becomes necessary accounting nickel crystal lattice structure. However, in this paper, this case is not considered.

Testing of the developed approach was carried out on supercomputer MVS-10P (JSCC RAS). The calculations were performed on the central processors Intel Xeon E5-2690 and vector processors Intel Xeon Phi 7110X. The number of CPU cores ranged from 16 to 2048. The number of VPU cores ranged from 60 to 7680. Calculations were carried out on grids $\Omega^{(k)}$ for k=2,3,4. The results of a computation speedup testing are shown in Figure 3. It demonstrates a good parallel solution used in the constructed software.

The preliminary computation of steady state flow on grid $\Omega^{(3)}$ shows the following. A zone of gas condensation with a high pressure appears near the nozzle. It prevents a further acceleration of gas and acts as a stabilizer of a flow. A similar density distribution is given in the experimental paper [32]. A comparison of these data shows a good enough qualitative agreement of our calculations with the measurements (see Figure 4). On the other hand, in [32] the nozzle parameters are not specified and moreover, the measured temperature distribution

is given for two-temperature approximation that is not used in our predictions. Therefore, it is not possible to conduct a quantitative comparison of flow parameters for this experiment.

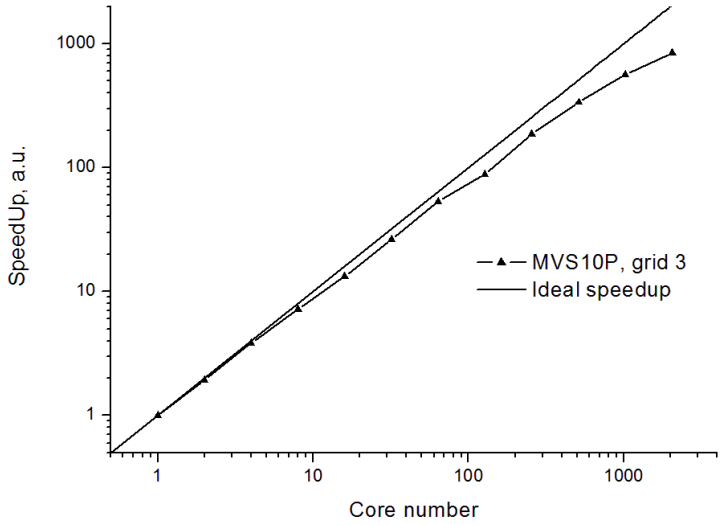


Figure 3: The speedup computations for grid $\Omega^{(3)}$.

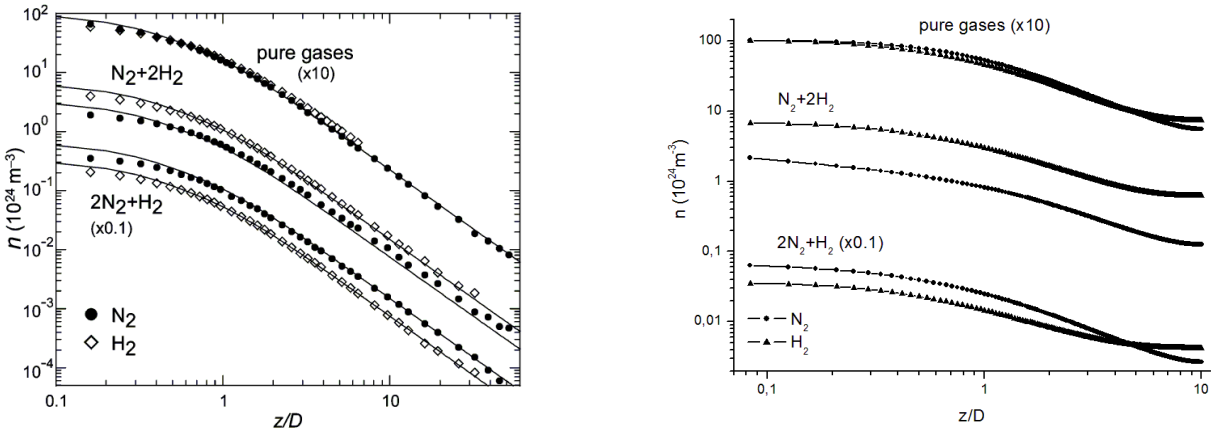


Figure 4: The distributions of the number density in experiments [32] (on the left side) and the data from our computations (on the right side).

5 CONCLUSIONS

- The problem of modeling the gas flows in microchannels of technical systems in the conditions of many scales of computational domain is considered. For a solution the two-scale approach combining the solution of QGD equations and correction of gasdynamic parameters by MD method is offered.
- The general numerical algorithm is based on solving the QGD equations system by a grid finite-volume method. The system of MD equations is used independently in each cell of a grid and is solved by means of Verlet scheme.
- The parallel implementation of approach is based on the methods of splitting into physical processes and domain decomposition. The parallel program is focused on the use of modern computing systems with the central and hybrid architectures. The program is realized by means of MPI, OpenMP and CUDA technologies.

- The research of proposed approach showed the following. The developed parallel algorithm is quite effective and steady to using the data of MD computations correcting the flow. By means of the developed program it is possible to calculate dependences of kinetic coefficients of QGD system on temperature and other factors. Transition from macrolevel to microlevel and back is carried out correctly.
- Comparing the calculated macroparameters of a three-dimensional flow agrees well with the results of natural experiments.

6 AKNOWLEDGEMENTS

This work was supported by Russian Foundation for Basic Researches (projects No. 15-07-06082-a, 16-07-00206-a, 15-29-07090-ofi_m).

REFERENCES

[1] A. Papyrin, V. Kosarev, S. Klinkov, A. Alkhimov, V. Fomin, *Cold Spray Technology*. Amsterdam, Elsevier Science, 2007.

[2] V.O. Podryga, Molecular Dynamics Method for Simulation of Thermodynamic Equilibrium. *Mathematical Models and Computer Simulations*, **3**, 381–388, 2011.

[3] V.O. Podryga, S.V. Polyakov, Molecular Dynamics Simulation of Thermodynamic Equilibrium Establishment in Nickel. *Mathematical Models and Computer Simulations*, **5**, 456–466, 2015.

[4] V.O. Podryga, S.V. Polyakov, Correction of the Gas Flow Parameters by Molecular Dynamics. E. Oñate, M. Bischoff, D.R.J. Owen, P. Wriggers, T. Zohdi eds. *4th International Conference on Particle-Based Methods – Fundamentals and Applications (PARTICLES 2015)*, Barcelona, Spain, September 28-30, 2015.

[5] V.O. Podryga, S.V. Polyakov, D.V. Puzyrkov, Supercomputer molecular modeling of thermodynamic equilibrium in gas-metal Microsystems. *Vychislitel'nye metody i programmirovanie*, **1**, 123–138, 2015.

[6] T.A. Kudryashova, V.O. Podryga, S.V. Polyakov, Simulation of gas mixture flows in microchannels. *Vestnik RUDN. Seriya: Matematika. Informatika. Fizika*, **3**, 154–163, 2014.

[7] D. Frenkel, B. Smit, *Understanding Molecular Simulation. From Algorithm to Applications*. New-York, Academic Press, 2002.

[8] T.G. Elizarova, *Quasi-gas dynamic equations*. Springer, 2009.

[9] T.G. Elizarova, A.A. Zlotnik, B.N. Chetverushkin, On quasi-gasdynamic and quasi-hydrodynamic equations for binary mixtures of gases. *Doklady Mathematics*, **3**, 1–5, 2014.

[10] G.A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Oxford, Clarendon Press, 1994.

[11] J.W. Buddenberg, C.R. Wilke, Calculation of Gas Mixture Viscosities. *Ind. Eng. Chem.*, **7**, 1345–1347, 1949.

[12] G.N. Abramovich, *Applied gas dynamics. Part 1*. Moscow, Publishing of Science, 1991.

[13] G. Von Mie, Zur kinetischen theorie der einatomigen korper. *Ann. Phys., Leipzig*, **8**, 657–697, 1903.

[14] L.R. Fokin, A.N. Kalashnikov, The transport properties of an N2-H2 mixture of rarefied gases in the EPIDIF database. *High Temperature*, **5**, 643–655, 2009.

[15] X.W. Zhou, R.A. Johnson, H.N.G. Wadley, Misfit-energy-increasing dislocations in vapor-deposited CoFe/NiFe multilayers. *Phys. Rev. B*, **69**, 144113, 2004.

[16] J.E. Lennard-Jones, Cohesion. *Proceedings of the Physical Society*, **5**, 461–482, 1931.

[17] P.M. Morse, Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys. Rev.*, **34**, 57–64, 1929.

[18] S. Maruyama, Molecular dynamics method for microscale heat transfer. *Advances in Numerical Heat Transfer*, **2**, 189–226, 2000.

[19] D.V. Sivukhin, *The general course of physics. Vol. II. Thermodynamics and molecular physics*. Moscow, Publishing of Science, 1975.

[20] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren et al., Molecular dynamics with coupling to an external bath. *J. Chem. Phys*., **81**, 3684–3690, 1984.

[21] D.W. Heerman, *Computer simulations methods in theoretical physics*. Springer, 1986.

[22] I.V. Fryazinov, Balance method and variational-difference schemes. *Differentsial'nye uravneniya*, **7**, 1332–1343, 1980.

[23] A.A. Samarskiy, A.V. Koldoba, Yu.A. Poveshchenko, V.F. Tishkin, A.P. Favorskiy, *Difference schemes on irregular grids*. Minsk, Publishing of Closed Joint Stock Company “Criterion”, 1996.

[24] R. Eymard, T.R. Gallouet, R. Herbin, The finite volume method. *Handbook of Numerical Analysis, Amsterdam, North Holland,* **7**, 713–1020, 2000.

[25] I.V. Popov, I.V. Fryazinov, Method of adaptive artificial viscosity for the equations of gas dynamics on triangular and tetrahedral grids. *Mathematical Models and Computer Simulations*, **1**, 50–62, 2013.

[26] L. Verlet, Computer «experiments» on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev*., **159**, 98–103, 1967.

[27] A. Toselli, O. Widlund, Domain *Decomposition Methods - Algorithms and Theory*. Springer Series in Computational Mathematics, Springer, Berlin, Vol. 34, 2005.

[28] S.V. Polyakov, Yu.N. Karamzin, O.A. Kosolapov, T.A. Kudryashova, S.A. Sukov, Hybrid supercomputer platform and applications programming for the solution of continuous mechanics problems by grid methods. *Izvestiya YuFU. Tekhnicheskie nauki*, **6**, 105–115, 2012.

[29] Official documentation and manuals on MPI. URL: http://www.mcs.anl.gov/research/projects/mpi/

[30] Official documentation and manuals on OpenMP. URL: http://openmp.org/wp/

[31] CUDA Toolkit Documentation v. 7.5. URL: http://docs.nvidia.com/cuda/#

[32] A. Ramos, G. Tejeda, J.M. Fernandez, S. Montero, Nonequilibrium Processes in Supersonic Jets of N2, H2, and N2 + H2 Mixtures: (I) Zone of Silence. *J. Phys. Chem. A*, **113**, 8506–8512, 2009.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

HIGHER-ORDER ASYMPTOTIC HOMOGENIZATION OF PERIODIC MATERIALS WITH LOW SCALE SEPARATION

Maqsood Ameen, Ron Peerlings and Marc Geers

Eindhoven University of Technology
Eindhoven, The Netherlands
e-mail: m.ameen, r.h.j.peerlings, m.g.d.geers@tue.nl

Keywords: Linear Elastic Composites, Homogenization methods, Higher-order Periodic Homogenization, Size Effects.

Abstract. *In this work, we investigate the limits of classical homogenization theories pertaining to homogenization of periodic linear elastic composite materials at low scale separations and demonstrate the effectiveness of higher-order periodic homogenization in alleviating this limitation. Classical homogenization techniques are known to be very effective for materials with large scale separation between the scale of the heterogeneity and the macro-scale dimension, but inaccurate at low scale separations. Literature suggests that asymptotic homogenization is capable of pushing the limit to smaller scale separation by taking on board higher-order terms of the asymptotic expansion. We studied infinite two-dimensional elastic two-phase composite materials consisting of stiff inclusions in a soft matrix, subjected to a periodic body force, for various scale ratios between the period of the body force and that of the inclusions. We created reference solution using direct numerical simulation and used ensemble averaging for the complete family of all possible microstructures to obtain the reference homogenized solution. We show that the response predicted using zeroth order classical homogenization deviates from this reference homogenized solution for scale ratios below 10. The higher-order asymptotic homogenization solution still gives a very good approximation even in the low scale separation regime and it becomes better as more higher-order terms are included. The higher-order theory results in a size-dependent macroscopic model, which indeed allows one to push the limitations of homogenization in the direction of less scale separation.*

1 INTRODUCTION

All matter is heterogeneous at some scale, but frequently it is convenient to treat it as homogeneous. Some of the well-known examples are metal alloys, concrete, porous structures and fibrous composites. The distinct features of their microstructures respond quite differently to mechanical loading and hence their deformation is heterogeneously distributed at the scale of that microstructure. It is the combination of the different microstructural features which governs the overall response of the material to the loading.

Homogenization is a mathematical technique for studying partial differential equations with rapidly oscillating coefficients, which are typical of the equations that govern the physics of heterogeneous materials. An important aspect in the analysis of multiphase materials is to deduce their effective behavior (e.g. mechanical stiffness, thermal expansion properties, etc.) from the corresponding single-phase behaviors and the geometrical arrangement of the phases. This concept of rendering "homogeneous" a heterogeneous material is what we call homogenization.

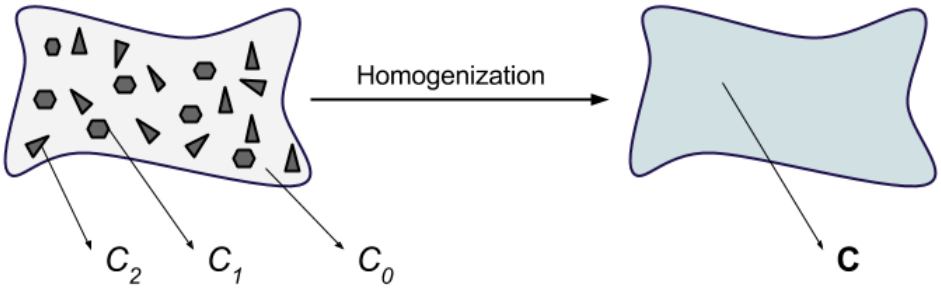


Figure 1: Homogenization: schematic

Conventional homogenization methods are based on *a separation of scales*, given by: $l << L$, where l is the size of the heterogeneity and L represents the macroscopic length scale. However, if the microstructural size is not much smaller, or even of the same order as the macroscopic length scale, then most of the classical homogenization schemes break down. Literature [1-4] suggests that the asymptotic homogenization method is capable of pushing the limit to smaller scale separation, by generating a hierarchy of problems which can be solved sequentially to generate a solution that asymptotically converges to the exact (homogenized) solution.

The present work investigates the scale separation limits of classical homogenization theories and demonstrates the effectiveness of higher order periodic homogenization for small scale separation. A qualitative and quantitative assessment of the scale separation limits of the classical and the higher order periodic homogenization methods is performed on two-dimensional elastic two-phase composites. The details of the problem are described in Section 2. We study the limits of classical homogenization and look into the contribution of each higher order term in rectifying this limitation.

2 PROBLEM DESCRIPTION

An infinite two-dimensional elastic two-phase composite material consisting of stiff circular inclusions in a soft matrix material, is subjected to anti-plane shear by means of a periodic body force. This anti-plane shear problem [1] can be described by the following partial differential

equation:

$$\frac{\partial}{\partial x_1}\left(G\frac{\partial u_3}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(G\frac{\partial u_3}{\partial x_2}\right)+F=0 \tag{1}$$

where $u_3 = u(x_1, x_2)$ is the resulting out-of-plane displacement, $G(x_1, x_2)$ is the Shear Modulus distribution function, defined below, and $F(x_1, x_2)$ is a body force which we define here as $F = F_0 \sin\left(\frac{2\pi x_1}{L}\right) \sin\left(\frac{2\pi x_2}{L}\right)$.

The period of the material's microstructure is *l* and that of the body force is *L*, and the ratio *L/l* hence characterises the scale separation. The homogenized properties are defined not for a specific microstructural configuration with respect to a period of the body force, but for an ensemble averaged microstructural configuration as described in Section 3.

The microstructure consists of a matrix material having Shear Modulus G_m and a circular inclusion having radius r and Shear Modulus G_p. Rather than considering a sharp interface, we define the Shear Modulus function $G(x)$ such that it smoothly varies from G_m to G_p by a cubic distribution. For $a \leq x \leq b$, it is given by:

$$G(x) = \frac{(3a-b-2x)(b-x)^2 G_m + (a-3b+2x)(a-x)^2 G_p}{(a-b)^3} \tag{2}$$

where a and b represents the interface boundaries such that the thickness of the interface is given by $(b-a)$. Fig.2 shows the Shear Modulus distribution for a microstructure having the following parameters: $G_m = 1$, $G_p = 20$, $r = 0.3$, $a = 0.25$ and $b = 0.35$. The amplitude of the body force, F_0, is taken as 1.

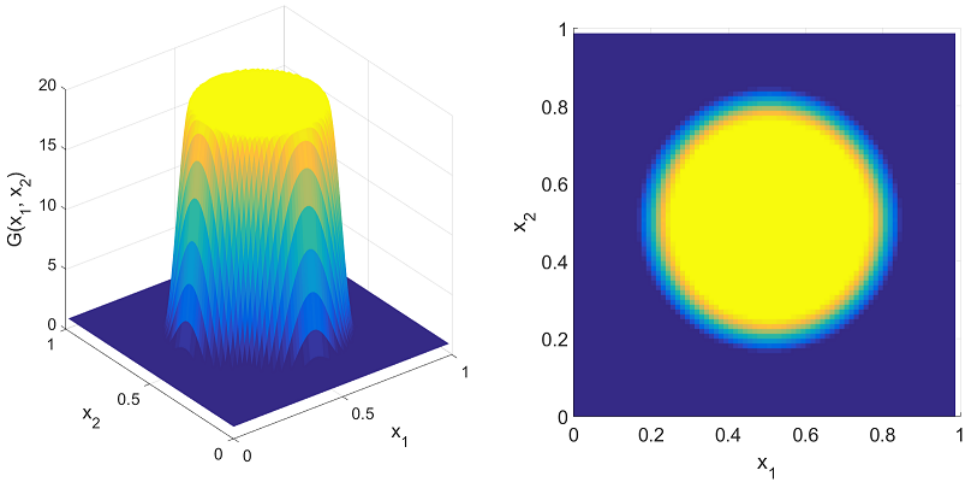


Figure 2: Shear Modulus distribution within a microstructure of size $l = 1$

3 METHODOLOGY

On the one hand, reference solutions are created using direct numerical simulation of a family of microstructural configurations for a range of scale ratios. On the other hand, asymptotic homogenization is used to obtain homogenized properties for zeroth order and higher-order effective continua. Predictions made using these homogenized properties are then compared against the reference solutions. Fig.3 shows a schematic of the methodology.

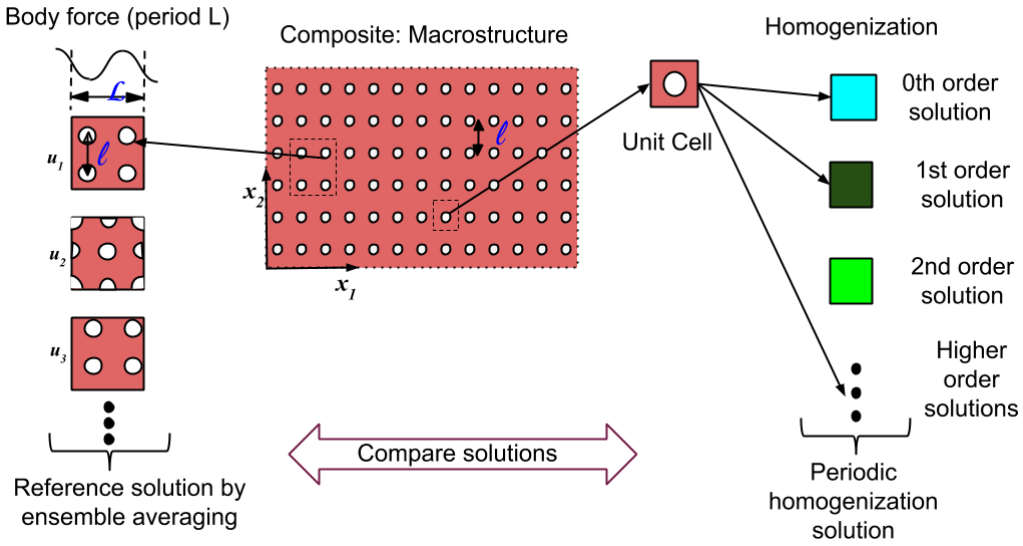


Figure 3: Schematic of the problem: Comparison of periodic homogenization solution with reference solution created by full-scale simulation for various scale ratios

Reference solutions are created using a brute force method, by performing a full-scale numerical simulation. Since the exact location of the microstructure with respect to that the body force is unknown, each relative position is assumed to have the same probability of occurrence. The homogenized response is defined not for a specific microstructural configuration with respect to a period of the body force, but by taking an ensemble average for the complete family of all possible microstructures i.e., by averaging the displacement solutions computed for all shifts. The result is plotted in Fig.4 in terms of the peak averaged displacement (infinity norm) versus the ratio *L/l* for circular inclusions. For scale ratios larger than 10, the curves would be horizontal as the particle sizes become very small relative to the period of the body force and hence do not cause any significant influence on the average displacement solution. As the scale ratio becomes smaller than 10, the relative size of particles becomes high enough to cause size effects, which is apparent in the plot. The curvature of these curves is dependent on the geometry of the microstructure and stiffness contrast.

Periodic homogenization is a rigorous method used to extract effective or homogenized properties from heterogeneous media. For our problem, we first analyze a representative unit cell and derive the effective Shear Moduli. This is then inserted into the macroscopic model to compute the displacement field of the homogenized material subjected to the antiplane shear load defined above.

The microscopic scale described by $\vec{y}$ is determined by the microstructure with a characteristic length l. The macroscopic length $\vec{x}$ described by the wavelength of the applied loading on the material (or the boundary conditions), has a characteristic length L. The small parameter η is defined as the ratio of the two length scales. Hence we have $\eta = l/L$ and $\vec{y} = \eta^{-1}\vec{x}$.

The equilibrium equation can be written in terms of $\vec{x}$ and $\vec{y}$:

$$\vec{\nabla} \cdot (\mathbf{G}(\vec{y}) : \vec{\nabla} u) + \vec{f}(\vec{x}) = 0 \tag{3}$$

Periodic homogenization makes use of an asymptotic expansion of the unknown variable in terms of the powers of the small parameter η:

$$u(\vec{x}) = u_0(\vec{x}, \vec{x}/\eta) + \eta u_1(\vec{x}, \vec{x}/\eta) + \eta^2 u_2(\vec{x}, \vec{x}/\eta) + \mathcal{O}(\eta^3) \tag{4}$$

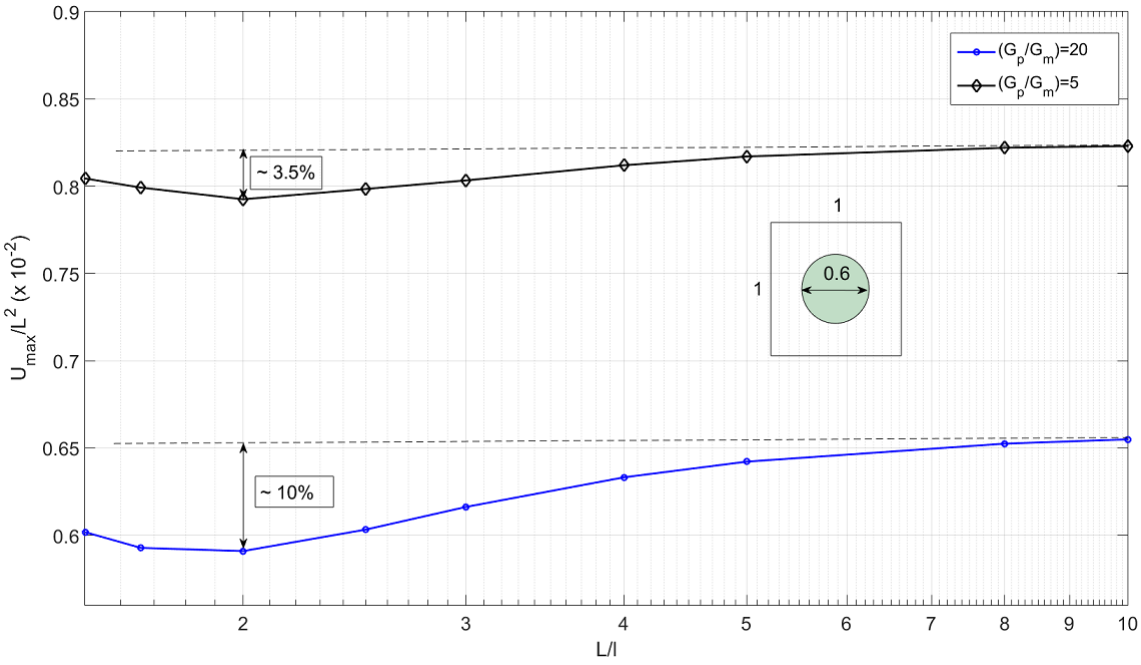


Figure 4: Peak average displacement versus scale ratio for circular inclusions with different stiffness contrast

where the functions $u_i(\vec{x}, \vec{x}/\eta)$ have to be determined and must be periodic in $\vec{y}$. This periodic dependence introduces the fast displacement fluctuations (*microfluctuations*) at the microscale η, while the dependence on $\vec{x}$ is slow. For small η, the equation can be regarded as consisting of a leading term $u_0(\vec{x}, \vec{x}/\eta)$, followed by a series of rapidly diminishing correction terms. As η increases, the contributions from the higher order terms increase and are not negligible anymore.

The essence of this asymptotic method lies in requiring the new equilibrium equation ((3) after replacing u with (4)) to be satisfied at each order of η separately and for independent $\vec{x}$ and $\vec{y}$. Thus, we generate a hierarchy of problems which needs to be solved sequentially to compute the unknown functions u_i. One can show that the result must be of the form [2]:

$$u(\vec{x}) = v_0(\vec{x}) + \eta(v_1(\vec{x}) + N_1(\vec{y}) : \vec{\nabla} v_0) + \eta^2(v_2(\vec{x}) + N_1(\vec{y}) : \vec{\nabla} v_1 + N_2(\vec{y}) \vdots (\vec{\nabla}\vec{\nabla} v_0)^T) + \mathcal{O}(\eta^3) \tag{5}$$

where $N_i(\vec{y})$ is called the microfluctuation function of $(i-1)^{th}$ order, which can be obtained by solving a unit cell problem for each order of η. Here we solve this problem numerically, using a finite difference discretization Fig.5 shows some of the microfluctuation functions computed. Note that the order of magnitude of these fluctuation functions decreases for higher order solutions.

Once the microfluctuation functions are computed, the effective constants of the corresponding order can be calculated. These homogenized constants are then passed on to the macroscale in order to compute the variables v_i.

4 RESULTS

The results obtained using periodic homogenization solution are now compared with the reference solution obtained using direct numerical simulation. Fig.6 shows the average peak displacement, normalized by that predicted by the classically homogenized solution, as a function of the scale ratio $\eta^{-1} = L/l$. Shown are the reference solution, the classical homogenization

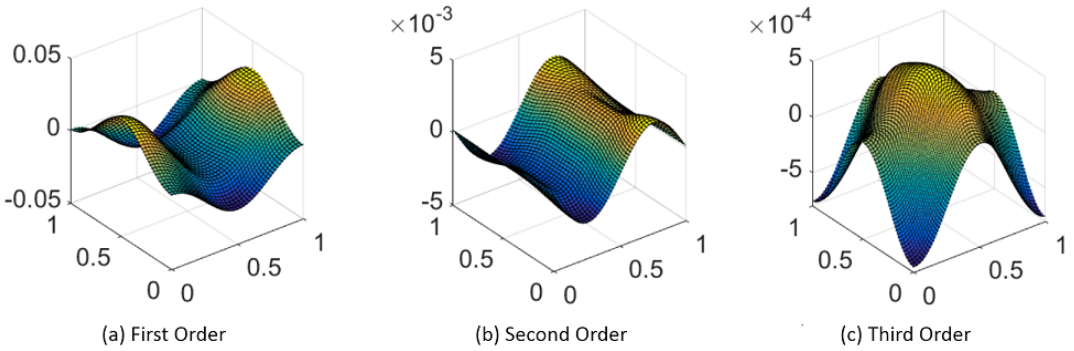


Figure 5: Microfluctuation functions of different orders

solution and higher order solutions, for a phase contrast of 20. The zeroth order classical homogenization solution is independent of the scale ratio and hence is a straight line as shown in the plot. For scale ratios $L/l > 10$, the reference solution converges to this constant value, but at low scale separations it deviates from it significantly. The higher order periodic homogenization solutions closely match with the reference solution even for low scale ratios. The second order solution starts to deviate from the reference solution at $L/l \approx 4.5$, while the fourth order solution can still give a good approximation for even lower scale ratios.

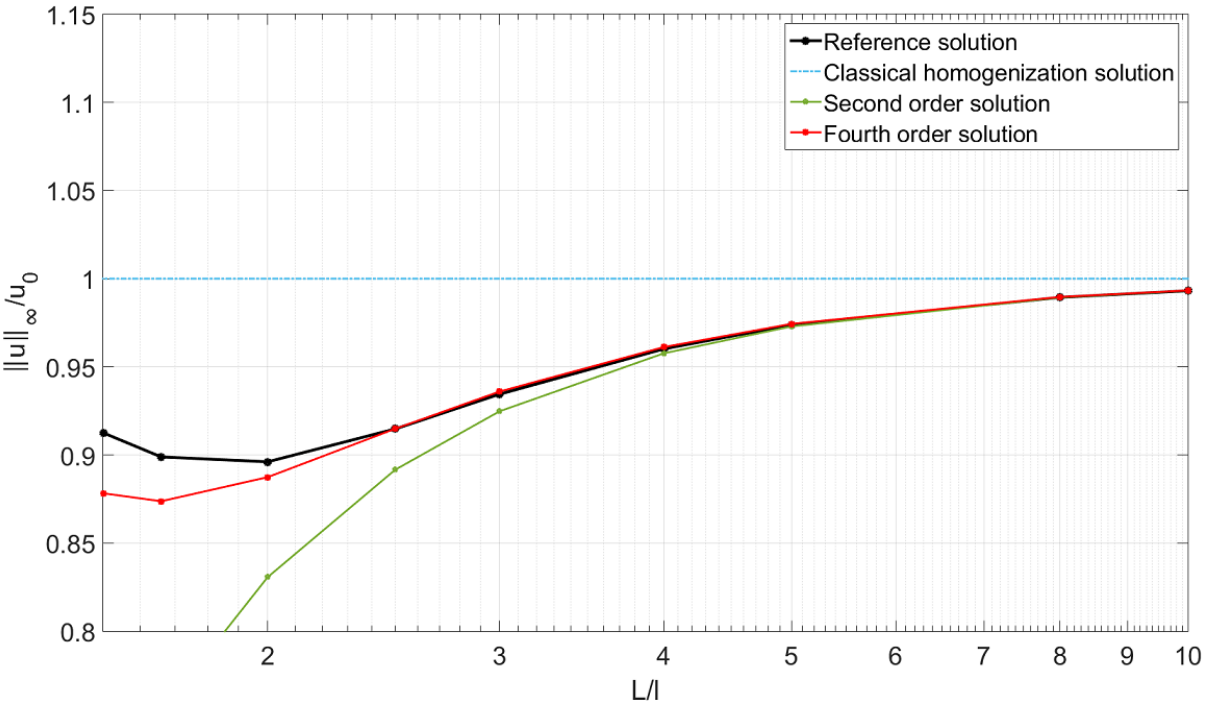


Figure 6: Scale separation plot: Norm of the displacement solution vs. scale separation, for various cases

5 CONCLUSIONS

The result gives a very good insight into the low scale separation regime of an elastic periodic two-phase composite. Some of the observations from this work are as follows:

- The classical homogenization solution is accurate only for cases where $l << L$. However,

quantitatively it can still predict homogenized solution accurately for scale separations above $L/l = 10$.

- Higher order periodic homogenization is an accurate method for tackling problems in linear elasticity for the cases with low scale separation. The approximation in the low scale separation regime becomes better as more higher-order terms are included.

- We can conclude that higher order periodic homogenization is not constrained by the conventional separation of scales. It rather has a more flexible law $l < L$, where if the material property and geometry are specified, a clear criteria stating the order required for a certain accuracy can be made.

REFERENCES

[1] Smyshlyaev V. P., Cherednichenko K. D.: Strain gradient effects in periodic heterogeneous media. JMPS 48:1325-1357, 2000.

[2] Boutin C.: Microstructural effects in elastic composites. IJSS 33: 1023-1051, 1996.

[3] Drugan W., Willis J.: A micromechanics-based estimates of RVE size for elastic composites. JMPS 44:497-524, 1996.

[4] Peerlings R. H. J., Fleck .N. A.: Computational evaluation of strain gradient elasticity constants. IJMCE 2, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MESO-SCALE MODELING OF HYBRID INDUSTRIAL/RECYCLED STEEL FIBER-REINFORCED CONCRETE

Antonio Caggiano[1], Diego Said Schicchi[2], Guillermo Etse[1] and Enzo Martinelli[3]

[1] CONICET and Universities of Buenos Aires and Tucuman
e-mail: acaggiano@fi.uba.ar, getse@herrera.unt.edu.ar

[2] Stiftung Institut für Werkstofftechnik (IWT), Bremen, Germany
and Instituto Nacional de Tecnología Industrial, Buenos Aires, Argentina
schicchi@iwt.uni-bremen.de

[3] University of Salerno
Fisciano, Salerno, Italy
e.martinelli@unisa.it

Keywords: HyFRC, Meso-scale, Discrete Crack Approach, Zero-thickness Interface, Embedded Short Beams, Pull-out.

Abstract. *This paper investigates the mechanical behavior of fiber-reinforced concrete (FRC) and focusses on the quantifying the effect of replacing Industrial Steel Fibers (ISFs), commonly adopted as spread reinforcement in FRC, with Recycled Steel Fibers (RSFs) recovered from waste tires. More specifically, it analyses the bending behavior of FRC beams reinforced with a constant volume fraction of steel fibers and variable proportions of ISFs and RSFs. First, a numerical model is formulated by assuming that FRC behaves as a multi-phase medium, where the nonlinear material behavior of the concrete matrix is simulated by following a discrete-crack approach for meso-scale analysis. Then, steel fibers are modeled as short cables, randomly distributed and embedded within the concrete matrix. The internal forces in the steel fibers are obtained by considering both bond-slip behavior and dowel effect. Comparisons between experimental results, obtained by the authors in a previous study, and numerical simulations, performed by means of the proposed numerical model, are discussed: the significant predictive capability of the latter confirms the soundness of the mechanical assumptions on which the model is based. Moreover, the possibility of predicting the behavior of FRC with Hybrid Recycled/Industrial Fibers paves the way toward the actual application of this sustainable material in real applications. Finally, it is worth highlighting that the theoretical formulation proposed in this work stems out of the activities foreseen by the SUPERCONCRETE Project (H2020-MSCA-RISE-2014 – n. 645704), funded by the European Union as part of the H2020 Programme.*

1 INTRODUCTION

In recent years the disposal of exhaust tires has emerged as a big issue in waste management and the increasing amount of this waste actually represents a serious threat for both environment preservation and human health [1]. Based on the ‘‘Council Directive 1999/31/EC’’ of the European Commission on the Landfill of Waste, as of 2003 post-consumer ‘‘whole tires’’ could no longer be landfilled and, since July 2006, such regulations must be applied to both ‘‘whole’’ and ‘‘shredded’’ tires [2]. Therefore, there are strong motivations for recycling waste tires, which can easily be turned into an eco-friendly source of secondary raw materials.

Recycling processes of waste tires mainly consist of separating the internal steel reinforcement from the rubber covering. Rubber scraps and short steel fibers are obtained via these processes and, among other alternative solutions, they can be employed in partial-to-total replacement of ordinary concrete constituents. Particularly, rubber scraps find an interesting field of application as a partial replacement of ordinary stone aggregates for obtaining the so-called rubberized concretes [3][4]. Furthermore, Recycled Steel Fibers (RSFs) can replace Industrial Steel Fibers (RSFs) for producing a cementitious composite generally referred to as Recycled Steel Fiber Reinforced Concretes (RSFRCs) [5][6].

Various experimental studies available in the scientific literature deal with the mechanical characterization of FRC in post-cracking range. Several recently issued works have investigated the mechanical performance and the post-cracking response of the most common composite materials, i.e., Steel-FRC [7], Polypropylene FRC [8] or Hybrid-FRC [9]. Furthermore, some experimental campaigns on “ecofriendly fiber reinforced concrete composites”, such as those made with either recycled steel fibers obtained from waste tires (RSFRC) [10] or natural fibers (NFRC) [11] are also available in literature.

Several theoretical models, intended to simulate the failure behavior and post-cracking response of FRC at both material and structural levels, are also available in the literature. They range from empirical design relationships [12][13] to more complex meso-mechanical models [14][15]. The latter take into account explicitly the interaction among the different phases of the composite (i.e., fibers, matrix, coarse aggregates and their interfaces) [16] and, hence, they require a sound knowledge of fiber-matrix interaction [17][18].

This work proposes a meso-scale model aimed at simulating the failure behavior and post-cracking response of Hybrid Industrial/Recycled Steel Fiber-Reinforced Concretes (HyIRSFRCs). More specifically, a zero-thickness interface model for plain concrete is employed in the framework of a meso-scale discrete-crack approach. This discontinuous model assumes the fracture-based model originally proposed by Carol et al. [19]. Moreover, a novel and promising approach to account for the fiber effect in concrete composites and mortar has been proposed in this paper. Particularly, this new advanced extension deals with the assumption that Industrial Steel Fibers (ISFs) and Recycled Steel Fibers (RSFs) can be considered as embedded short beams randomly distributed within the concrete matrix.

The paper is organized as follows. Section 2 reports the main assumptions of the FRC modeling and approaches. After this, Section 3 describes the fracture energy-based plasticity formulation for plain mortar/concrete interfaces while Section 4 highlights the fiber bond-slip formulation and dowel mechanisms for the embedded beams. Section 5 shows a relevant application of the proposed model aimed at simulating the post-cracking response of the HyIRSFRC. Finally, some concluding remarks are given in Section 6.

2 OUTLINE OF THE MODELING APPROACH

The present model for FRC is inspired to a discrete-crack approach and considers fibers as beam elements "embedded" within the concrete matrix. Hence, it includes three internal formulations as follows:

i. *A fracture energy-based plasticity formulation for plain mortar/concrete joints*: the constitutive model for iso-parametric interface elements relates normal and tangential stress components with the corresponding relative displacements; the three-parameter hyperbolic failure surface by Carol et al. [19] is assumed as maximum strength criterion, while the ratio between fracture work and energy controls the post-cracking response: Section 3 reports further details on this model;

ii. *Fiber bond-slip* developed in the fiber (beam) direction. Pull-out mechanisms of steel fibers crossing cracks (these latter represented through opened joints) is formulated by means of an elastoplastic model as indicated in Section 4;

iii. *Fiber dowel action* based on elastic foundation concepts to obtain the dowel force-displacement relationship developed in the transversal direction of the considered short beam which crosses an active fracture: this model is detailed in Section 5.

3 FRACTURE ENERGY-BASED INTERFACE PLASTICITY FORMULATION

This section presents a rate-independent fracture-based model for concrete interfaces: Table 1 summarizes its main features, whereas further details are available in the literature [20].

	Fracture - based energy interface model
Constitutive relationships	$\dot{\mathbf{u}} = \dot{\mathbf{u}}^{el} + \dot{\mathbf{u}}^{cr}$ $\dot{\mathbf{u}}^{el} = \mathbf{C}^{-1} \cdot \dot{\mathbf{t}}$ $\dot{\mathbf{t}} = \mathbf{C} \cdot \left(\dot{\mathbf{u}} - \dot{\mathbf{u}}^{cr} \right)$
Yield condition	$f\left(\mathbf{t}, \kappa\right) = \sigma_T^2 - (c - \sigma_N \tan\phi)^2 + (c - \chi \tan\phi)^2$
Flow rule	$\dot{\mathbf{u}}^{cr} = \dot{\lambda}\mathbf{m}$ $\mathbf{m} = \mathbf{A} \cdot \mathbf{n}$
Cracking work-evolution	$\dot{\kappa} = \dot{w}_{cr}$ $\dot{w}_{cr} = \sigma_N \cdot \dot{u}^{cr} + \sigma_T \cdot \dot{v}^{cr} \qquad \sigma_N \geq 0$ $\dot{w}_{cr} = \left[\sigma_T - \lvert\sigma_N\rvert tan(\phi) \right] \cdot \dot{v}^{cr} \qquad \sigma_N < 0$
Evolution law	$p_i = \left(1 - \left(1 - r_{pi}\right) S[\xi_{p_i}]\right) p_{0i}$
Kuhn - Tucker loading/unloading and consistency conditions	$\dot{\lambda} \geq 0, \quad f \leq 0, \quad \dot{\lambda} f = 0$ **Kuhn-Tucker** $\dot{f} = 0$ **Consistency**

Table 1: Overview of the interface model for plain concrete and mortars.

In Table 1, $\mathbf{C}$ defines the uncoupled normal/tangential elastic stiffness matrix, $\dot{\mathbf{u}}^{el}$ and $\dot{\mathbf{u}}^{cr}$ the vectors of the elastic and cracking displacement rate (according to the non-associated flow rule), respectively, $f\left(\mathbf{t}, \kappa\right)$ is the hyperbola defining the yield condition of the model on the bases of the three-parameters (Figure 1) χ (the tensile strength), c (the cohesion) and ϕ (the friction angle) [19].

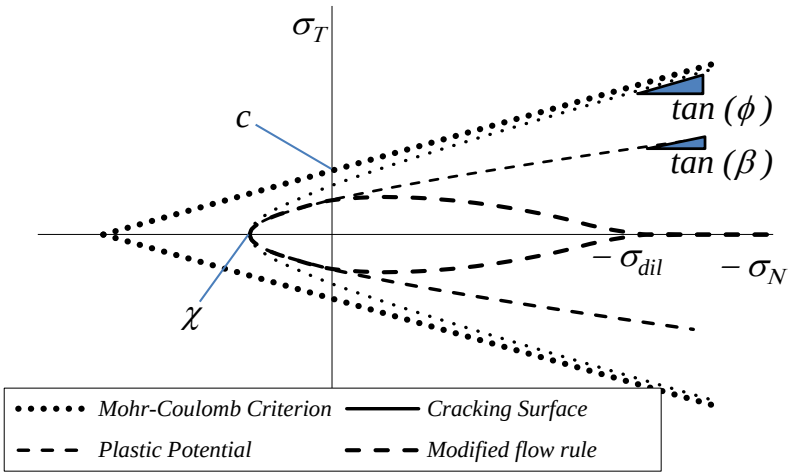


Figure 1: Failure hyperbola by Carol et al. [19], Mohr–Coulomb surface, plastic potential and modified flow rule.

The cracking displacement rate, $\dot{\mathbf{u}}^{cr} = \left[\dot{u}^{cr}, \dot{v}^{cr}\right]^t$ (where $\dot{u}^{cr}$ and $\dot{v}^{cr}$ represent the normal and the tangential components, respectively, while t interprets the transposition vector operator), is given through a general non-associated flow rule which controls the direction, **m**, of the interface fracture displacements by means of the transformation matrix operator **A** applied to the associated normal flow derivate, **n**; $\dot{\lambda}$ is the non-negative plastic multiplier derived by means of the Kuhn-Tucker loading/unloading and consistency conditions.

Furthermore, Table 1 highlights the incremental fracture work, $\dot{w}_{cr}$, which controls the evolution of the yielding surface in a generic fracture (post-elastic) process. A unified decay function is considered for each internal parameter, p_i (alternatively equals to χ, c or $\tan\phi$), of the yield condition: p_{0i} represents the initial value for p_i, $r_p p_{0i}$ the residual one and $S[\xi_{p_i}]$ the scaling function where ξ_{p_i} measures the ratio between the current work spent and the available fracture energies in mode I and II.

4 EMBEDDED SHORT BEAMS: PULL-OUT AND DOWEL EFFECTS

FRC is modeled as a meso-scale medium composed by one homogeneous matrix (aggregates and paste) plus another "phase" represented by the steel fibers. Hence, the fracture process is modeled through interface elements, in the framework of a discrete crack approach, while the stress transferred between cracks, due to fibers bridging effect, is modelled with embedded cable (beam) elements. The spatial position and orientation of fibers are generated randomly in a two-dimensional finite element mesh. The number of fibers is calculated as a function of volume fraction and geometric characteristics of the analyzed reinforcement. Then, the contributions of the steel fibers are considered through two sub-models. Two plasticity-based models are employed for modelling the fibers bond-slip behavior and dowel mechanism, respectively, in the axial and transversal direction of the fiber.

4.1 Bond-slip

This section deals with the proposed one-dimensional plasticity model for the stress-strain response. The model deals with a resulting bilinear stress-strain (σ_f–ε_N) rule which models the fiber bond–slip response (Table 2). It is based on the additive decomposition of the total strain rate $\dot{\varepsilon}_N$ into elastic $\dot{\varepsilon}_N^{el}$ and plastic $\dot{\varepsilon}_N^{pl}$ components. $\dot{\sigma}_f$ is the total stress rate while E_f repre-

sents the uniaxial elastic modulus which considers both the uniaxial response of the steel fiber and the bond-slip effect of the short steel reinforcement in mortar/concrete substrate; f_f is the yield condition, being $\sigma_{y,f}$ the initial yield stress and Q_f the internal softening variable in post-elastic regime. The evolution law is defined in terms of the incremental plastic multiplier $\dot{\lambda}_f$ and the softening module H_f.

	1D bond-slip model
Constitutive relationships	$\dot{\varepsilon}_N = \dot{\varepsilon}_N^{el} + \dot{\varepsilon}_N^{pl}$ $\dot{\sigma}_f = E_f(\dot{\varepsilon}_N - \dot{\varepsilon}_N^{pl})$
Yield condition	$f_f = \lvert \sigma_f \rvert - (\sigma_{y,f} + Q_f) \leq 0$
Flow rule	$\dot{\varepsilon}_N^{pl} = \dot{\lambda}_f \partial f_f / \partial \sigma_f = \dot{\lambda}_f sign[\sigma_f]$
Evolution law	$\dot{Q}_f = \dot{\lambda}_f H_f$
Kuhn - Tucker loading/unloading and consistency conditions	$\dot{\lambda}_f \geq 0,\ f_f(\sigma_f, Q_f) \leq 0,\ \dot{\lambda}_f f_f(\sigma_f, Q_f) = 0$ $\dot{f}_f(\sigma_f, Q_f) = 0$

Table 2: 1D bond-slip model.

The complete derivation of this numerical model and its validation against bond-slip experimental tests are proposed in previous works published by the authors [17][21].

4.2 Dowel action

A numerical sub-model for the dowel action has been accounted defining both stiffness and strength of the generic fiber embedded in the concrete matrix and subjected to a possible transverse force/displacement at the fracture level. The well-known Winkler beam theory is used to describe the dowel force-displacement relationship, which transformed in terms of dowel stress vs. relative displacement, allows to describing the equivalent shear beam stiffness. Further details of the numerical model for dowel mechanisms can be found in [21].

5 NUMERICAL SIMULATIONS

The simulations of 150×150×600 mm^3 pre-cracked concrete beams tested under four-point bending according to Martinelli et al. [10] are performed. Specimens of various mixtures, characterized by the same volume fraction of fibers, but different proportions of industrial and recycled reinforcements were tested in bending. The specimens presented a vertical notch (of approximately 2.0 mm wide) at the bottom (mid-length) of the beam characterized by a depth of about 45 mm. The actual beam span length is 450 mm (Figure 2).

Four FRC mixtures have been analyzed, always using 0.5% of fibers in volume of matrix plus a reference concrete without reinforcements, and also combining the aforementioned ISFs and RSFs:

- Plain concrete.
- RSFRC 0-05: with only ISFs (RSFs = 0%).
- RSFRC 50-05: with 50% of ISFs replaced by an equal amount of RSFs.
- RSFRC 100-05: with all RSFs.

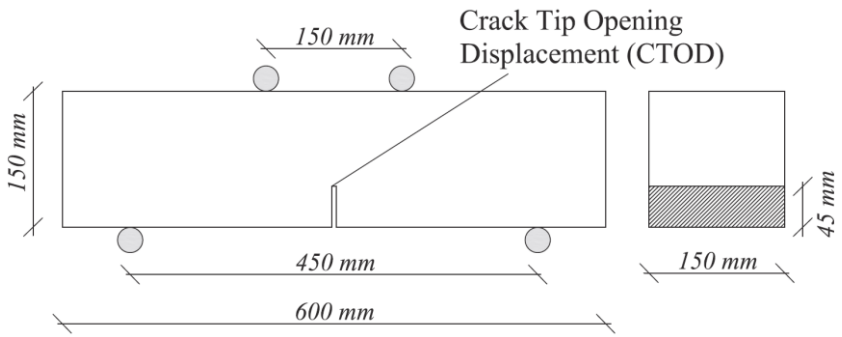
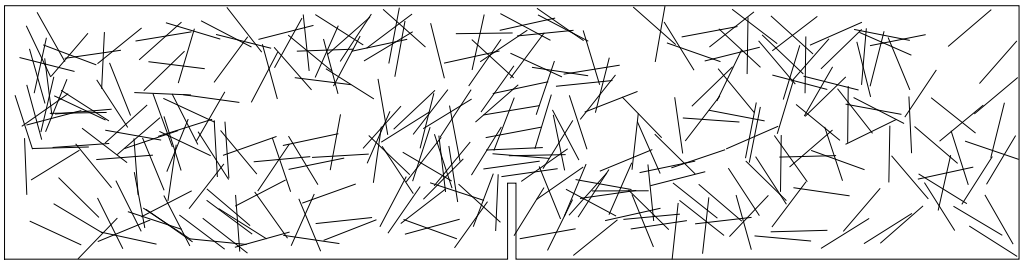
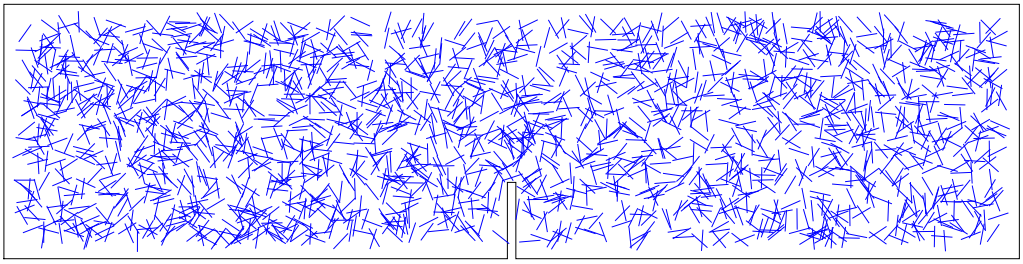


Figure 2: Geometry of the notched beam tested in four-point bending.

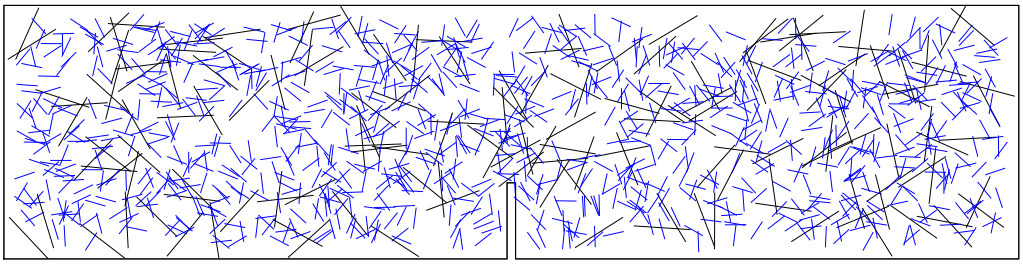
The explicit mesoscopic geometry is determined by means of a random 2D generation of both ISFs and RSFs as shown in Figure 3.



(a) RSFRC 0-05: number of ISFs=316, fiber length ISF=33 mm, diameter ISF=0.55 mm.



(b) RSFRC 100-05: number of RSFs=2076, fiber length (mean value) RSF=12 mm, diameter (mean value) RSF=0.23 mm [10].

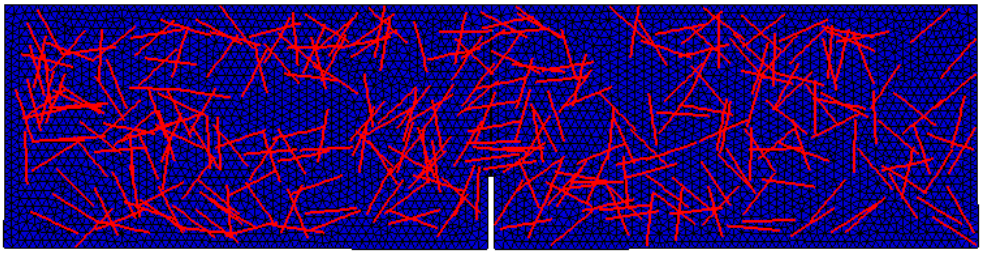


(c) RSFRC 50-05: number of ISFs=158, of RSFs=1038.

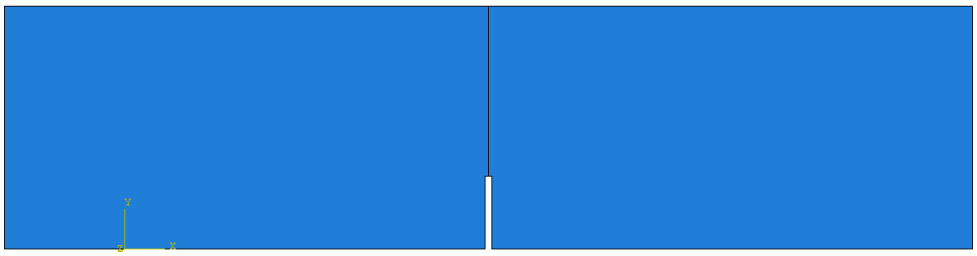
Figure 3: Two-dimensional finite element geometry: concrete phases and fibers as embedded short beams.

Figure 4 reports the 2-D geometry of the considered structure and highlights the FE discretization employed in the analyses. Plane stress hypothesis and displacement-based control are assumed. Moreover, 3-node linear elastic plane stress elements have been adopted in the FE mesh, whereas all non-linearities are concentrated within zero-thickness interface elements defined throughout the adjacent edges of the finite elements in the notch zone. Non-linear

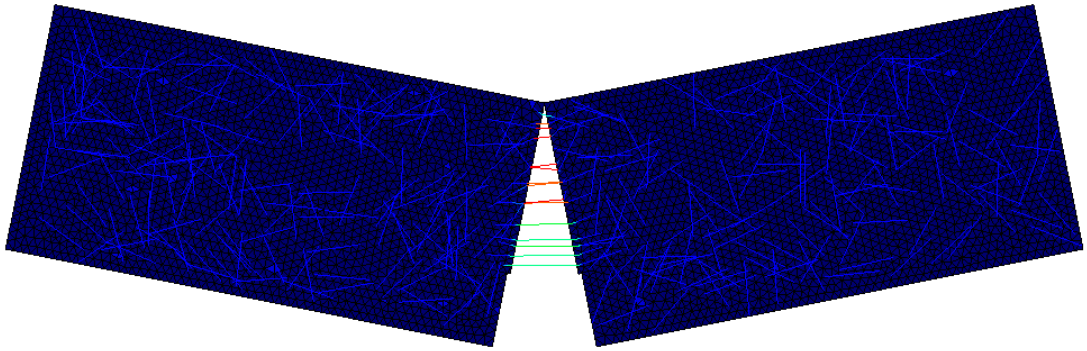
fracture-based law was introduced in those interface elements according to the formulation outlined in Section 3. Then, the constitutive models for considering the stress transferred between cracks through embedded beam elements are outlined in Section 4.



(a) Adopted FE mesh;



(b) Interface zone at the notch section;



(c) Cracked configuration of the RSFRC 0-05 specimen.

Figure 4: Adopted finite element mesh, interfaces position and possible cracked configuration of the 4-point bending beam. As example there is plotted the results of the RSFRC 0-05 specimen.

Figures 5 to 8 show the numerical response in terms of vertical load vs. Crack Tip Opening Displacements (CTOD) against the corresponding experimental results. It can be observed that the proposed model leads to very accurate simulations of the post-cracking response observed in the experimental tests. As expected, the load–CTOD responses of concrete with or without fibers emphasize the significant influence of fibers on the both peak strength and post-peak response of the FRC specimens tested in bending: the brittle behavior of the concrete matrix became significantly tougher when fibers are added as spread reinforcement.

The numerical simulations, as well as the experimental observations, confirm that the post-cracking response of FRC specimens with only ISFs is characterized by the highest toughness, as a result of the superior bond properties and dowel action of these fibers with respect to RSFs.

The model formulation is also capable to capture the effect of replacing increasing amount of ISFs with an equal quantity of RSFs. Both experimental and numerical results highlight as the post-cracking behavior of FRC is generally characterized by a more pronounced softening range in specimens having a greater quantity of RSFs in substitution of ISFs. It should be not-

ed that the local bond–slip and dowel laws were determined through inverse identification on test results obtained on specimens reinforced with only ISFs and/or RSFs. Numerical predictions of hybrid industrial/recycled steel fiber-reinforced concrete have been obtained by just changing the fiber types/contents.

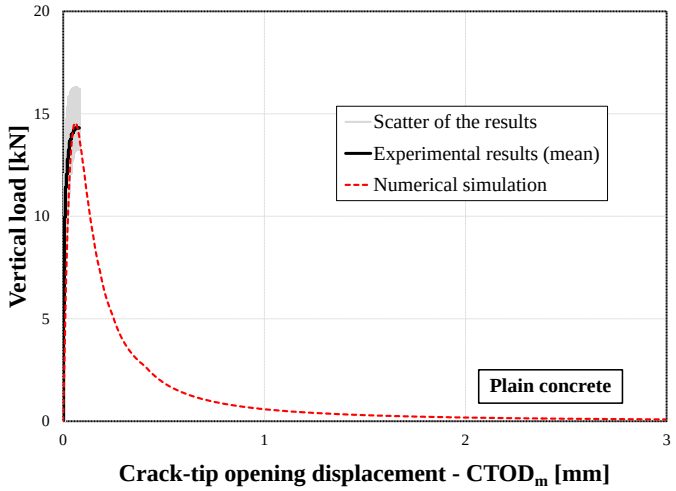


Figure 5: Vertical force – CTOD$_m$ curves of plain concrete.

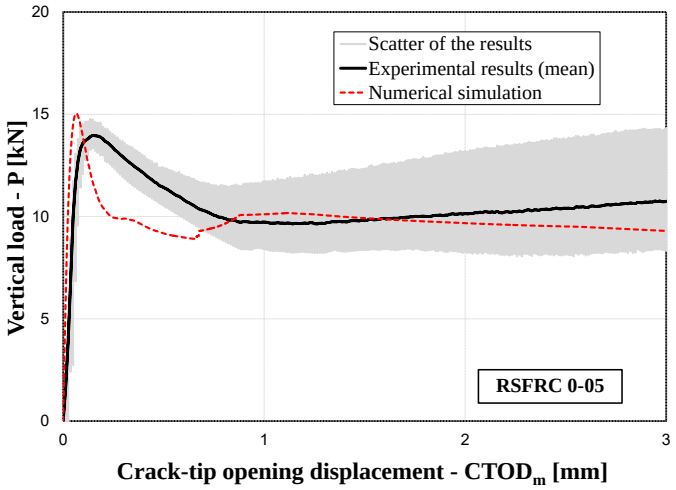


Figure 6: Vertical force – CTOD$_m$ curves of RSFRC 0-05.

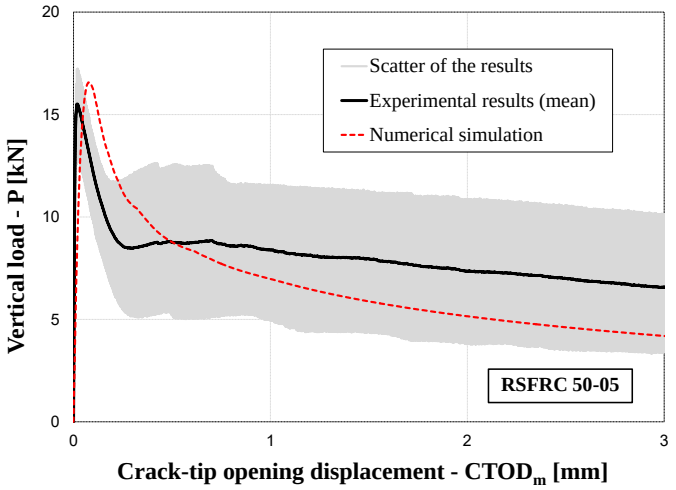


Figure 7: Vertical force – $CTOD_m$ curves of RSFRC 50-05.

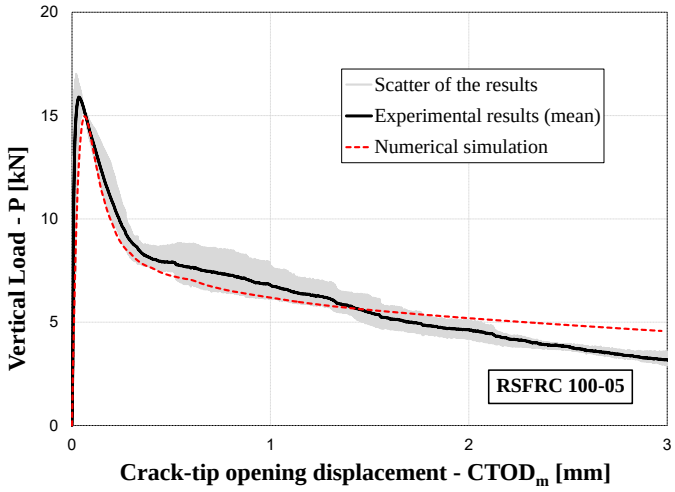


Figure 9: Vertical force – $CTOD_m$ curves of RSFRC 100-05.

6 CONCLUSIONS

This paper has presented a meso-mechanical model for simulating the fracture response of HRSFRCs in bending. It is based on a discontinuous crack finite element approach and assumes dispersed fibers "embedded" within the mesh of the cementitious matrix.

The numerical results demonstrate the capability of the proposed model in simulating the experimental observations derived by four-point bending tests on hybrid recycled/industrial steel fiber-reinforced concrete specimens. Particularly, the model is capable to capture the significant influence of steel fiber contents and types on both the maximum strength and post-peak toughness exhibited by the aforementioned specimens.

Finally, although further experimental comparisons are needed for achieving a full validation, the proposed model paves the way toward predicting the fracture behavior of HRSFRCs as a result of the mechanical properties and geometric distribution of their key constituents.

7 ACKNOWLEDGMENTS

The model proposed in this paper stems out of the activities of the SUPERCONCRETE Project (H2020-MSCA-RISE-2014, n. 645704): the Authors wish to acknowledge the financial contribution of the European Union, as part of the H2020 Programme.

REFERENCES

[1] M. Sienkiewicz, J. Kucinska-Lipka, H. Janik, A. Balas, Progress in used tyres management in the European Union: a review, *Waste Manage.* **32**(10), 1742-1751, 2012.

[2] EU-Directive1999/31/EC of the council of 26 April 1999 on the landfill of waste, *Off. J. Eur. Union L*, **182**, 1-19, 1999.

[3] E. Ganjian, M. Khorami, A.A. Maghsoudi, Scrap-tyre-rubber replacement for aggregate and filler in concrete, *Constr. Build. Mater.*, **23**(5), 1828-1836, 2009.

[4] G. Centonze, M. Leone, M. Aiello, Steel fibers from waste tires as reinforcement in concrete: a mechanical characterization, *Constr. Build. Mater.*, **36**, 46-57, 2012.

[5] A.G. Graeff, K. Pilakoutas, K. Neocleous, M.V.N. Peres, Fatigue resistance and cracking mechanism of concrete pavements reinforced with recycled steel fibres recovered from post-consumer tyres, *Eng. Struct.*, **45**, 385-395, 2012.

[6] W.H. Yung, L.C. Yung, L.H. Hua, A study of the durability properties of waste tire rubber applied to self-compacting concrete, *Constr. Build. Mater.*, **41**, 665-672, 2013.

[7] A. Caggiano, M. Cremona, C. Faella, C. Lima, E. Martinelli, Fracture behavior of concrete beams reinforced with mixed long/short steel fibers, *Constr. Build. Mater.*, **37**, 832-840, 2012.

[8] M. Elser, E. Tschegg, S. Stanzl-Tschegg, Fracture behaviour of polypropylenefibre-reinforced concrete under biaxial loading: an experimental investigation, *Compos Sci Technol*, **56**(8), 933-945, 1996.

[9] N. Banthia, F. Majdzadeh, J. Wu, V. Bindiganavile, Fiber synergy in hybrid fiber reinforced concrete (hyfrc) in flexure and direct shear, *Cem Concr Compos*, **48,** 91-97, 2014.

[10] E. Martinelli, A. Caggiano, H. Xargay. An experimental study on the post-cracking behaviour of Hybrid Industrial/Recycled Steel Fibre-Reinforced Concrete, *Construction and Building Materials*, **94**, 290-298, 2015.

[11] F. de Andrade Silva, B. Mobasher, R.D.T. Filho, Fatigue behavior of sisal fiber reinforced cement composites, *Mater Sci Eng: A*, **527**(21–22), 5507-5513, 2010.

[12] M. Nataraja, N. Dhang, A. Gupta, Stress-strain curves for steel-fiber reinforced concrete under compression, *Cem Concr Compos*, **21**(56), 383-390, 1999.

[13] A. Brandt, M. Glinicki, J. Potrzebowski, Application of FRC in construction of the underground railway track, *Cem Concr Compos*, **18**(5), 305-312, 1996.

[14] A. Caggiano, G. Etse, E. Martinelli, Zero-thickness interface model formulation for failure behavior of fiber-reinforced cementitious composites, *Comput Struct*, **98-99**, 23-32, 2012.

[15] V.M. Cunha, J.A. Barros, J.M. Sena-Cruz, A finite element model with discrete embedded elements for fibre reinforced composites, *Comput Struct*, **94-95**, 22-33, 2012.

[16] G. Etse, A. Caggiano, S. Vrech. Multiscale failure analysis of fiber reinforced concrete based on a discrete crack model. *International J of Fracture*, **178**.1-2, 131-146, 2012.

[17] A. Caggiano, E. Martinelli, A unified formulation for simulating the bond behaviour of fibres in cementitious materials, *Mater Des*, **42**, 204-213, 2012.

[18] A. Caggiano, H. Xargay, P. Folino, E. Martinelli, Experimental and numerical characterization of the bond behavior of steel fibers recovered from waste tires embedded in cementitious matrices, *Cement and Concrete Composites*, **62**, 146-155, 2015.

[19] I. Carol, P. Prat, C. Lopez, Normal/shear cracking model: applications to discrete crack analysis. *ASCE-J Engrg Mech*, **123**, 765-773, 1997.

[20] A. Caggiano, G. Etse, Coupled thermo–mechanical interface model for concrete failure analysis under high temperature. *Computer Methods in Applied Mechanics and Engineering*, **289**, 498-516, 2015.

[21] S. Vrech, G. Etse, A. Caggiano. Thermodynamically consistent elasto-plastic microplane formulation for fiber reinforced concrete. *International Journal of Solids and Structures*, **81**, 337-349, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THE KIRKENDALL SHIFT AND FRENKEL EFFECT DURING MULTI-COMPONENT DIFFUSION PROCESS

Bartek Wierzba[1]

[1] Rzeszow University of Technology, Faculty of Mechanical Engineering and Aeronautics,
al. Powstańców Warszawy 12, 35-959 Rzeszow, Poland.
e-mail: bwierzba@prz.edu.pl

Keywords: Kirkendall plane, voids, diffusion, mass conservation.

Abstract. *In this paper the phenomenological process related to the evolution of the ternary and higher system is discussed. The two coupled phenomena related to the difference in diffusion coefficients will be presented by numerical investigations - mainly the Kirkendall and Frenkel effects. Such a difference leads to the: lattice shift, stress generation and relaxation, non-equilibrium distribution of vacancies and voiding. The generalized Darken method (bi-velocity) of the multi-component system is formulated. The approximation method allows for determination the evolution of the voids. Thus the relaxation time for the vacancies in terms of mean migration length and vacancy diffusion coefficient in case of multicomponent system will be formulated. Moreover, the void evolution in Fe-Pd system will be calculated.*

1 INTRODUCTION

The Navier-Stokes set of equations is a theoretical model for continuum fluid dynamics where the local thermodynamic equilibrium holds. However, this equations encounters some challenging difficulties for non-continuum flows. This phenomena was widely studied in the gas community, where the gas density is usually very low. In recent years, non-continuum flows have also attracted much attention with the rapid development in microelectromechanical systems. However, due to the finite Knudsen number effect, the continuum-equilibrium assumption may break down and the model will fail to work for these flows [1, 2, 3].

In the recent years Brenner proposed new model based on the fluid velocity [4, 5, 6]. He introduced one single additional term into each of the momentum and energy equations. Generally Brenner assumed that the mass velocity in the hydrodynamics equations should be replaced by volume velocity which relates to the flux of volume rather than mass [4, 5, 6].

On the other hand, diffusion community knows very well an alternative way of providing volume conservation. Namely, it is a lattice drift (Kirkendall effect) caused by the vacancy flux divergence leading to dislocation climb and subsequently to the construction of extraplanes in accumulation region and dismantling atomic planes in depleting region [7, 8, 9]. The drift velocity is than generated. This constraint means zero divergence of overall volume flux density:

$$\frac{\partial}{\partial x}\sum_{i=1}^{r}\Omega_i^m J_i = 0$$

When the only driving force is chemical potential gradient, $\mu_i = \mu_i^{ch}$, the volume flux is given by:

$$\Omega_i^m J_i = D_i^* \varphi \frac{\partial \rho_i}{\partial x} + \rho_i \upsilon .$$

In this paper the intrdiffusion in solid state description will be formulated. The mass, heat, volume continuity equation, flux definition and vacancy evolution equation will be presented. We will focus on the kinetic effects that are related to difference of mobilities, namely on the stress generation and relaxation [3,10]. Interdiffusion leads to the accumulation of matter at the side of slower component of the diffusion couple. Due to the volume constraint, nature just must find some ways to reduce this accumulation to zero. This reduction, will be realized here by introduction of: 1) Kirkendall effect; 2) Backstress effect and 3) non-

equilibrium vacancy distribution. The last paragraph will present the results of the model mainly, the voids evolution during interdiffusion process in Fe-Pd system.

2. THE MASS AND HEAT EQUATIONS

Mass conservation law. Consider multicomponent mixture, where ρ_i denote the *i*-th component density. The evolution of the density is described by the equation:

$$\frac{\partial \rho_i}{\partial t} + \operatorname{div}\left(\rho_i v_i\right) = 0 \qquad (1)$$

where $J_i = \rho_i v_i$ is the overall flux of the *i*-th component and v_i denote its volume velocity (the medium velocity). We do not take into account mass production or consumption in this equation.

Heat balance. Diffusive flux can influence the heat balance. The heat flux can be defined after Fourier (for isotropic material) as:

$$J^q = -k \operatorname{grad} T \qquad (2)$$

where k denote the thermal conductivity and T is the temperature. We can define the enthalpy of the system according to internal energy density as:

$$h = u + \frac{\operatorname{tr}(\sigma)}{\rho} \qquad (3)$$

The differential form of the above equation is:

$$\left.\frac{D\rho h}{Dt}\right|_v = \left.\frac{D\rho u}{Dt}\right|_v + \left.\frac{D\operatorname{tr}(\sigma)}{Dt}\right|_v \approx \sum_i \left.\frac{D\rho_i u_i}{Dt}\right|_{v_i} + \left.\frac{D\operatorname{tr}\left(\sum_i \Omega_i^m \rho_i \sigma\right)}{Dt}\right|_{v_i} \qquad (4)$$

where $\left.\frac{DT}{Dt}\right|_v$ denote the Lagrange derivative defined over the velocity v:

$$\left.\frac{DT}{Dt}\right|_v = \frac{\partial T}{\partial t} + v \operatorname{grad} T, \qquad (5)$$

Assuming, that the changing enthalpy in time is proportional to its specific heat and temperature, the following equation can be written:

$$\left.\frac{Dh}{Dt}\right|_v = c_p \left.\frac{DT}{Dt}\right|_v \qquad (6)$$

Finally, by introducing Eqs. (2) - (6) into the energy conservation law and the mechanical parts are neglected, than the final heat transport equation can be derived:

$$\rho c_p \left.\frac{DT}{Dt}\right|_v \approx \mathrm{div}\left(k \ \mathrm{grad}\ T\right) \tag{7}$$

The fluxes. The overall flux and velocity of the mass are defined by the volume drift and diffusion velocities:

$$\rho\Omega^m v = \sum_i \left(\rho_i \Omega_i^m v^{drift} + \rho_i \Omega_i^m v_i^d\right) \qquad \text{where} \quad v := v^{drift} + v^d = v^{drift} + \sum_i \rho_i \Omega_i^m v_i^d \,, \tag{8}$$

where $i = 1,\ldots, r$ and r denotes the number of components and $N_i^m = \rho_i / \rho$ denote the molar fraction.

The volume drift velocity. During an arbitrary transport process when volume is affected by the distribution of every mixture component and the stress field, from the Liouville theorem it follows:

$$\sum_i \frac{\partial \rho_i \Omega_i^m}{\partial t} + \sum_i \mathrm{div}\left(\rho_i \Omega_i^m v_i\right) = \mathrm{div} v^\sigma \,. \tag{9}$$

We consider here a multicomponent solution in a closed system, where partial molar volumes and elastic properties do not depend on composition, $\Omega_i \neq f\left(N_1,...,N_r\right)$. Thus, the volume can be only affected by the external forces:

$$\mathrm{d}\left|\Omega\right| / \mathrm{d}t = \int_{\partial|\Omega|} v^\sigma \,\mathrm{d}x \,. \tag{10}$$

where: υ^σ is a velocity generated by external forces (elastic deformation). Introducing Euler relation ($\sum_i \rho_i \Omega_i^m = \rho\ \Omega^m \equiv 1$) and Eq. (8), (9) allow to calculate the drift velocity of the mixture:

$$\sum_i \mathrm{div}\left(v^{drift} + \rho_i \Omega_i^m v_i^d\right) = \mathrm{div} v^\sigma \tag{11}$$

Denoting the Darken velocity, v^D as:

$$v^D = -\rho_i \Omega_i^m v_i^d \tag{12}$$

The final form of the drift velocity can be rewritten in the form:

$$\mathrm{div}\ v^{drift} = \mathrm{div}\left(v^\sigma + v^D\right) \tag{13}$$

Equation (13) defines the drift velocity of the system. In one dimension the drift velocity can be expressed by analytical function as:

$$v^{drift} \approx v^\sigma + v^D \tag{14}$$

In this paper we will consider, additionally four more effects: 1) Kirkendall effect, 2) Backstress 3) vacancy generation effect and 4) thermal gradient

The **Kirkendall effect** means the movement of lattice from slower diffusant side towards the faster diffusant side with some drift velocity, v^{drift}. Thus this effect results in diffusion flux as follow:

$$J_i = \rho_i v_i^d + \rho_i v^{drift}, \qquad (15)$$

The diffusion flux is defined by the Nernst-Planck flux equation [11,12] which in general form reads:

$$J_i = \rho_i B_i \operatorname{grad}\mu_i + \rho_i v^{drift}, \qquad (16)$$

where μ_i is the generalized diffusion potential of the *i*-th component.

Backstress effect - the diffusion potential is affected by the internal stress effect - stress gradient appearing due to attempt of matter accumulation. Each diffusing atom of both species is affected by common stress force:

$$\operatorname{grad}\mu_i^{Int} = -\operatorname{grad}\left(\Omega_i^m p^{Int}\right) \qquad (17)$$

Non-equilibrium vacancy distribution - the diffusion potential is a difference of component chemical and common vacancy potentials, $\mu_i = \mu_i^{ch} - \mu^V$, and equalization of the diffusion fluxes instead of lattice shift, is provided by the non-equilibrium vacancy gradient appearing due to attempt of matter accumulation. Role of effective force here is played by the gradient of vacancy chemical potential, proportional to the gradient of deviation of vacant sites fraction from its local equilibrium value:

$$\mu^V = -kT \ln \frac{\rho_V}{\rho_V^{eq}} \qquad (18)$$

Temperature distribution - the isothermal heat transmitted by moving the atom in the process of jumping a lattice site less the intrinsic enthalpy [13, 14]:

$$\operatorname{grad}\mu^T = -\frac{1}{T} Q^* \operatorname{grad} T \qquad (19)$$

where: Q^* is defined as heat of transport. Finally, the diffusion potential is a sum of component chemical and common vacancy and stress potentials, $\mu_i = \mu_i^{ch} - \mu^V + \mu_i^{Int} + \mu^T$. Consequently, the gradient of the diffusion potential is defined as follow:

$$\operatorname{grad}\mu_i = -kT\frac{1}{\rho_i}\operatorname{grad}\rho_i - \frac{kT}{\rho_V^{eq}}\operatorname{grad}\left(\rho_V - \rho_V^{eq}\right) - \operatorname{grad}\left(\Omega_i^m p^{Int}\right) - \frac{1}{T}Q^*\operatorname{grad}T \tag{20}$$

Internal pressure - the internal pressure is a result of the difference in the diffusion coefficients of the components and difference in the lattice in diffusion couple [15]:

$$\frac{\partial p^{Int}}{\partial t} = -\frac{E}{3(1-2\upsilon)}\operatorname{div}\left(v^{\sigma} + v^{d}\right), \tag{21}$$

where: the overall diffusion velocity is defined as:

$$v^d = \sum_i \rho_i \Omega_i^m v_i^d \tag{22}$$

Vacancy exchange. The last equation defining the model is the vacancy exchange. The equation is defined as:

$$\frac{\partial \rho_v}{\partial t} - \operatorname{div}\sum_{i=1}^{r}\rho_i B_i \operatorname{grad}\mu_i + \frac{\rho_v - \rho_v^{eq}}{\tau_v} = 0 \tag{23}$$

where N_v is the vacancy molar fraction, j_v is the vacancy flux. The N_v^{eq} and τ_v denote the vacancy equilibrium molar fraction and relaxation time, respectively.

3. RESULTS.

In this section the results of interdiffusion and voids formation will be shown. The voids formation will be analyzed in two-dimmensional space. For experimental verification the Fe-Pd binary system was used. The voids radii will be estimated and the results for different calculation times will be shown. The following set of the equation will be solved:

$$\frac{\partial \rho_i}{\partial t} + \operatorname{div}\left(\rho_i v_i^d + \rho_i v^D\right) = 0 \quad \text{on} \quad \Omega \setminus \Omega_V \qquad \text{- mass conservation}$$

$$\frac{\partial \rho_v}{\partial t} - \operatorname{div}\sum_{i=1}^{r}\rho_i B_i \operatorname{grad}\mu_i + \frac{\rho_v - \rho_v^{eq}}{\tau_v} = 0 \quad \text{on} \quad \Omega \setminus \Omega_V \qquad \text{- vacancy exchange}$$

$$\frac{\partial \rho_i}{\partial t} = \frac{\partial \rho_V}{\partial t} = 0 \quad \text{on} \quad \Omega_V$$

where: $\operatorname{grad}\mu_i = -kT\frac{1}{\rho_i}\operatorname{grad}\rho_i - \frac{kT}{\rho_V^{eq}}\operatorname{grad}\left(\rho_V - \rho_V^{eq}\right)$. This set of the equation will be supplemented with boundary conditions:

$$\rho_i v_i^d = 0 \quad \text{on} \quad \partial\Omega \qquad \text{- diffusion flux on the boundary}$$

The subset Ω_V denote the position of the voids. The voids are initially introduced onto the

calculation mesh. Thus, only the growth will be simulated. In general, the voids growth on the faster diffusion couple side. In Fe-Pb system the voids growth on the iron reach side of the diffusion couple. We assume that in Fe-Pd system the mean migration length for vacancy was $L_V = 10^{-8} m$. The diffusion coefficient was calculated from Boltzmann-Matano analysis (from known experimental results of diffusion in Fe-Pd system at 1273K for 150h [16]), i.e. at the concentration point $N_{Fe} = 0.75$ the diffusion coefficients was estimated as: $D_{Fe} = 1.17 \cdot 10^{-15}$ and $D_{Pd} = 9.27 \cdot 10^{-16}\ m^2 s^{-1}$. The results for different calculation times are shown on figures 1-2.

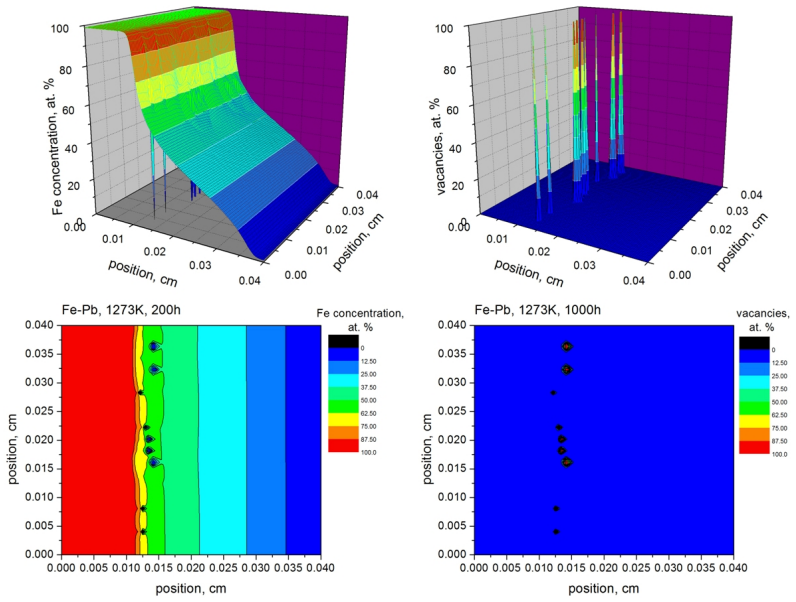


Figure 1. The simualtion results of the a) - c) Fe concentration and b) - d) voids formation (vacancies evolution) in Fe-Pd system after annealing at 1273K for 200h

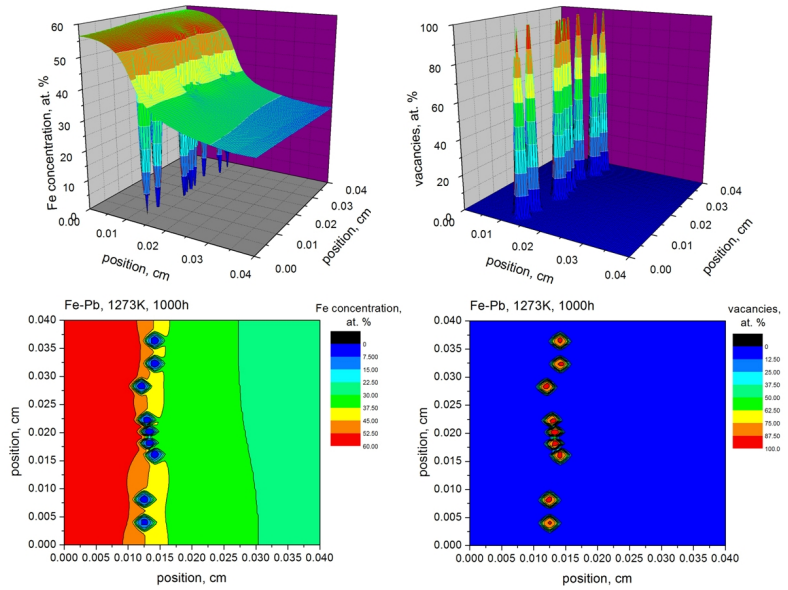


Figure 2. The simualtion results of the a) - c) Fe concentration and b) - d) voids formation (vacancies evolution) in Fe-Pd system after annealing at 1273K for 1000h

4. CONCLUSIONS.

Present paper shows the method which allow for calculation of the physical models of interdiffusion. The drift velocity was calculated in case of presence of the vacancy, backstress and chemical force fields. Moreover, it was shown, that the interdiffusion process can be influenced by the vacancy and heat fields. The model was checked by simulating voids growth during diffusion couple experiments. It was presented, that the voids are formed at the faster diffusion side of the diffusion couple. Moreover the radii of the void changes parabolically with time. In the future the simulation should take into account, that the voids are moving during the diffusion process.

Acknowledgements. This work has been supported by the National Science Centre (NCN) in Poland, decision number 2013/09/B/ST8/00150.

REFERENCES

[1] E. P. Muntz, Annu. Rev. Fluid Mech., **21**, 387, 1989.
[2] C. M. Ho, Y. C. Tai, Annu. Rev. Fluid Mech., **30**, 579, 1998.
[3] Z. Guo, and K. Xu, Adv. Appl. Math. Mech., **1,** 391, 2009.
[4] H. Brenner, Physica A, **349**, 11, 2005.
[5] H. Brenner, Physica A, **349**, 60, 2005.

[6] H. Brenner, Physica A, **370**, 190, 2006.
[7] L. S. Darken, Trans. AIME, **174**, 184, 1948.
[8] A.D. Smigelskas, E.O. Kirkendall, Trans. AIME, **171,** 130, 1947.
[9] J. Bardeen, C. Herring, Diffusion in alloys and the Kirkendall effect. In Shockley W (ed) Imperfections in Nearly Perfect Crystals, Wiley, New York, 1952.
[10] R. W. Balluffi, S. M. Allen, Carter W.C. and Kemper R. A., Kinetics of materials, Wiley, Hoboken, (NJ) 2005.
[11] W. Nernst, Z Phys Chem (N F) **4**, 129, 1889.
[12] M. Planck, Annu Rev Phys Chem **40**, 561, 1890.
[13] C. Basaran, M. Lin, International Journal of Solids and Structures **44**, 4909, 2007.
[14] C. Chen, H.-Y. Hsiao, Y.-W. Chang, F. Ouyang, K.N. Tu, Materials Science and Egineering R **73**, 85, 2012.
[15] B. Wierzba, M. Danielewski, Physica A **390**, 2325, 2011.
[16] B. Wierzba, W. Skibiński, S. Wędrychowicz, P. Wierzba, Physica A **433**, 268, 2015

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CRACK PATH FIELD AND STRAIN INJECTION TECHNIQUES IN DYNAMIC FRACTURE SIMULATIONS

Oriol Lloberas-Valls[1], Alfredo E. Huespe[2], J. Oliver[1] and Ivo F. Dias[3]

[1]CIMNE – Centre Internacional de Metodes Numerics en Enginyeria
Campus Nord UPC, Edifici C-1, c/Jordi Girona 1-3, 08034 Barcelona, Spain
e-mail: olloberas@cimne.upc.edu, xavier.oliver@upc.edu

[2] INTEC-UNL-CONICET
Gemes 3450, Santa Fe, Argentina
e-mail: ahuespe@intec.unl.edu.ar

[3] Laboratório Nacional de Engenharia Civil (LNEC)
Avenida Brasil 101, 1700 Lisbon, Portugal
e-mail: idias@lnec.pt

Keywords: Fracture dynamics, strong discontinuity approach, crack path field, strain injection techniques.

Abstract. *Dynamic fracture phenomena are studied employing low cost computational tools based on Finite Elements with Embedded strong discontinuities (E-FEM). Fracture nucleation and propagation are accounted for through the injection of discontinuous strain and displacement modes inside the finite elements. The Crack Path Field technique is employed to compute the trace of the strong discontinuity during fracture propagation.*

Unstable crack propagation and crack branching are observed upon increasing loading rates. The variation in terms of crack pattern and energy dissipation is studied and a good correlation is found between the maximum experimental crack speed and maximum dissipation at the onset of branching. Comparable results are obtained against simulations employing supra-elemental techniques, such as phase-field and gradient damage models, considering coarser discretizations which can differ by two orders of magnitude.

1 INTRODUCTION

Dynamic fracture processes are challenging phenomena to be studied experimentally. Expensive equipment needs to be employed in order to accurately capture crack growth in a small time frame and dynamic loading conditions are not always straightforward to reproduce in a laboratory.

For this reason, numerical techniques such as Finite element (FE) methods are regarded as valuable alternatives for solving general dynamic fracture problems. However, they show some limitations mainly related with the objectivity, computational cost and the overall ability to predict experimental tests [36]. Different alternatives have been adopted to model crack nucleation and propagation in a dynamic setting. For instance works involving fracture in dynamic problems have been addressed by Falk et al. [7], Pandolfi et al. [26], Song and Belytschko [35] and Linder and Armero [16], where cracks are inserted between FE (inter-elemental) and inside the FE support (intra-elemental). Other techniques tackle the strain localization phenomena by considering supra-elemental bands such as phase-field modeling [2, 11, 19] and gradient damage models [15]. Such techniques led to impressive 2D and 3D results but at the cost of extremely fine FE discretizations which can be regarded prohibitive in a dynamic context where a large amount of time steps need to be resolved in order to capture crack growth with a sufficient time resolution. Softening visco-elastic visco-plastic damage continuum model has been employed for dynamic fracture of concrete up to intermediate loading rates in [27, 28]. At high loading rates, where fragmentation and spalling take place, erosion or element deletion models proposed by Camacho and Ortiz [3], Li et al. [14] have proven to be competitive methods. Other alternative numerical methodologies have been addressed with some success, e.g. peridynamics [10, 32], discrete methodologies such as lattice models [13] and mesh-free methodologies [1], to mention a few.

One of the main objectives of such numerical approaches is to help understanding dynamic fracture phenomena mainly driven by inertial forces which play a dominant role over possible viscous behaviors at high loading rates, e.g. crack curving and branching phenomena detected experimentally when a critical crack tip velocity is exceeded [9, 29–31]. In this scenario, computationally affordable intra-element techniques need to be ready to account for complex fracture patterns in which the dominant crack paths may involve branching and sudden changes of the crack propagation direction.

The present contribution introduces and assesses a finite element method for modeling crack propagation problems with the presence of a dominant crack in brittle or quasi-brittle materials. Problems involving fragmentation or spalling are left outside the scope of this work. Our approach is based on the Embedded Finite Element technology (cf. [4, 23]) which has already been utilized for the study of fracture in quasi-brittle materials and successfully applied to the study of tensile crack growth in gravity dams (cf. [5, 6]). The specific formulation developed by Oliver et al. [24] and assessed for quasi-static multiscale fracture problems has been adopted and tailored for dynamic fracture propagation problems. One of the main advantages of the present approach is that cracks, represented by strong discontinuities embedded into the finite elements, may propagate through the mesh in arbitrary directions reproducing complex failure patterns and, therefore, significantly coarser meshes can be employed compared to other supra-elemental techniques such as phase-field or gradient models.

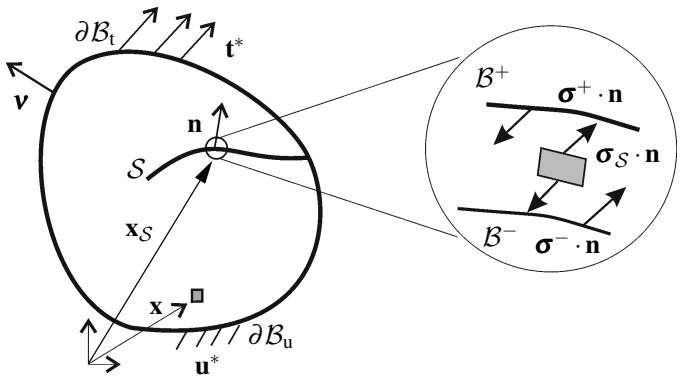


Figure 1: IVBP at solid $\mathcal{B}$ with evolving crack. The disjoint domains of $\mathcal{B}$ generated by $\mathcal{S}$ are denoted $\mathcal{B}^+$ and $\mathcal{B}^-$.

2 MODEL DESCRIPTION

The governing equations of the dynamic fracture Initial Boundary Value Problem (IBVP) at time $t \in [0,T]$ of the solid $\mathcal{B}$ (cf. Figure 1) can be stated as: find $\mathbf{u}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$, satisfying:

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}+\mathbf{b}=\rho\ddot{\mathbf{u}}\,; \quad \forall\mathbf{x}\in\mathcal{B}\backslash\mathcal{S}\,; \quad t\in[0,T]\,; \quad \text{Momentum equation} \tag{1}$$

$$\boldsymbol{\varepsilon}=\boldsymbol{\nabla}^{\mathrm{s}}\mathbf{u}\,; \quad \forall\mathbf{x}\in\mathcal{B}\backslash\mathcal{S}\,; \quad t\in[0,T]\,; \quad \text{Compatibility equation} \tag{2}$$

$$\boldsymbol{\sigma}=\boldsymbol{\Sigma}(\boldsymbol{\varepsilon},r)\,; \quad \forall\mathbf{x}\in\mathcal{B}\,; \quad t\in[0,T]\,; \quad \text{Constitutive equation} \tag{3}$$

$$\mathbf{u}=\mathbf{u}^*(\mathbf{x},t)\,; \quad \forall\mathbf{x}\in\partial\mathcal{B}_{\mathrm{u}}\,; \quad t\in[0,T]\,; \quad \text{Displ. boundary conditions} \tag{4}$$

$$\boldsymbol{\sigma}\cdot\mathbf{n}=\mathbf{t}^*(\mathbf{x},t)\,; \quad \forall\mathbf{x}\in\partial\mathcal{B}_{\mathrm{t}}\,; \quad t\in[0,T]\,; \quad \text{Traction boundary conditions} \tag{5}$$

$$\mathbf{u}(\mathbf{x},0)=\mathbf{0}\,; \quad \forall\mathbf{x}\in\mathcal{B}\,; \quad \text{Initial displ. condition} \tag{6}$$

$$\dot{\mathbf{u}}(\mathbf{x},0)=\dot{\mathbf{u}}_0(\mathbf{x})\,; \quad \forall\mathbf{x}\in\mathcal{B}\,; \quad \text{Initial velocity condition} \tag{7}$$

$$\boldsymbol{\sigma}^+\cdot\mathbf{n}=\boldsymbol{\sigma}_S\cdot\mathbf{n}=\boldsymbol{\sigma}^-\cdot\mathbf{n}\,; \quad \forall\mathbf{x}_\mathcal{S}\in\mathcal{S}\,; \quad t\in[\tau,T]\,; \quad \text{Traction continuity across } \mathcal{S} \tag{8}$$

where $\mathbf{u}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$ correspond to the displacement, strain and stress fields and $\mathbf{b}$, ρ and $\boldsymbol{\Sigma}$ denote the volumetric forces, the density and the constitutive relation, respectively. The instant when the discontinuity surface $\mathcal{S}$ is introduced at $\mathbf{x}_\mathcal{S}$ is denoted by τ and $\boldsymbol{\sigma}_\mathcal{S}$, $\boldsymbol{\sigma}^+$, $\boldsymbol{\sigma}^-$ and $\mathbf{n}$ stand for the stress at an interface point of $\mathcal{S}$, the stresses at each side of this interface and the normal to the discontinuity $\mathcal{S}$.

According to the the Continuum Strong Discontinuity Approach (CSDA) introduced in [21] and [22], the stresses $\boldsymbol{\sigma}_\mathcal{S}$ are determined through the constitutive model $\boldsymbol{\Sigma}$ selected for the bulk material accounting for regularization issues.

The strong form of the fracture propagation problem stated in (1) to (8) is recast in a variational format following the methodology presented in [33] and by Oliver et al. [22] where both displacement and strain fields are defined as the addition of a smooth and an enhanced part representing the corresponding singularity in the kinematics. The equivalent variational statement reads:

$$\int_\mathcal{B}\boldsymbol{\sigma}:\boldsymbol{\nabla}^{\mathrm{s}}\boldsymbol{\eta}\,d\mathcal{B}+\int_\mathcal{B}(\mathbf{b}-\rho\ddot{\mathbf{u}})\cdot\boldsymbol{\eta}\;d\mathcal{B}+\int_{\partial\mathcal{B}_{\mathrm{t}}}\mathbf{t}^*\cdot\boldsymbol{\eta}\;d\Gamma=0 \qquad ;\ \forall\,\boldsymbol{\eta}\ \in\ \hat{\mathcal{V}}_{\mathrm{u}}, \tag{9}$$

where $\mathbf{u}$ and $\boldsymbol{\eta}$ stand for the displacements and displacement variations belonging to the spaces of kinematically admissible displacements $\mathcal{V}_{\mathrm{u}}$ and displacement variations $\hat{\mathcal{V}}_{\mathrm{u}}$, respectively.

2.1 Finite element discretization with different kinematic descriptions

The finite element technology employed in this study is designed to account for different kinematic descriptions which are found optimal to reproduce different stages of fracture nucleation and propagation. Strain localization is considered at the surface $\mathcal{S}$ (cf. Figure 1) which is $k-$regularized in the finite element model, i.e $\mathcal{S}$ is represented by a band of finite thickness k across which the displacements are assumed to be continuous.

Quadrilateral finite elements are employed in all considered discretizations and the resulting mesh is basically divided in two disjoint domains: $\mathcal{B}_{\text{std}}(t)$ and $\mathcal{B}_{\text{inj}}(t)$ ($\mathcal{B} = \mathcal{B}_{\text{std}}(t) \cup \mathcal{B}_{\text{inj}}(t)$) accounting for the standard and injected enhanced kinematics, respectively. These vary with time throughout the evolution of the fracture process as depicted in Figure 2. Domains with injected kinematics $\mathcal{B}_{\text{inj}}(t)$ are, in turn, composed by the union of the disjoint domains $\mathcal{B}_{\text{wd}}$, i.e. with weak discontinuity kinematics, and $\mathcal{B}_{\text{sd}}$, i.e. with strong discontinuity kinematics satisfying $\mathcal{B}_{\text{inj}}(t) = \mathcal{B}_{\text{wd}}(t) \cup \mathcal{B}_{\text{sd}}(t)$. Therefore, three types of domains with different kinematics can be identified in the partitions of $\mathcal{B}$:

1. The domain $\mathcal{B}_{\text{std}}$ corresponds to the union of those finite elements with standard kinematics, i.e. continuous displacements. The finite element formulation employed in $\mathcal{B}_{\text{std}}$ is also termed irreducible formulation since displacements are the only components of the solution field. Bilinear polynomials are used to interpolate the displacement field throughout the quadrilateral elements. The standard kinematics is employed in all FE (belonging to $\mathcal{B}_{\text{std}}$) that are found far away from the main fracture process and, for this reason, no localization phenomena is detected. Due to the standard character of its implementation no further details concerning the formulation of standard kinematics FE are given in this manuscript.

2. The domain $\mathcal{B}_{\text{wd}}$ corresponds to the union of those finite elements in $\mathcal{B}_{\text{inj}}$ where a weak displacement discontinuity is considered in the kinematics. Such kinematic enrichment is also referred to as constant stress-discontinuous strain mode (CS-DSM) in Section 2.1.1 where further formulation details are provided. The weak discontinuity kinematics is employed at regions susceptible of undergoing strain-localization phenomena, i.e. where strains exceed a certain critical threshold. These domains are typically found at a vicinity of the crack tip or in those locations where stress concentration may lead to the nucleation of fracture phenomena. The enhanced deformation of the elements belonging to $\mathcal{B}_{\text{wd}}$, amenable to carry a non-directional discontinuity, renders a flexible element which is particularly useful to determine the correct propagation of the crack and to accommodate the strain states at crack branching regions.

3. The domain $\mathcal{B}_{\text{sd}}$ corresponds to the union of those finite elements in $\mathcal{B}_{\text{inj}}$ in which a strong displacement discontinuity is considered in the kinematics. Such kinematic enrichment is also referred to as discontinuous displacement mode (DDM) in Section 2.1.2 where further formulation details are provided. The strong discontinuity kinematics is employed throughout the trace of the crack and the elements belonging to $\mathcal{B}_{\text{sd}}$ must exceed a certain strain threshold and, additionally, fulfill a number of conditions, detailed in Section 2.1.2.

The domain $\mathcal{B}_{\text{inj}}$ is, therefore, modeled with finite elements equipped with strain injections, including CS-DSM and DDM modes introduced in [18, 24]. Strain injections are included in finite elements through the concept of assumed enhanced strains [34] considering a three-field Petrov-Galerkin mixed formulation by assuming that displacements $\mathbf{u}$ and strains $\boldsymbol{\varepsilon}$ in (9) are

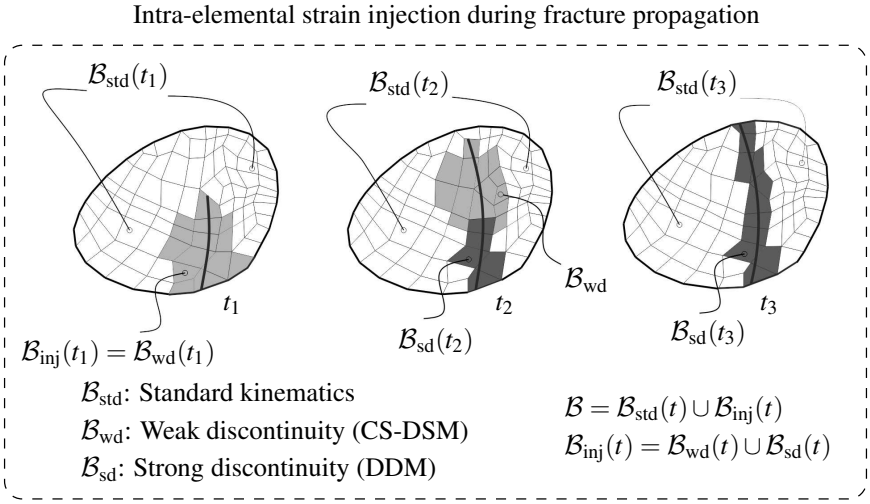


Figure 2: Subdomain categories of the discrete body along different times (t_1, t_2 and t_3) of the analysis.

independent fields. In both CS-DSM and DDM modes the strain field is partitioned in two terms: $\boldsymbol{\varepsilon} = \boldsymbol{\xi} + \boldsymbol{\gamma}$ denoting a compatible (and smooth) strain field $\boldsymbol{\xi}$ and an enhanced strain field $\boldsymbol{\gamma}$ which tackles possible singularities in the failure propagation kinematics.

The resulting Embedded Finite Element method (E-FEM) has been already presented in [23] and [24] and assessed for the case of quasi-static failure propagation. In the following subsections a brief description of its formulation is given but the reader is referred to the works in [18, 23, 24] for complete implementation details.

2.1.1 Weak discontinuity injection in $\mathcal{B}_{\text{wd}}$. Constant Stress-Discontinuous Strain Mode (CS-DSM)

The weak discontinuity kinematics is injected when a user defined strain threshold is exceeded (cf. [18]). Such critical strain threshold is determined based on a strain-like internal variable of the damage model, r, which records the maximum historical value of the equivalent strain, τ_ε, accounting for the positive strain counterpart, i.e. only tensile stress states contribute to the strain norm. For complete details on the k-regularized damage model employed in this study the reader is referred to the works in [8, 18, 20].

Considering the strain field decomposition into a compatible and enhanced strain, the resulting variational three-field problem can be stated as: Find $\mathbf{u} \in \mathcal{V}_{\text{u}}^{\text{wd}}$ and $\boldsymbol{\varepsilon} = \boldsymbol{\xi} + \boldsymbol{\gamma}$ with $\boldsymbol{\xi} \in \mathcal{V}_{\varepsilon}^{\text{wd}}$ and $\boldsymbol{\gamma} \in \mathcal{V}_{\gamma}^{\text{wd}}$, satisfying:

$$\int_{\mathcal{B}} \boldsymbol{\sigma}(\boldsymbol{\varepsilon}) : \boldsymbol{\nabla}\boldsymbol{\eta}\, d\mathcal{B} - \underbrace{\int_{\mathcal{B}} (\mathbf{b} - \rho\ddot{\mathbf{u}}) \cdot \boldsymbol{\eta}\, d\mathcal{B} - \int_{\partial\mathcal{B}_{\text{t}}} \mathbf{t}^* \cdot \boldsymbol{\eta}\, d\Gamma}_{\mathbf{W}_{\eta}^{\text{ext}}(\boldsymbol{\eta},\mathbf{b},\ddot{\mathbf{u}},\mathbf{t}^*)} = 0; \qquad \forall \boldsymbol{\eta} \in \hat{\mathcal{V}}_{\text{u}}^{\text{wd}}, \tag{10}$$

$$\int_{\mathcal{B}^e} \phi^e \hat{\boldsymbol{\xi}}^e : \left(\phi^e \boldsymbol{\xi}^e - \sum_{i=1}^{n^e_{\text{node}}} (\boldsymbol{\nabla}^{\text{s}} N_i^e \otimes \mathbf{d}_i^e) \right) d\mathcal{B} = 0; \qquad \forall \hat{\boldsymbol{\xi}}^e \in \hat{\mathcal{V}}_{\varepsilon}^{\text{wd}}, \tag{11}$$

$$\int_{\mathcal{B}^e} \hat{\boldsymbol{\gamma}}^e : (\boldsymbol{\chi}_S^{(h^e,k^e)} \boldsymbol{\sigma})\, d\mathcal{B} = \int_{\mathcal{S}^e} \hat{\boldsymbol{\gamma}}^e : [\![\boldsymbol{\sigma}]\!]\, d\mathcal{B} = 0; \qquad \forall \hat{\boldsymbol{\gamma}}^e \in \hat{\mathcal{V}}_{\gamma}^{\text{wd}}, \tag{12}$$

where $\mathcal{V}_{\square}^{\text{wd}}$ and $\hat{\mathcal{V}}_{\square}^{\text{wd}}$ denote the trial and test function spaces, respectively. The subindices u, ε and γ refer to the spaces of displacements $\mathbf{u}$, strains $\boldsymbol{\varepsilon}$ and enhanced strain fields $\boldsymbol{\gamma}$, respectively (cf. [18] for a detailed definition of these function spaces). Displacements $\mathbf{u}$ and displacement variations $\boldsymbol{\eta}$ are element-wise interpolated using standard bilinear shape functions while the compatible and enhanced strain components are interpolated using the spatially constant function ϕ and the dipole function χ, detailed in [18, 24]. The term denoted $\mathbf{W}_{\eta}^{\text{ext}}$ in (10), refers to the virtual work due to the external and inertial forces.

As described in [18] the non-directional character of the strain enhancement renders the CS-DSM element an excellent candidate to be employed at regions undergoing complex fracture phenomena such as crack branching and intersection. It is designed to sense strain localization phenomena and, therefore, is seen particularly useful to anticipate a strong discontinuity at the crack tip region in a certain well-captured propagation direction.

2.1.2 Strong discontinuity injection in $\mathcal{B}_{\text{sd}}$. Discontinuous Displacement Mode (DDM)

Upon increasing strain localization in a particular direction, the strong discontinuity kinematics is injected after the the injection of the weak discontinuity. The element to be enriched with the strong discontinuity kinematics belongs to $\mathcal{B}_{\text{inj}}$ for which the strain threshold referred in Section 2.1.1 must have been exceeded. Additionally, all elements in $\mathcal{B}_{\text{sd}}$ must fulfill the following conditions (cf. [18] for a more detailed explanation):

1. The constitutive tangent satisfies the bifurcation condition, i.e. the determinant of the corresponding acoustic tensor must vanish.

2. The local dissipation per unit of surface exceeds a user-defined fraction of the material fracture energy G_{f}

3. The trace of one strong discontinuity $\mathcal{S}$ (cf. Section 2.2) intersects the element.

When the above mentioned conditions are accomplished, the strong discontinuity kinematics can be injected with confidence propagating the crack following a correct direction and avoiding possible stress locking effects arising from any kinematic incompatibilities.

Considering the strain field decomposition into a compatible and enhanced strain field, as proposed in the assumed enhanced strains methodology, the mixed variational problem in $\mathcal{B}_{\text{sd}}$ can be written as: find $\mathbf{u} \in \mathcal{V}_{\text{u}}^{\text{sd}}$ and $\boldsymbol{\varepsilon}_{\text{sd}} = \boldsymbol{\xi}_{\text{sd}} + \boldsymbol{\gamma}_{\text{sd}}$ with $\boldsymbol{\xi}_{\text{sd}} \in \mathcal{V}_{\varepsilon}^{\text{sd}}$ and $\boldsymbol{\gamma}_{\text{sd}} \in \mathcal{V}_{\gamma}^{\text{sd}}$, satisfying:

$$\int_{\mathcal{B}} \boldsymbol{\sigma}(\boldsymbol{\varepsilon}) : \nabla^{\text{s}} \boldsymbol{\eta} \, d\mathcal{B} - \mathbf{W}_{\eta}^{\text{ext}}(\boldsymbol{\eta}, \mathbf{b}, \ddot{\mathbf{u}}, \mathbf{t}^{*}) = 0; \qquad \forall \boldsymbol{\eta} \in \hat{\mathcal{V}}_{\text{u}}^{\text{sd}}, \tag{13}$$

$$\int_{\mathcal{B}^e} \chi_S^{+}(\mathbf{x}) \, (\llbracket \hat{\mathbf{u}} \rrbracket^e \otimes^{\text{s}} \mathbf{n} : \boldsymbol{\sigma}) \, d\mathcal{B} = \int_{\mathcal{S}^e} \llbracket \hat{\mathbf{u}} \rrbracket^e \cdot \llbracket \boldsymbol{\sigma} \cdot \mathbf{n} \rrbracket \, d\Gamma = 0; \qquad \forall \llbracket \hat{\mathbf{u}} \rrbracket^e \in \mathbb{R}^{n_{\text{dim}}}, \tag{14}$$

$$\int_{\mathcal{B}^e} \phi^e \hat{\boldsymbol{\xi}}_{\text{sd}}^{\,e} : \left(\phi^e \boldsymbol{\xi}_{\text{sd}}^{\,e} - \left[\nabla^{\text{s}} N_i^e \otimes \mathbf{d}_i^e) - \nabla \boldsymbol{\varphi}^e \otimes^{\text{s}} \llbracket \mathbf{u} \rrbracket^e \right] \right) d\mathcal{B} = 0; \qquad \forall \hat{\boldsymbol{\xi}}^{\,e} \in \mathbb{S}^{n_{\text{dim}} \times n_{\text{dim}}}, \tag{15}$$

where $\mathcal{V}_{\square}^{\text{sd}}$ and $\hat{\mathcal{V}}_{\square}^{\text{sd}}$ denote the trial and test function spaces, respectively. The subindices u, ε and γ refer to the spaces of displacements $\mathbf{u}$, strains $\boldsymbol{\varepsilon}$ and enhanced strain fields $\boldsymbol{\gamma}$, respectively (The reader is referred to the work in [18] for a detailed definition of these function spaces).

Displacements variations $\boldsymbol{\eta}$ are element-wise interpolated using standard bilinear shape functions while the displacements $\mathbf{u}$ are interpolated as an addition of a standard interpolation of the smooth term and a term accounting for the displacement jump $[\![\mathbf{u}]\!]$ through the Heaviside step function and an element auxiliary function described in [18, 24]. The compatible strain and compatible strain variations are interpolated using the spatially constant function ϕ. The enhanced strains are element-wise interpolated employing the dirac delta shifted to $\mathcal{S}$ and the enhanced strain variations are computed through the generalized dipole-like function χ_S^+ detailed in [18, 24].

2.2 Crack Path Field tracking algorithm

The trace of the strong discontinuity is computed through the crack path field (CPF) technique introduced by Oliver et al. [23]. Essentially, a directional maximum of a conveniently chosen localizing field r (cf. [18, 23, 24]) is calculated defining the so-called crack path set Γ, i. e. predicting the trace of the strong discontinuity $\mathcal{S}$.

The direction in which the maximum of r is identified is taken according to the vector field

$$\hat{\mathbf{e}}(\mathbf{x},t) = \frac{\boldsymbol{\nabla}\tilde{u}(\mathbf{x},t)}{||\boldsymbol{\nabla}\tilde{u}||} \tag{16}$$

where $\tilde{u}$ denotes a scalar value of the displacement field $\mathbf{u}$ (cf. [18]). The orientation of the strong discontinuity $\mathbf{n}$ is directly taken from the vector $\hat{\mathbf{e}}$. The directional maximum of r, identifying the crack path set Γ, is computed through the zero-level set of the so called crack path field $\mu(\mathbf{x},t)$, defined as:

$$\mu(\mathbf{x},t) = \frac{\partial \tilde{r}}{\partial e} = \boldsymbol{\nabla}\tilde{r}\cdot\hat{\mathbf{e}}, \tag{17}$$

where $\tilde{r}$ is a sufficiently smooth field of r. The zero-level set of μ defines the crack path set Γ as

$$\Gamma(t) := \{\mathbf{x} \quad \big| \mu(\mathbf{x},t) = \frac{\partial \tilde{r}}{\partial e} = 0\}. \tag{18}$$

Note that $\mathcal{S}$ is contained in the crack path set Γ (cf. Figure 3) and is identified as the portion of Γ belonging to $\mathcal{B}_{\text{sd}}$. Once the intersections between $\mathcal{S}$ and the finite element boundaries are known, the auxiliary function φ^{e} can be constructed.

The spirit of the CPF technique is to facilitate the computation of the trace of the strong discontinuity specially in those cases in which crack propagation may be biased by mesh alignment or structured discretizations (cf. [23]).

2.3 Implicit time integration

The enhanced degrees of freedom ($\bar{\boldsymbol{\xi}}^e$ and $\bar{\boldsymbol{\gamma}}^e$ in (10)–(11), or $\bar{\boldsymbol{\xi}}^e$ and $[\![\mathbf{u}]\!]^e$ in (13)–(15)) are condensed out at the element level and, therefore, the unknowns in the discrete model consist on the vector of smooth displacements $\mathbf{u}$ and accelerations $\ddot{\mathbf{u}}$. Consequently, the elements with enhanced kinematics do not lead to an increase of the standard degrees of freedom. Since the enhanced degrees of freedom do not have an associated mass, their condensation is performed as done in quasi-static problems.

The resulting global system of equations in terms of $\mathbf{u}$ for each time step can be written in a matrix form as:

$$\mathbf{M}\ddot{\mathbf{u}}(t) + \mathbf{F}_{\text{int}}(\mathbf{u}(t)) - \mathbf{F}_{\text{ext}}(t) = \mathbf{0}, \tag{19}$$

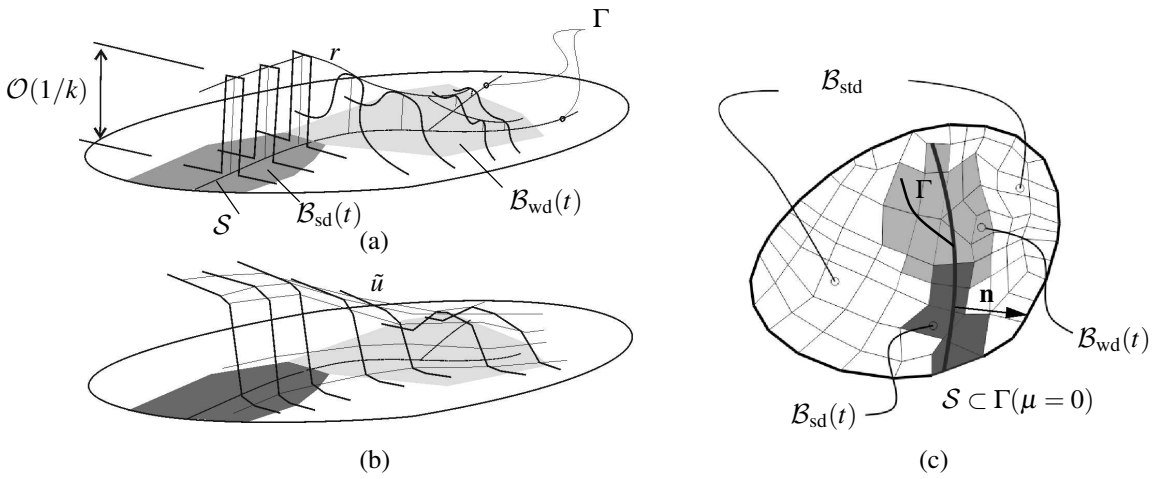


Figure 3: Crack Path Field (CPF) strategy for the crack tracking algorithm. Γ denotes the crack path set. (a) Localized field r around a strong discontinuity band. (b) Associated scalar displacement field $\tilde{u}$. (c) Γ (zero-level set of the crack path field function $\tilde{\mu}$) and the trace of the strong discontinuity $\mathcal{S}$.

where $\mathbf{M}$ and $\mathbf{F}_{\text{ext}}$ represent the mass matrix and external force vector, without the inertial forces, respectively. The mass matrix is computed in a standard way, i.e. lumping the contributions to a diagonal matrix, and is only associated to the degrees of freedom of the smooth displacement field. The internal force vector and $\mathbf{F}_{\text{int}}$ results from the evaluation of the first term, on the l.h.s. of (9), (10) or (13), corresponding to $\mathcal{B}_{\text{std}}$, $\mathcal{B}_{\text{wd}}$ and $\mathcal{B}_{\text{sd}}$, respectively.

A standard implicit time integration (Newmark) scheme is employed to find the solution of (19) at time step $t+1$. Considering the solution $\mathbf{u}^{t+1}$ the system of equations in (19) can be expressed as

$$\mathbf{R}(\mathbf{u}^{t+1}) = \mathbf{0}. \tag{20}$$

Due to the evolution of crack phenomena, the integration domains vary with time, as well as the spatial integration rules and the residual force vector has to be evaluated in an incremental way as follows:

$$\mathbf{R}(\mathbf{u}^{t+1}) = \mathbf{R}(\mathbf{u}^{t}) + \Delta\mathbf{R}\Big(\Delta\mathbf{u}^{t+1}, \mathcal{B}^{t}_{\text{std}}, \mathcal{B}^{t}_{\text{wd}}, \mathcal{B}^{t}_{\text{sd}}\Big). \tag{21}$$

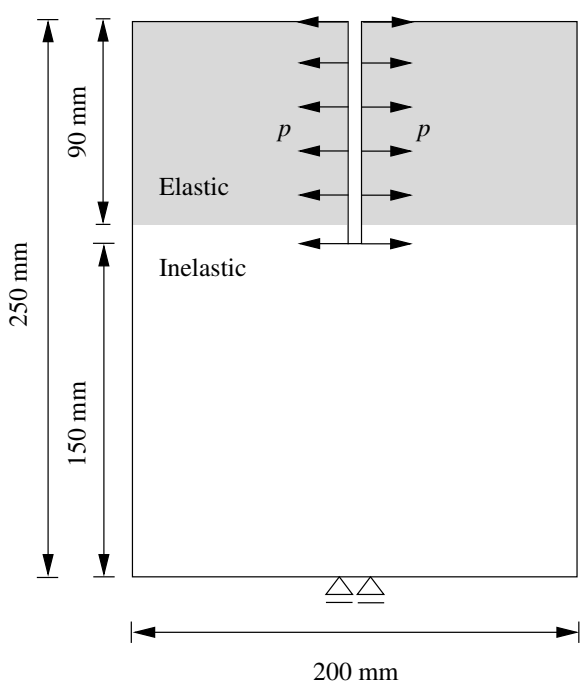
The integration rules for the evaluation of the residual force vector, $\mathbf{R}$, are specific for each domain $\mathcal{B}^{t+1}_{\text{std}}$, $\mathcal{B}^{t+1}_{\text{wd}}$ and $\mathcal{B}^{t+1}_{\text{sd}}$. An alternative approach is utilized in our implementation which consists in redefining the stresses, which take place in the internal force expressions such that one quadrature rule suffices for evaluating $\mathbf{R}(\mathbf{u}^{t+1})$ in the complete domain $\mathcal{B}$ (cf. Oliver et al. [23, 24] for details of the effective stress definition and their updating scheme).

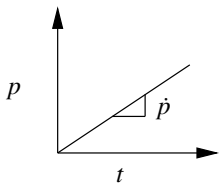
3 Representative simulations

The following examples illustrate the performance of the crack path field and strain injection techniques in dynamic simulations with different loading rates and material brittleness. Plane strain conditions are assumed in all two-dimensional examples.

3.1 Dynamic simulations of quasi-brittle fracture at different loading rates

Dynamic fracture simulations of a compact tension (CT) test are performed considering a quasi-brittle material such as concrete (cf. Figure 4). The influence of the loading regime is studied by varying the pressure rates applied at the loading walls of the CT specimen. The damage model described in [17, 24] is employed with a linear softening law and an only tension





	$\dot{p}\left[\frac{N}{mm^2 \cdot s}\right]$
Low loading rate	1.077×10^3
Medium loading rate	$7.180 - 10.000 \times 10^3$
High loading rate	$13.067 - 21.538 \times 10^3$

Figure 4: Geometry and boundary conditions for the dynamic CT test.

$\sigma_u\ [N/mm^2]$	E $[N/mm^2]$	ν	$G_f\ [N/mm]$	$\rho\ [Kg/m^3]$
3.5	30.0×10^3	0.18	0.09	2400

Table 1: Material parameters for the dynamic compact tension (CT) test.

failure criterion. A constant time discretization is considered with $\Delta t = 10^{-6}$ s and the material parameters are shown in Table 1.

In order to prevent damage nucleation at the vicinity of the wall where the pressure load is applied, the upper part of the specimen (cf. Figure 4) is kept elastic. The Rayleigh wave speed v_R for a concrete material is considered approximately equal to 2100 m/s (cf. [25]).

Upon increasing loading rate, the failure pattern changes from a vertical mode-I crack to mixed mode with multiple cracks, i.e. the fracture surface increases (cf. Figure 5). This tendency is also observed in Ožbolt et al. [25] where a similar test is studied. The injection of weak and strong discontinuities in combination with the Crack Path Field (CPF) tracking algorithm automatically captures branching phenomena without the need for a special strong discontinuity kinematics at the branched element as proposed in [16]. In the proposed methodology, the crack path sets cannot intersect itself and, therefore, the elements at the branching region are injected with the CS-DSM which has no assumed direction of localization.

The total energy dissipation and the dissipation plots are shown for different loading rates $\dot{p}$ ranging 1.077×10^3 to $2.154 \times 10^4\ N/mm^2/s$ in Figure 6. Such dissipated energy $W^{\mathcal{D}}$ at time t is calculated as:

$$W^{\mathcal{D}}(t) = \int_0^t P_{\text{ext}}(t)\, dt - \Psi(t) - \mathcal{K}(t), \tag{22}$$

where $P_{\text{ext}}(t)$ denote the external mechanical power, performed by the external loads excluding the inertial forces, $\Psi(t)$ the total internal energy and $\mathcal{K}(t)$ the kinetic energy, at time t. Addi-

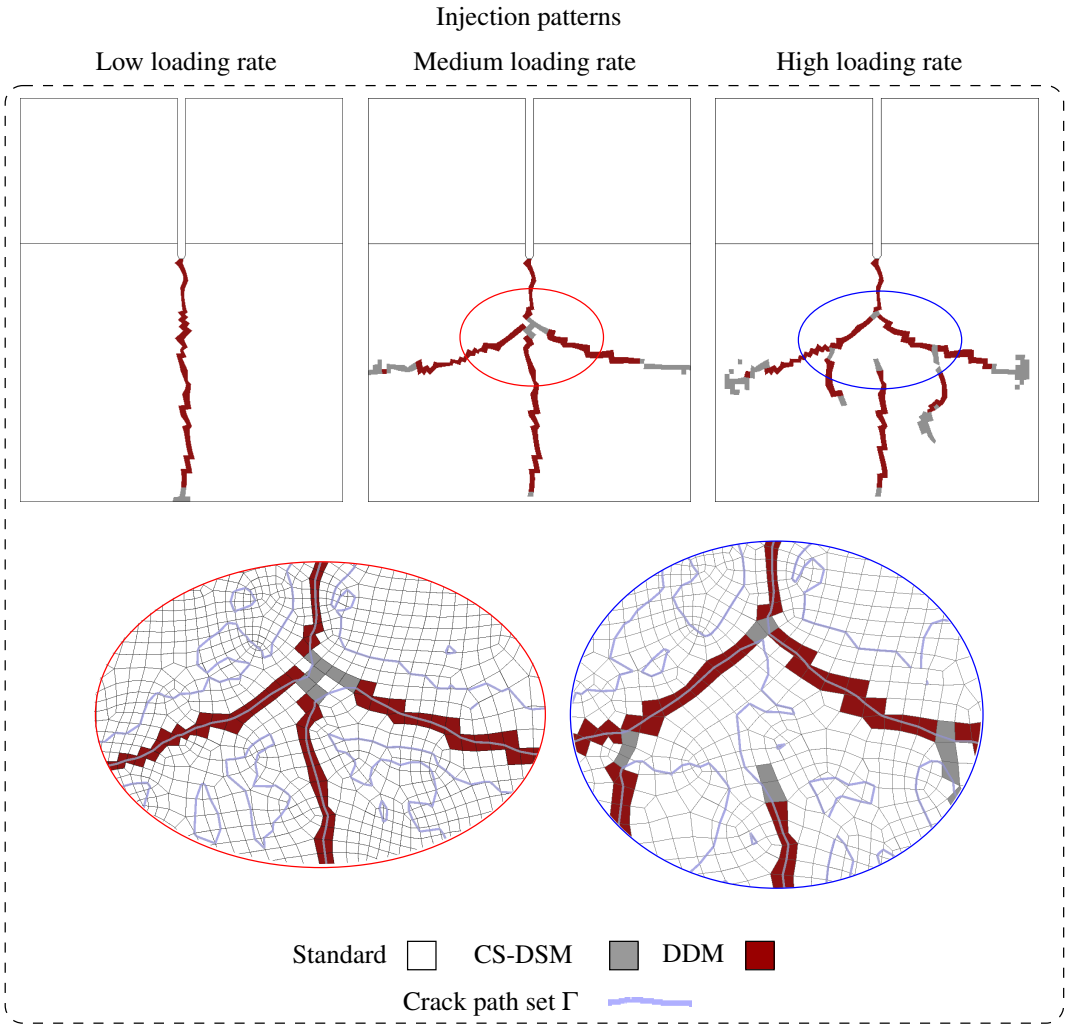


Figure 5: Injection patterns upon increasing loading rate at the dynamic compact tension test.

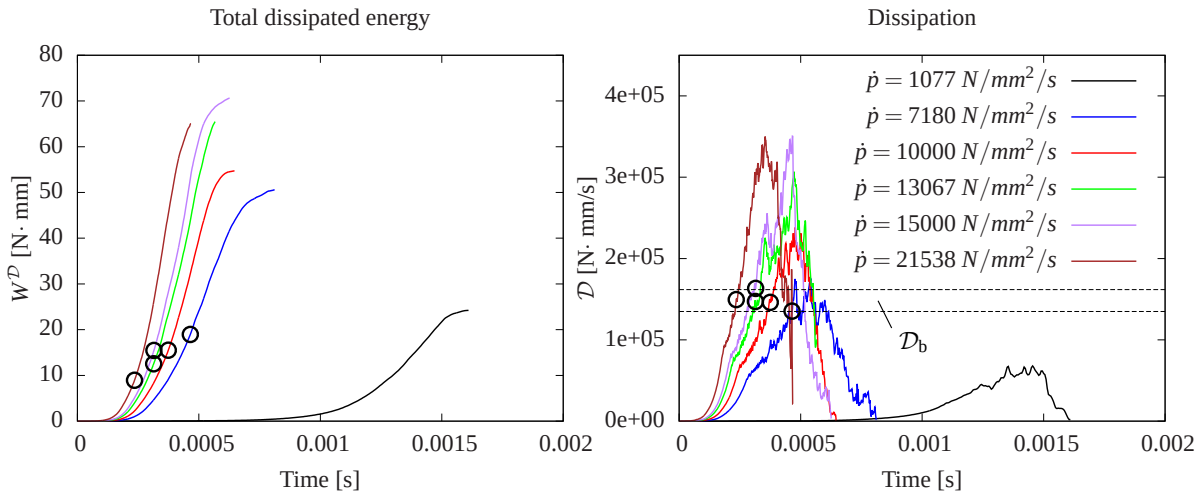


Figure 6: Time evolution of the total dissipated energy (left) and dissipation (right) for different loading rates. The black circles indicate the first observed branching episode. A thickness of 1mm is considered for the energy plots.

tionally, the dissipation $\mathcal{D}(t)$ can be computed as the derivative with respect to time of $W^{\mathcal{D}}(t)$ since

$$W^{\mathcal{D}}(t) = \int_0^{\chi=t} \mathcal{D}(\chi)\, d\chi. \tag{23}$$

The total dissipated energy and crack surface increase upon increasing loading and the corresponding dissipation presents a higher peak value (cf. Figure 6). The branching episodes are indicated in the plots with black circles at each of the curves. It is observed that all dissipations measured during branching events cluster in a region between 1.35×10^5 and 1.62×10^5 $N \cdot mm/s$. These limits $\mathcal{D}_\mathrm{b}$ can be estimated as

$$\mathcal{D}_\mathrm{b} = v_\mathrm{b} G_\mathrm{f} k, \tag{24}$$

where v_b refers to the crack velocity at the onset of branching reported in experiments, i.e. between 500 and 600 m/s for concrete material [25], G_f stands for the material fracture energy and k denotes the considered bandwidth of the localization zone in the DDM (cf. [18, 23]).

The expression in (24) implicitly sets a maximum dissipation for a single crack considering that the energy is instantaneously dissipated. In this view, all registered dissipations greater than the estimated upper limit in (24) necessarily involve more than one crack. In all results shown in Figure 6 the dissipations registered upon the onset of branching fit reasonably well between the above mentioned limits computed with the experimental maximum crack velocities 500 and 600 m/s.

For this reason, the expression in (24) can be seen as a methodology to infer the crack velocity at the onset of branching knowing the total dissipation assuming that the energy is instantaneously released during crack propagation. Considering the upper and lower registered dissipation limits at the onset of branching, the corresponding limiting velocities are 501.82 and 545.58 m/s according to (24). This alternative approach is based on the global dissipation values field $\mathcal{D}$ which is remarkably smoother and reliably evaluated than the crack tip velocity field.

In [11], crack branching is detected when a critical crack surface rate is reached which is analogous to the dissipation criteria. However, the advantage of monitoring the rate of dissipated energy is that it can be performed by accounting for a global energy balance instead of

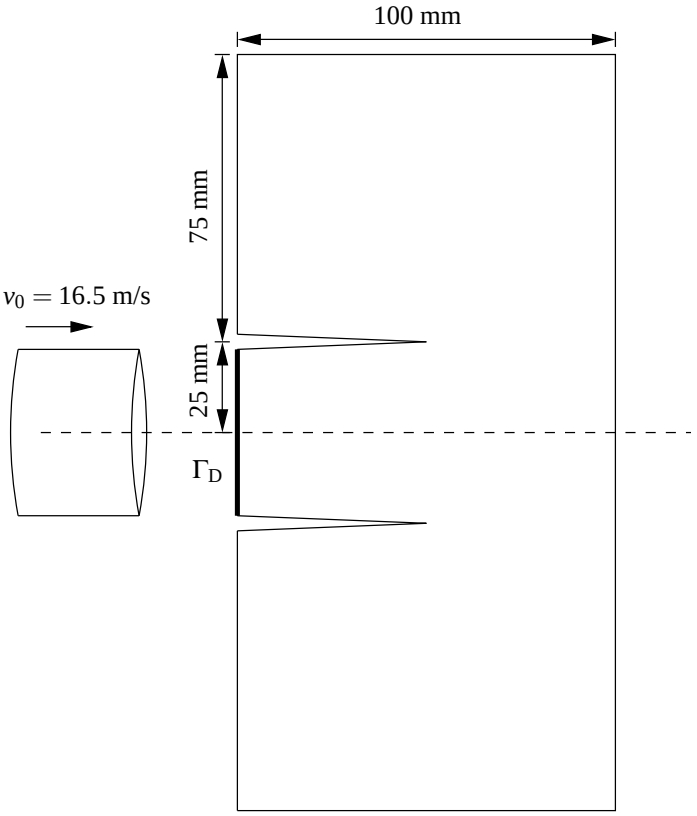


Figure 7: Geometry and boundary conditions for the Kalthoff experiment.

$\sigma_u\ [N/mm^2]$	E $[N/mm^2]$	ν	$G_f\ [N/mm]$	$\rho\ [Kg/m^3]$
844.0	190.0×10^3	0.3	22.17	8000

Table 2: Material parameters and wave velocities for the Kalthoff experiment.

locally studying crack surface growth. In other words, the total energy dissipation is computed as the difference between the external loads energy minus the deformation and kinetic energies. In this way, no additional model-dependent criteria are needed to locate the crack tip position.

3.2 Towards simulations of brittle failure involving complex fracture phenomena

The strain injection and crack path field techniques are employed to reproduce the Kalthoff experiment [12] (cf. Figure 7) consisting of an edged-cracked plate impacted by a projectile. It is reported experimentally that, at an impact velocity $v_0 = 16.5\ m/s$, a mode I crack propagates from the notch towards the superior and inferior specimen edges at an angle of proximately 70 degrees.

An applied velocity at the prescribed contour Γ_D mimics the boundary condition resulting from the impact. The test is imposed on a metallic plate with material parameters summarized in Table 2. Both unstructured and structured meshes have been studied with approximately 15000 elements and a time step discretization $\Delta t = 10^{-7}$ s is adopted.

Injection patterns for both meshes shown in Figure 8 reveal that both results are very much comparable and in agreement with the experimental results which report crack propagation an-

gles close to 70 degrees. The patterns are remarkably symmetric as expected from the imposed boundary conditions and are found in agreement with those provided by different numerical techniques, e.g. phase-field [11] and gradient damage models [15] but at the expense of a mesh discretization which is around two orders of magnitude finer (cf. Figure 9).

4 Conclusions

A FE formulation is presented which is capable of tackling fracture dynamics problems through the injection of weak and strong displacement discontinuities. Since the kinematic enhancement simulating the fracture is performed inside the element (intra-elemental) and not represented by bandwidth of several elements (supra-elemental), the necessary FE discretization can be adopted significantly coarser than strategies such as phase field or gradient damage models.

The crack path field global tracking technique in combination with the strong discontinuity injection procedure automatically account for complex fracture phenomena encountered upon increasing loading rates such as branching. Crack branching regions are injected with the non-directional CS-DSM element while the rest of the trace of discontinuity is modeled with the strong discontinuity injection.

Assuming that the fracture energy is instantly released, the maximum dissipation for one propagating crack can be estimated as the product of the fracture energy, the maximum crack tip velocity and the crack band width. In this view, the crack branching velocities can be estimated by monitoring the dissipation at the onset of branching $\mathcal{D}_{\mathrm{b}}$ and have been found to be in good agreement with crack tip velocities at the onset of branching reported experimentally.

The simulation of the Kalthoff experiment with both structured and unstructured meshes provides results in agreement with other supra-elemental numerical schemes but employing a finite element discretization which can be up to two orders of magnitude coarser.

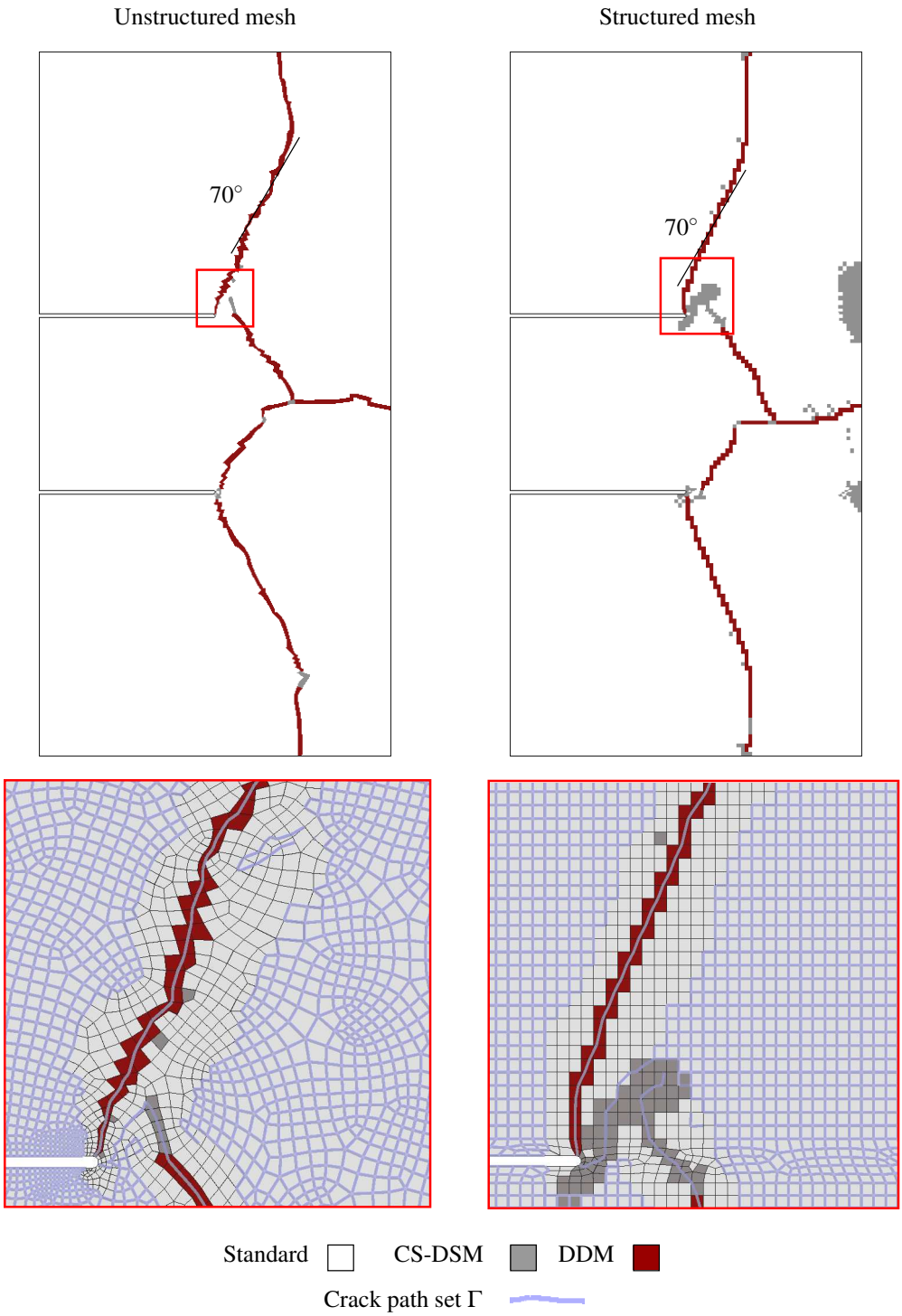


Figure 8: Injection patterns for the unstructured and structured meshes.

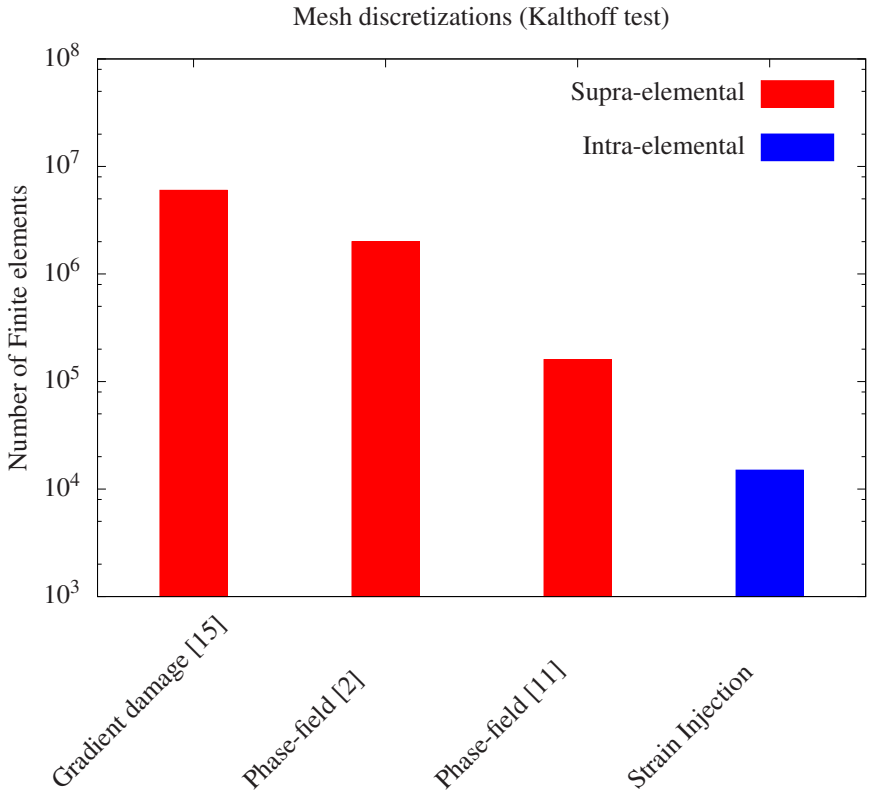


Figure 9: Equivalent mesh discretizations for the Kalthoff experiment employing different numerical techniques.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 320815 (ERC Advanced Grant Project "Advanced tools for computational design of engineering materials" COMP-DES-MAT). Oriol Lloberas-Valls gratefully acknowledges the funding received from the Spanish Ministry of Economy and Competitiveness through the "Juan de la Cierva" Postdoctoral Junior Grant: JCI-2012-13782 and the National Research Plan 2014: MAT2014-60919-R. Ivo Dias gratefully acknowledges the financial support from Laboratório Nacional de Engenharia Civil (LNEC) through the postdoctoral research Grant (CoMatFail project).

References

[1] T. Belytschko, D. Organ, and C. Gerlach. Element-free galerkin methods for dynamic fracture in concrete. *Computer Methods in Applied Mechanics and Engineering*, 187(3–4):385–399, 2000.

[2] M. J. Borden, C. V. Verhoosel, M. A. Scott, T. J. R. Hughes, and C. M. Landis. A phase-field description of dynamic brittle fracture. *Computer Methods in Applied Mechanics and Engineering*, 217220:77 – 95, 2012.

[3] G. T. Camacho and M. Ortiz. Computational modelling of impact damage in brittle materials. *International Journal of Solids and Structures*, 33(2022):2899–2938, 1996.

[4] I. F. Dias, J. Oliver, and A. E. Huespe. *Strain Injection Techniques in Numerical Modeling*

of Propagating Material Failure. Monograph. International Center for Numerical Methods in Engineering, 2012. ISBN 978-84-940243-7-5.

[5] I. F. Dias, J. Oliver, J. V. Lemos, and O. Lloberas-Valls. Advanced numerical techniques for modeling tensile crack propagation in gravity dams. In *Second International Dam World Conference*. LNEC, Lisbon, Portugal, 2015.

[6] I. F. Dias, J. Oliver, J. V. Lemos, and O. Lloberas-Valls. Modeling tensile crack propagation in concrete gravity dams via crack-path-field and strain injection techniques. *Engineering Fracture Mechanics*, 2016. In press.

[7] M. L. Falk, A. Needleman, and J. R. Rice. A critical evaluation of cohesive zone models of dynamic fractur. *Le Journal de Physique IV*, 11(PR5):Pr5–43, 2001.

[8] R. Faria, J. Oliver, and M. Cervera. A strain-based plastic viscous-damage model for massive concrete structures. *International Journal of Solids and Structures*, 35(14):1533–1558, 1998.

[9] L. B. Freund. *Dynamic Fracture Mechanics*. Cambridge University Press, 1998.

[10] Y. Ha and F. Bobaru. Studies of dynamic crack propagation and crack branching with peridynamics. *International Journal of Fracture*, 162(1–2):229–244, 2010.

[11] M. Hofacker and C. Miehe. A phase field model of dynamic fracture: Robust field updates for the analysis of complex crack patterns. *International Journal for Numerical Methods in Engineering*, 93(3):276–301, 2013.

[12] J. Kalthoff. Modes of dynamic shear failure in solids. *International Journal of Fracture*, 101(1-2):1–31, 2000.

[13] L. Kosteski, R.B DAmbra, and I. Iturrioz. Crack propagation in elastic solids using the truss-like discrete element method. *International journal of fracture*, 174(2):139–161, 2012.

[14] B. Li, A. Pandolfi, and M. Ortiz. Material-point erosion simulation of dynamic fragmentation of metals. *Mechanics of Materials*, 80, Part B:288 – 297, 2015.

[15] T. LI, J.-J. Marigo, D. Guilbaud, and S. Potapov. Gradient damage modeling of brittle fracture in an explicit dynamics context. Submitted. hal-01248263, 2015.

[16] C. Linder and F. Armero. Finite elements with embedded branching. *Finite Elements in Analysis and Design*, 45(4):280 – 293, 2009.

[17] D. L. Linero. *A model of material failure for reinforced concrete via Continuum Strong Discontinuity Approach and mixing theory*. PhD thesis, E.T.S. Enginyers de Camins, Canals i Ports, Technical University of Catalonia (UPC), Barcelona, 2006. Cimne Monograph Nbr.M106.

[18] O. Lloberas-Valls, A. E. Huespe, J. Oliver, and I. F. Dias. Strain injection techniques in dynamic fracture modeling. *Computer Methods in Applied Mechanics and Engineering*, 2016. Submitted.

[19] C. Miehe, M. Hofacker, and F. Welschinger. A phase field model for rate-independent crack propagation: Robust algorithmic implementation based on operator splits. *Computer Methods in Applied Mechanics and Engineering*, 199(4548):2765–2778, 2010.

[20] J. Oliver. Continuum modelling of strong discontinuities in solid mechanics using damage models. *Computational Mechanics*, 17(1-2):49–61, 1995.

[21] J. Oliver. Modelling strong discontinuities in solids mechanics via strain softening constitutive equations. part 1: Fundamentals. *International Journal for Numerical Methods in Engineering*, 39(21):3575–3600, 1996.

[22] J. Oliver, A. E. Huespe, M. D. G. Pulido, and E. Chaves. From continuum mechanics to fracture mechanics: the strong discontinuity approach. *Engineering Fracture Mechanics*, 69:113–136, 2002.

[23] J. Oliver, I. F. Dias, and A. E. Huespe. Crack-path field and strain-injection techniques in computational modeling of propagating material failure. *Computer Methods in Applied Mechanics and Engineering*, 274:289–348, 2014.

[24] J. Oliver, M. Caicedo, E. Roubin, A. E. Huespe, and J. A. Hernández. Continuum approach to computational multiscale modeling of propagating fracture. *Computer Methods in Applied Mechanics and Engineering*, 294:384 – 427, 2015.

[25] J. Ožbolt, A. Sharma, and H.-W. Reinhardt. Dynamic fracture of concrete – compact tension specimen. *International Journal of Solids and Structures*, 48(10):1534 – 1543, 2011.

[26] A. Pandolfi, P. Krysl, and M. Ortiz. Finite element simulation of ring expansion and fragmentation: the capturing of length and time scales through cohesive models of fracture. *International Journal of Fracture*, 95(1–4):279–297, 1999.

[27] R. R. Pedersen, A. Simone, and L. J. Sluys. An analysis of dynamic fracture in concrete with a continuum visco-elastic visco-plastic damage model. *Engineering Fracture Mechanics*, 75(13):3782–3805, 2008.

[28] R. R. Pedersen, A. Simone, and L. J. Sluys. Mesoscopic modeling and simulation of the dynamic tensile behavior of concrete. *Cement and Concrete Research*, 50:74–87, 2013.

[29] M. Ramulu and A. S. Kobayashi. Mechanics of crack curving and branching a dynamic fracture analysis. *International Journal of Fracture*, 27(3-4):187–201, 1985.

[30] E. Sharon and J. Fineberg. Confirming the continuum theory of dynamic brittle fracture for fast cracks. *Nature*, 397(6717):333–335, 1999.

[31] E. Sharon, S. P. Gross, and J. Fineberg. Local crack branching as a mechanism for instability in dynamic fracture. *Physical Review Letters*, 74(25):5096–5099, 1995.

[32] S.A. Silling. Dynamic fracture modeling with a meshfree peridynamic code. In K. J. Bathe, editor, *Computational Fluid and Solid Mechanics 2003*, pages 641–644. Elsevier Science Ltd, Oxford, 2003.

[33] J. C. Simo and J. Oliver. A new approach to the analysis and simulation of strain softening in solids. *Fracture and damage in quasibrittle structures*, pages 25–39, 1994.

[34] J. C. Simo and M. S. Rifai. A class of mixed assumed strain methods and the method of incompatible modes. *International Journal for Numerical Methods in Engineering*, 29(8): 1595–1638, 1990.

[35] J.-H. Song and T. Belytschko. Cracking node method for dynamic fracture with finite elements. *International Journal for Numerical Methods in Engineering*, 77(3):360–385, 2009.

[36] J.-H. Song, H. Wang, and T. Belytschko. A comparative study on finite element methods for dynamic fracture. *Computational Mechanics*, 42(2):239–250, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

HYBRID FE/XFE FINITE ELEMENT MODEL FOR SIMULATION OF BRITTLE-DUCTILE FRACTURES IN DUAL-PHASE STEEL GRADES.

K. Perzyński and L. Madej

AGH University of Science and Technology Mickiewicza 30 av., 30-059, Krakow, Poland.
e-mail: kperzyns@agh.edu.pl

Keywords: XFEM, FEM, brittle fracture, ductile fracture, digital material representation

Abstract. *Development of hybrid finite element/eXtended finite element numerical solution combined with the digital material representation (DMR) approach to simulate brittle/ductile fracture is the overall subject of the realized research. This particular work is divided into three main parts. In the first, proposed procedures for development of DMR model of investigated two phase steel are described. Then, details on the procedure of creation of representative digital material representation model to simulate interrelations between ductile and brittle fractures occurring within particular phases are presented. Finally, based on evaluated model parameters examples of its application to numerical simulation of failure in DP steels are presented.*

The main outcome from this work is the hybrid FEM/XFEM model, which can be easily applied to various deformation processes. The approach fulfils industrial demand to create a computer aided design of processes were fracture is an issue. Presented model is designed for practical research and industrial applications.

1 INTRODUCTION

Modelling fracture in dual phase (DP) steels is a complex task because of the composite character of the investigated microstructure. There are two phases with significantly different mechanical properties: soft ferrite as a matrix and hard constituent phase as inclusions. Recently modelling of fracture initiation and propagation based on the digital material representation (DMR) approach is becoming popular and different models, which take into account mainly ductile fracture with various fracture criteria in DP steels are being developed [1, 2].

That is why, the main aim of the approach proposed by the Authors is to create a robust model of both brittle and ductile failure in DP steels based on modern numerical approaches that take microstructure explicitly into account during simulation of deformation.

Authors developed a combined model of ductile and brittle fracture that occur in ferrite and martensite, respectively. Ductile fracture is modelled by the Ductile Fracture criterion implemented within classical Finite Element (FE) model, while brittle one is predicted by more sophisticated eXtended Finite Element Method (XFEM). Proper data transfer protocols between these two methods were proposed to create a hybrid numerical model.

2 DIGITAL MATERIAL REPRESENTATION

The concept of digital material representation was suggested approx. a decade ago and is dynamically evolving [3, 4, 5, 6]. The main objective of the DMR is to create digital representation of microstructure with its morphological features (i.e. grains, grain orientations, inclusions, cracks, different phases etc.) represented explicitly. Generation of material microstructure with specific geometrical features and properties is one of the most important algorithmic parts of systems based on the DMR. Such DMR is further used in numerical simulations of material processing. Recently a lot of research is put on the development of methods responsible for creation of the 2D and 3D representations of analysed microstructures.

To obtain an accurate description of the 2D microstructure an image processing methods are usually applied. As an input data for this analysis SEM/EBSD results can be used. In this case not only information on microstructure geometry is obtained but also information on initial crystallographic orientation is provided. This approach was successfully used in [7, 8]. Unfortunately the approach is time consuming and expensive because each numerical simulation based on DMR requires a SEM/EBSD analysis. That is why image processing is also applied to the optical microscopy images that are more affordable. However, in this case only information regarding grain morphology is obtained see e.g. [9]. Details of such procedure applied to the DP steel can be found in [10]. The approach is generally composed of the three main stages responsible for image filtering (denoising), phases distinction and phases analysis. These subsequently applied methods aim to detect phases, phase boundaries, and, finally, to prepare the material microstructure in the DMR form for further numerical simulations (Figure 1).

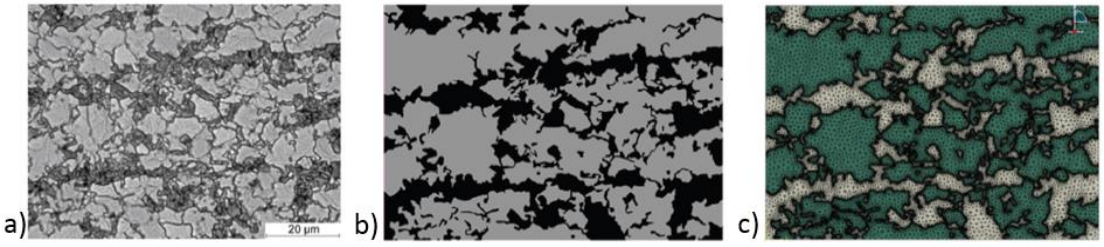


Figure 1: a) Real microstructure b) binary representation c) digital material representation with finite element discretization of dual phase steel [6].

However, the crucial aspect that has to be addressed, prior the numerical simulation, is evaluation of representativeness aspects of such digital models.

3 SELECTION OF THE REPRESENTATIVE VOLUME ELEMENT OF DUAL PHASE STEELS

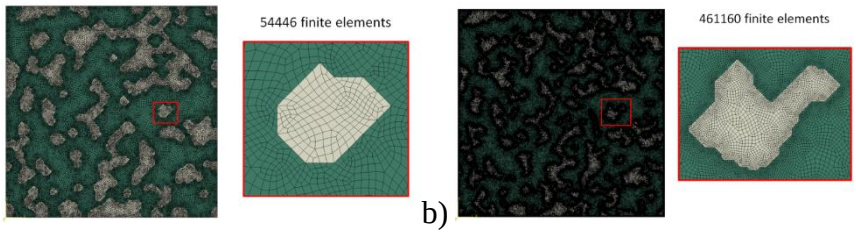
In the literature [11, 12] there are two different approaches to the digital material representation. The DMR model can be considered as:

- the Representative Volume Element (RVE) – and in this sense it is a model of the material that can be used to determine the corresponding effective properties for the homogenized macroscopic model,
- the Unit Cell (UC) – in this sense it is a part of the RVE that enables obtaining results for particular part of the material. Thus, the Unit Cell is not representative for the whole numerical model

Consequently, before numerical investigation based on the RVE DMR approach a specific minimum size of the DMR model that can be considered as the RVE should be established. Research on this subject is presented within the chapter. However, aspects related with the mesh sensitivity have to be investigated first to eliminate numerical errors affecting final solution.

3.1 Mesh sensitivity analysis

The finite element method is mesh sensitive and obtained results depend to some extend on quality of the finite element mesh used for calculation. This is especially important for DMR models as local solutions are of particular interest. Thus, seven DMR models of the same microstructure ($100 \times 100 \mu m$) with different levels of the discretization were analysed to evaluate this issue. For all models the same sets of material properties for the martensite and ferrite phases, respectively, were used. Periodic boundary conditions were applied to maintain continuity of the computational domain. Number of finite elements used for discretization was set from 54 446 to 461 160. Four node bilinear plane strain quadrilateral reduced integration, hourglass control quad finite elements (CPE4R) were applied during discretization. DMR models were incorporated into the commercial ABAQUS application and calculated using the implicit solver. Selected microstructures used for mesh sensitivity analysis are presented in Figure 2.



a) b)

Figure 2: Microstructures discretized from a) 54 446 to b) 461 160 finite elements.

For low quality mesh, geometrical features of the martensitic phase are not replicated properly. All models were subjected to tensile test with displacement equal 20%. Examples of obtained results are presented in Figure 3. The figure confirms that mesh density has influence on obtained results. When mesh density is higher than 122 000 elements no visible differences in material behavior between models are visible. After the threshold value, obtained flow stress data converge to the same curve. So it can be stated that 224 000 finite elements per 100 μm^2 have to be used during the calculation to minimize the effect of mesh on quality of obtained results. Thus, this minimal value of FE *elements*/μm^2 is used in subsequent calculations related

to evaluation of minimal size of the DMR model that can be considered as representative for the entire macro sample.

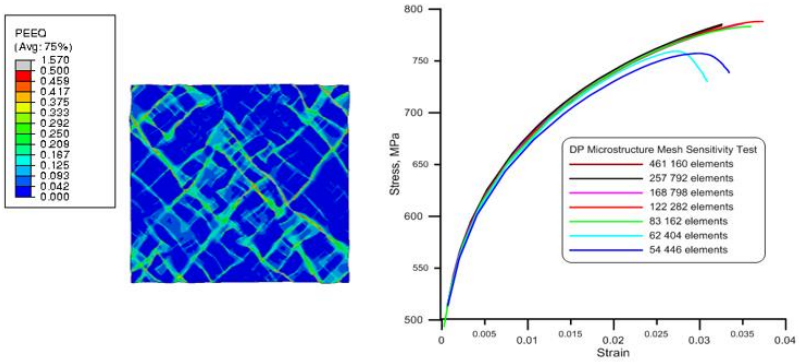


Figure 3: Homogenized strain–stress curves for 7 investigated microstructures.

3.2 Influence of DMR model size

To evaluate the minimal size of the DMR model that can be considered as RVE, four digital microstructure models were developed with the sizes: 50×50, 100×100, 200×200 and $300 \times 300 \mu m$. Authors decided to extract three smaller models from the largest microstructure in order to receive the same martensite volume fraction but different geometrical sizes as seen in Figure 4.

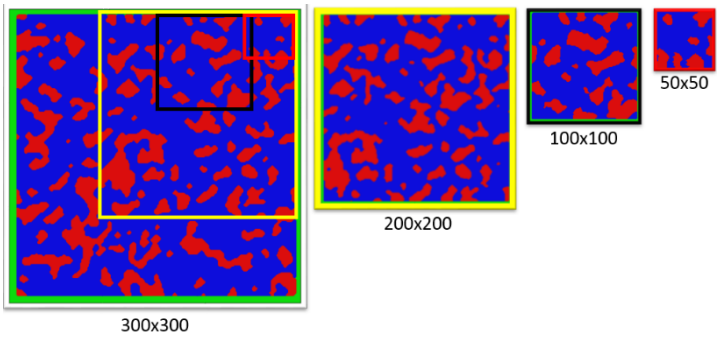


Figure 4: Digital material representation models with the same amount of martensite and different sizes: $50 \times 50, 100 \times 100, 200 \times 200$ and $300 \times 300\ \mu m$.

Digital models were incorporated into the finite element software. Periodic boundary conditions were applied and digital microstructures were deformed approx. $30 - 40\%$ of engineering strain. Number of finite elements used for discretization was set above the evaluated threshold value and varied between 190 000 to 335 000 nodes depending on microstructure complexity. Four node bilinear plane strain quadrilateral reduce integration, hourglass control quad finite elements (CPE4R) were used for the discretization purposes.

Because representative macroscopic response of the DMR models is under investigation, particular attention is put on homogenized flow curves obtained during deformation. Obtained homogenized stress–strain curves are shown in Figure 5.

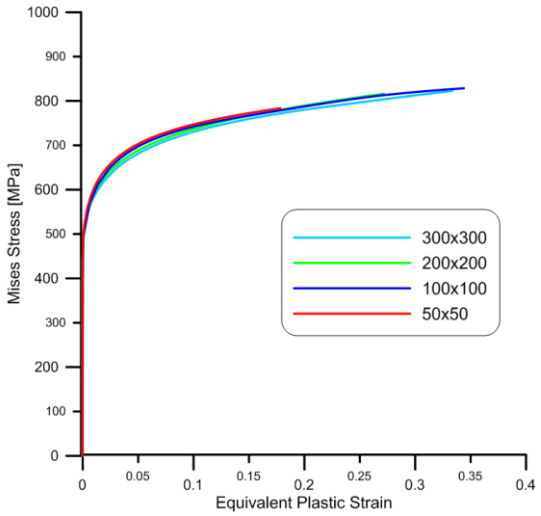


Figure 5: Homogenized stress–strain curves for investigated DMR models.

As seen in Figure 5 all microstructures provide quite similar material response to the applied deformation. Only the smallest microstructure $50 \times 50\ \mu m$ did not converge to the rest of the results due to the large local deformations in some regions – finite elements reached large degradation, which ended the simulation with solver error. Remaining models in a global sense behave consistently despite the investigated size. Additionally, to evaluate local material response strain distributions across the microstructure were analysed as seen in Figure 6.

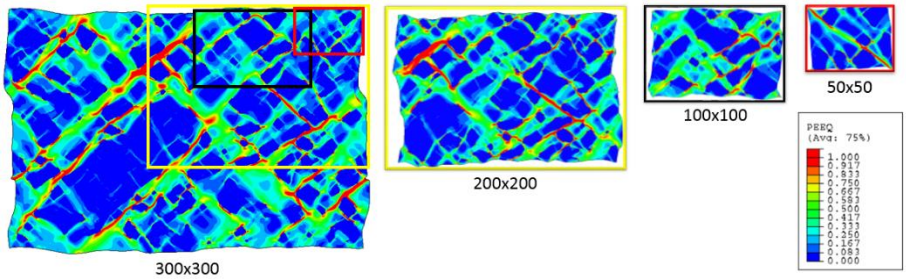


Figure 6: Equivalent strain distribution received in all DMR's after 30% of deformation.

As seen in Figure 6, maximum strain have similar values in all investigated microstructures. This information is crucial when the problem of cracks initiation in material is under consideration. Thus, it can be concluded that for the calculation accuracy microstructure larger than $100 \times 100 \mu m$ should be used in further calculations.

These recommended DMR model parameters are then used during the numerical simulation realized with the developed model.

4 BRITTLE-DUCTILE FRACTURE MODEL

As mentioned, modelling of brittle-ductile fractures based on presented DMR approach requires combination of two numerical methods. In present work brittle martensite fracture was resolved with XFEM method implemented in the ABAQUS package. The most important model parameters are: martensite fracture initiation criterion based on the critical stress parameter σ_c and fracture propagation criterion based on critical energy condition (E_c). Model parameters

were taken from literature investigation presented in [13] and were 1370 MPa, and 17.39 $\frac{J}{m^2}$, respectively.

Ferrite phase fracture during large plastic deformation fails by ductile failure mechanisms. Thus, a ductile failure criterion where the equivalent plastic strain at the onset of damage, ε_{iD}^{p} is a function of stress triaxiality and strain rate [14] was used:

$$\varepsilon_{iD}^{p}\left(\eta,\dot{\varepsilon}_{i}^{p}\right)=\frac{\varepsilon_{i}^{+}\,sinh[k_0(\eta^{-}+\eta)]+\varepsilon_{i}^{-}sinh[k_0(\eta-\eta^{+})]}{sinh[k_0(\eta^{-}+\eta^{+})]} \tag{1}$$

where: ε_{i}^{+} and ε_{i}^{-} – equivalent plastic strain for equibiaxial tensile and equibiaxial compressive deformation respectively, η – stress triaxiality (a ratio of the equivalent mean stress σ_m to the Misses equivalent stress σ_i), k_0 – parameter obtained experimentally. These parameters depend on the material, strain rate and temperature of the process. Finally, failure occurs when state variable w_D reaches 1:

$$w_D=\int\frac{d\varepsilon_{iD}^{p}}{\varepsilon_{iD}^{p}\left(\eta,\dot{\varepsilon}_{i}^{p}\right)}=1 \tag{2}$$

The fracture initiation parameter for ferrite failure was adopted from literature [15].

Data transfer between the two models was realized with the developed python script. Example of results obtained from numerical simulation of deformation, under tensile stress state, based on developed model is shown in Figure 7.

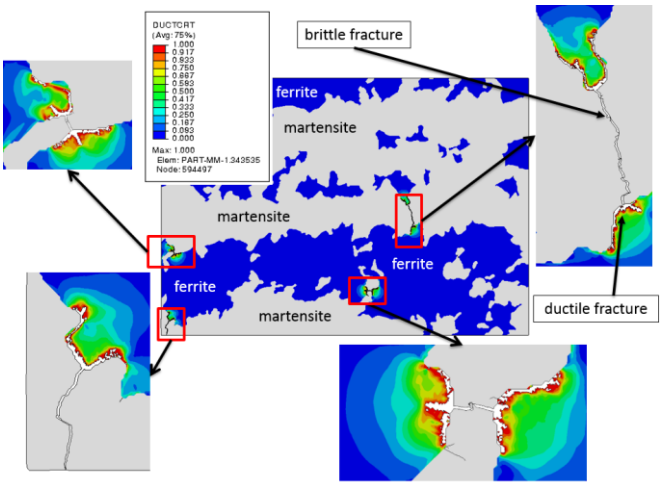


Figure 7: Brittle-ductile fractures propagating across martensite and ferrite.

5 CONCLUSIONS

- DMR model allow to take into account microstructural inhomogeneities, which have large influence on damage behaviour in DP steel.
- The representativeness of the DP DMR models is not strongly related to the size of the model, but mainly depends on proper replication of martensite volume fraction.
- Minimum number of finite elements used for the discretization of the DP DMR models should be approximately 224 000 FE elements peer 100 μm^2 to minimize mesh sensitivity effect during calculations.
- The use of the proposed method for modelling fracture in DP steels provides the possibility to take into account all relevant mechanisms of fracture: brittle fracture in

martensite phases, decohesion between martensite-ferrite phases and ductile failure in ferrite.

6 ACKNOWLEDGEMENTS.

Financial assistance provided by the National Science Centre under the 2014/14/E/ST8/00332 project is gratefully acknowledged. Numerical calculations were performed at the ACK Cyfronet: MNiSW/IBM_BC_HS21/AGH/076/2010.

REFERENCES

[1] V. Uthaisangsuk, U. Prahl, W. Bleck, Failure modeling of multiphase steels using representative volume elements based on real microstructures, *Procedia Engineering,* **1,** 171-176, 2009.

[2] V. Uthaisangsuk, U. Prahl, W. Bleck, Modelling of damage and failure in multiphase high strength DP and TRIP steels, *Engineering Fracture Mechanics,* **78**, 469-486, 2011.

[3] C. F. Cornwell, R. W. Noack, E. J. Abed, *Three-dimensional digital microstructures, project report documentation prepared by High Performance Technologies*, Inc Aberdeen Proving Ground, MD 21005, 2006.

[4] M. Bernacki, Y. Chastel, H. Digonnet, H. Resk, T. Coupez, R. Logé, Development of numerical tools for the multiscale modelling of recrystallisation in metals, based on a digital material framework, *Computer Methods in Material Science,* **7**, 142-149, 2007.

[5] R. Logé, M. Bernacki, H. Resk, L. Delannay, H. Digonnet, Y. Chastel,T. Coupez, Linking plastic deformation to recrystallization in metals, using digital microstructures, *Philosophical Magazine,* **88**, 3691-3712, 2008.

[6] L. Madej, Realistic description of dual phase steel morphology on the basis of the Monte Carlo method, *Computer Methods in Matrials Science,* **12**, 197-206, 2012.

[7] D. Raabe, R. Becker, Recrystalization simulation by coupling of a crystal plasticity FEM with a cellular automaton method, *Modelling and Simulation in Materials Science and Engineering*, **8**, 445-462, 2010.

[8] R. Logé, H. Resk, Z. Sun, L. Delannay, M. Bernacki, Modelling of plastic deformation and recrystalization of polycrystals using digital microstructures and adaptive meshing techniques, *Steel Research International*, **81**, 1420-1425, 2010.

[9] A. Milenin, P. Kustra, The multi scale FEM simulation of wire fracture phenomena during drawing of Mg alloy, *Steel Research International*, **79**, 717-722, 2008.

[10] L. Rauch, L. Madej, Application of the automatic image processing in modelling of the deformation mechanisms based on the digital representation of microstructure, *International Journal for Multiscale Computational Engineering,* **8**, 343-356, 2010.

[11] Z. Hashin, Analysis of composite materials – a survey, *Journal of Applied Mechanics,* **50**, 481-505, 1983.

[12] M. Serafi, W. Cecot, Numerical aspects of computational homogenization, *Computer Methods in Materials Science,* **13**, 213-218, 2013.

[13] A. Ramazania, A. Schwedt, A. Aretz, U. Prahl, W. Bleck, Characterization and modelling of failure initiation in DP steel, *Computational Materials Science,* **75**, 35-44, 2013.

[14] H. Hooputra, H. Gese, H. Dell, H. Werner, A comprehensive failure model for crashworthiness simulation of aluminium extrusions, *International Journal of Crashworthiness*, **9**, 449-464, 2004.

[15] N. Vajragupta, V. Uthaisangsuk, B. Schmaling, S. Münstermann, A. Hartmaier, W. Bleck, A micromechanical damage simulation of dual phase steels using XFEM, *Computational Materials Science,* **54**, 271-279, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EFFECTIVE MESO AND MACRO PROPERTIES FOR FIBRE-REINFORCED-POLYMER CURING COUPLED TO VISCO-ELASTICITY

R. Mahnken[1] **and C. Dammann**[2]

[1],[2]Chair of Engineering Mechanics (LTM)
University of Paderborn
Warburger Str. 100, 33098 Paderborn, Germany
[1]e-mail: mahnken@ltm.upb.de
[2]e-mail: dammann@ltm.upb.de

Keywords: Curing, shrinkage, thermal-mechanical-chemical-coupling, homogenization, RVE.

Abstract. *Our work presents a three-scale model for temperature-dependent visco-elastic effects accompanied by curing, which are important phenomena in a resin transfer molding (RTM) process. The effective bulk quantities in dependence on the degree of cure are obtained by homogenization for a representative unit cell (micro-RVE) on the heterogeneous microscale. To this end, an analytic solution is derived by extension of the well known composite spheres model. Voigt and Reuss bounds resulting from the assumption of a homogeneous matrix are compared to the effective quantities. During curing and subsequent mechanical loading, the periodic mesostructure defined by a visco-elastic polymeric matrix and linear-thermo-elastic carbon fibres is taken into account as a representative unit cell (meso-RVE) subjected to thermo-mechanical loading on the mesoscale. Homogenization of the mesoscale by volume averaging is applied to obtain the effective properties for the fully cured composite on the macroscale, e.g. the macroscopic anisotropic thermal expansion coefficient. In the examples we simulate the curing process as well as mechanical loading of the cured part with the finite- element-method.*

1 INTRODUCTION

In these days, applications for polymeric materials are found in carbon- and glass fibre-reinforced composite materials, epoxy laminates and (nano-) particle-reinforced polymer structures, cf. RUIZ and TROCHU [1] as well as LANGE [2]. An important production process of fibre-reinforced polymeric (FRP) materials is resin transfer molding (RTM). It is mainly characterized by curing, where the matrix in the initial uncured state, a mixture of resin and curing agent, exhibits a viscous liquid behavior until gelation. The process is highly temperature dependent and influences strongly the mechanical, thermal and chemical properties.

Due to increasing demand, over the last years considerable effort was made to develop three-dimensional constitutive models that account for a time- or degree of cure dependence, respectively, of the mechanical properties. Several suggestions for simulation of polymer curing on the basis of a physically and chemically sound approach have been presented by ADOLF and co-workers, see e.g. [3]. Furthermore, both geometrically linear [4, 5, 6] and nonlinear [7, 3, 8] constitutive models have been proposed by ADOLF *et al.* and HOSSAIN *et al.*, where the latter suggest a thermodynamically consistent framework for the simulation of curing polymers independent of the choice of the free energy density. In this way any phenomenological or micromechanical approach can be utilized. The influence of the fibre/matrix RVE size for a similar model is investigated in HEINRICH *et al.* [9].

In LION and HÖFER [8] a phenomenological thermo-visco-elastic curing model for finite strain deformations is proposed. The formulation is based on process dependent viscosities as in the previous works of HAUPT and LION [10, 11] within a thermodynamic framework. It accounts for thermally and chemically induced volume changes using a ternary multiplicative split of the deformation gradient into mechanical, thermal and chemical parts. Furthermore, a *coordinate of reaction* is introduced representing the degree of cure, and which is taken as in internal variable into the Helmholtz free energy. A advanced application of the finite element method for curing processes is presented in KLINGE *et al.* [12], in order to consider the aspect of microheterogeneity of the curing process by use of a multiscale finite element method.

MAHNKEN and DAMMANN [13] present a three-scale framework for fibre-reinforced polymer curing. It employs a meso-RVE consisting of the fibre and the matrix part and a micro-RVE for the latter consisting of the three components resin, curing agent and solidified material. Microscopic modeling and a comparative study render different *meso bulk material properties* including bounds (for the assumption of a homogeneous mixture for the matrix components) and effective material constants (for the assumption of a heterogeneous arrangement for the matrix components). Based on the preliminaries in [13], in MAHNKEN and DAMMANN [14] a periodic mesostructure is adapted. It is defined by a visco-elastic polymeric matrix and linear-thermo-elastic carbon fibres as a representative unit cell (meso-RVE) subjected to thermo-mechanical loading on the mesoscale. Basic relations in a thermodynamic framework which is specialized to the scenario of a mechanical-thermal-chemical process for polymeric materials are presented. To this end, the effective mesoscopic properties from [13] are applied to the prototype model in [15] resulting in the thermodynamically consistent model in [14].

In this work, we briefly summarize the three-scale framework for FRP curing as well as the *meso bulk properties* for the matrix according to the homogeneous and heterogeneous matrix approaches from [13]. We also outline the meso-model as a result from [14] in combination with the meso to macro transition by volume averaging. Moreover, we investigate the thermo-chemical curing process and subsequent mechanical loading in order to get the *effective bulk properties* for the fully cured composite on the macroscale.

The main aspects and the organization of this paper are summarized as follows: Section 2 introduces the preliminaries for two reaction mechanisms for thermosetting polymers and presents a three-scale framework for fibre reinforced polymer curing. Furthermore, we introduce the meso-RVE consisting of a matrix and a fibre part. Three components of the polymeric matrix are defined and

thus, mass phase fractions are introduced. Based on these preliminaries, two conceptions are distinguished in Section 3. Firstly, an equally distributed mixture on the microscale is assumed, secondly, an arrangement of a heterogeneous microstructure is taken into account. In Section 4 basic relations for the mesoscale are introduced. The constitutive relations of the prototype model including the homogeneous as well as the heterogeneous ansatz for the matrix properties are summarized followed by a comment on the fibre model as a special case. Preliminaries for the computational homogenization scheme which is used for the meso to macro transition are introduced in Section 5. Section 6 presents two numerical examples. In the first example, we illustrate the mechanical-thermo-chemical coupling for a meso-RVE in a three-dimensional finite-element simulation in Subsection 6.1. Some features on the coupling of temperature, curing and visco-elasticity of the proposed model are illustrated for a mesoscopic problem and its homogenized macroscopic response occurring in a curing process. We compare the meso-micro relations for the homogeneous matrix which yield (upper) Voigt and (lower) Reuss bounds with those derived for the heterogeneous matrix which yield effective parameters. Based on the curing process in Subsection 6.1, in the second example in Subsection 6.2 we determine the effective macroscopic anisotropic stiffness tensor and the effective macroscopic anisotropic thermal expansion coefficient for the fully cured fibre reinforced polymer.

Notations

Square brackets $[\bullet]$ are used throughout the paper to denote 'function of' in order to distinguish from mathematical groupings with parenthesis $(\bullet)$.

2 PRELIMINARIES FOR THE CURING OF THERMOSETTING POLYMERS

2.1 Polymerization mechanisms

Epoxy resin systems are included in the group of thermosetting polymers (thermosets). Starting with the initial uncured state the matrix (a mixture of resin and curing agent) undergoes a polymerization process during curing. As introduced by FLORY [16] two polymerization reaction mechanisms are distinguished: 1. by free-radicals or ions (*chain-growth polymerization*) and 2. by functional groups (*step-growth polymerization*).

Considering the scheme of *chain-growth polymerization* shown in Figure 1.a starting with an initiator the unsaturated monomer molecules add onto the active site of a growing polymer chain one

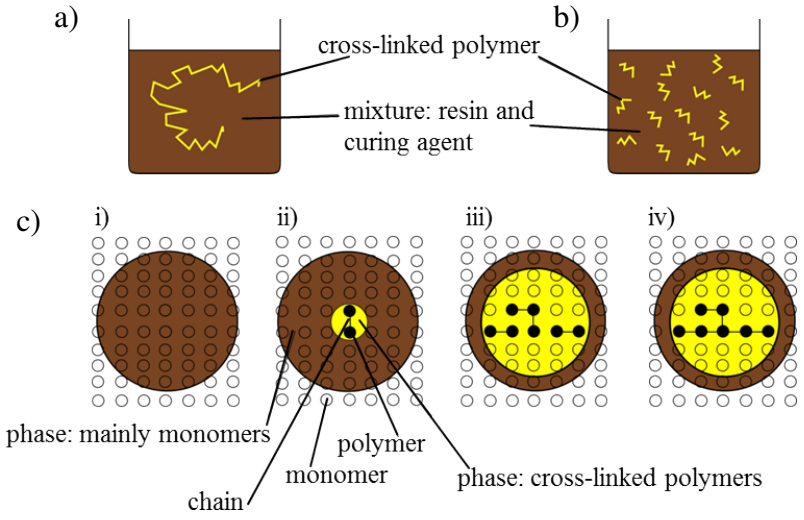


Figure 1: Polymerization in thermosets: a) schematic illustration of chain- and b) step-growth polymerization b) generic representation of a step-growth polymerization adapted to [17] i) uncured state, ii) two monomers react (monomer as predominant species), iii) state with larger chain length including dimer and trimer, iv) state where branching started.

at a time. Different steps operate at different stages of the mechanism (i.e. initiation, propagation, termination, and chain transfer). Following COWIE and ARRIGHI [17], the characteristics of *step-growth polymerization* are the growth throughout the matrix, whereas no initiator is necessary and the reaction between any two functional groups of monomers is stepwise. Thus, at multiple locations and at the same time similar steps are repeated throughout the reaction process as illustrated in Figure 1.b. In particular initiation, propagation and termination reactions are essentially identical in rate and mechanism as shown Figure 1.c. As described in STILLE [18], the chain length increases steadily and random growth takes place as the monomer reacts with both monomer or polymer species with equal ease until a high molecular weight polymer is obtained.

Taking into account both reactions mechanism is the decisive point for two conceptions for the matrix (the *homogeneous matrix* and the *heterogeneous matrix*) we distinguish later on. Especially for the *step-growth polymerization* we assume an idealization: For the locations from where on random growth starts we assume a probabilistic distribution. The random growth is then represented by multiple growing spherical inclusions with its origins at those multiple locations from where on the molecular functional groups of monomers start to react, i.e. the phase of cross-linked monomers in Figure 1.c at one of the multiple locations in Figure 1.b. This region is surrounded by a phase including mainly monomers.

2.2 A three-scale framework for fibre reinforced polymer curing

Based on the polymerization mechanisms described in the previous subsection Figure 2.a gives an overview of the three-scale framework. $\bar{\Pi} \subset \mathcal{R}^3$ denotes the homogeneous macrostructure. It is associated to the periodic mesosstructure $\Pi \subset \mathcal{R}^3$ of a layered fibre reinforced polymer consisting of the matrix part and the fibre part $\Pi = \Pi^M \cup \Pi^F$, respectively, each regarded as homogeneous solid constituents. The mesoscopic representative volume element (meso-RVE) Ω is associated to Π with the decomposition $\Omega = \Omega^M \cup \Omega^F$. We denote by $\boldsymbol{x}^J \subset \Pi^J$ the position vector of a material point $P^J \in \Pi^J$ for which either $J = M$ or $J = F$ holds at time t in the space-time-domain $\Pi^J \times]-\infty, T[$. Here T is the total time of interest. The microstructure $\pi \subset \mathcal{R}^3$ is related to the matrix part Π^M

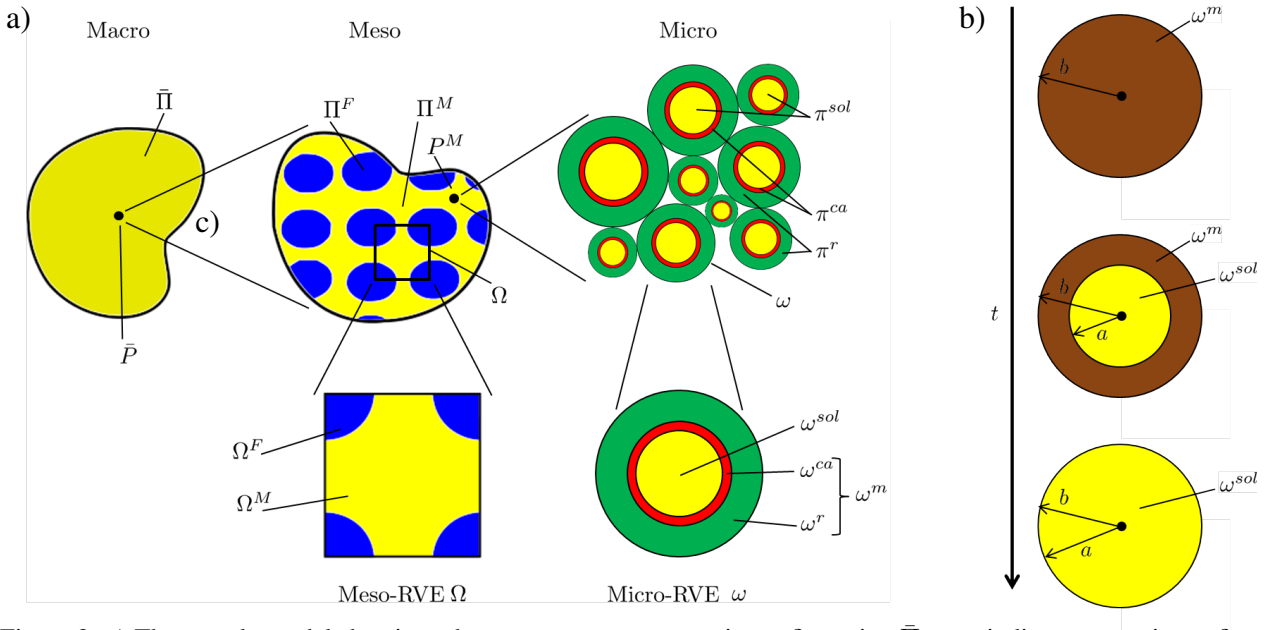


Figure 2: a) Three-scale model showing a homogeneous macroscopic configuration $\bar{\Pi}$, a periodic mesoscopic configuration Π consisting of homogeneous parts Π^F and Π^M, a corresponding meso-RVE Ω and a microscopic configuration as a multiphase system with phases π^{ca}, π^r and π^{sol} and a corresponding micro-RVE ω. b) Micro-RVE as a a two phase composite sphere over time: top) initial uncured, middle) partially cured and bottom) fully cured state.

at every point P^M at time t. $\pi = \pi^{ca} \cup \pi^r \cup \pi^{sol}$ consists of the phases π^j, $j = ca, r, sol$ (curing agent, resin and solidified material), respectively. The representative volume element (micro-RVE) ω is associated to the structure π with a decomposition of the micro-RVE in terms of three components $\omega = \omega^{ca} \cup \omega^r \cup \omega^{sol}$, or for simplification in terms of monomer and fully cured material $\omega = \omega^m \cup \omega^{sol}$. An illustration of such a two-phase system during curing over the time t is given in Figure 2.b.

In the sequel, for notational benefits, we omit the index J, as meso-micro relations are derived only for the behavior of Π^M. Thus, a quantity $\bullet^M$ related to the matrix reads $\bullet$.

2.3 Mass fractions

The matrix Π consists of three phases π^{ca}, π^r and π^{sol}. Following [19], with the time-dependent variables $dm^r[t]$, $dm^{ca}[t]$ and $dm^{sol}[t]$ for the masses of resin, curing agent and solidified material, respectively, the conservation of mass during the curing reaction requires that the masses of all constituents sum up to the total mass of the mixture dm_0. Using this relation, the mass phase fractions of resin, curing agent and solid, ζ^r, ζ^{ca}, ζ^{sol} in Table 1, are introduced such that the sum of all fractions is one.

3 MESOSCOPIC PROPERTIES

Following [13], in order to obtain mesoscopic material properties two different conceptions are presented:

- Homogeneous matrix: An equally distributed mixture on the microscale π is assumed (meso-model hom) for the three constituents which yields bounds for the meso bulk properties.
- Heterogeneous matrix: An arrangement of a heterogeneous microstructure π as shown in Figure 2.a (meso-model het) is taken into account. The microstructure is represented by the two-phase composite sphere as micro-RVE in Figure 2.b with radii a and b which are related to its constituents ω^m and ω^{sol}, respectively.

As a simplification, we assume linear thermo-chemo-elasticity for the microscale π and consequently introduce the heat- and curing-dilatation coefficients α^j and β^j as well as the shear- and compression moduli G^j and K^j for the phases ζ^j, $j = m, sol$. The material parameters are summed up in Table 1, where in addition the degree of cure z is simultaneously defined for the

- Homogeneous matrix: In dependence on the mass phase fractions for the monomer and the solid $\zeta^m[t] = \zeta^r[t] + \zeta^{ca}[t]$, $\zeta^{sol}[t]$.
- Heterogeneous matrix: In dependence on the radii a, b in accordance to the volumes ω^m and ω^{sol} in Figure 2.b.

The justification for this differing formulations is that in [13] for each constituent $m,\ sol$ the authors introduce volume phase fractions as well as mass phase fractions (cf. Table 1) and shown that they almost equal.

The equally distributed mixture yields upper (Voigt) and lower (Reuss) bounds, consequently in total three meso-models $homV,\ homR$ and het which lead to different evolution of the material properties in dependence on the degree of cure z for the mesoscale Π are distinguished in Table 1. The resulting meso-micro relations are summarized for the

- Homogeneous matrix: This meso-model denoted as hom is divided into two sub-models, the
  - Reuss model denoted as $homR$ which yields lower bounds for the meso bulk properties. In particular, we state the *bulk compression modulus* K_R in Table 1, Eq. (5) and the *bulk heat-dilatation coefficient* α_R in Table 1, Eq. (6) in dependence on the degree of cure as well as

the shrinking-dilatation coefficient $\beta(est)$ in Table 1, Eq. (2) which is not dependent on curing. The properties are in agreement with ad hoc assumptions from the literature, i.e. the *bulk curing-dilatation coefficient* corresponds to the derivations by [1], [15] and [20], whereas the *bulk compression modulus* and the *bulk heat-dilatation coefficient* coincide with the well known lower Reuss bounds. Thus, we combine these formulations for the volumetric quantities to a model which we denote as the Reuss model.

– Voigt model denoted as $homV$ which yields upper bounds for the meso bulk properties, in particular the bulk compression modulus K_V in Table 1, Eq. (3) and the bulk heat-dilatation coefficient α_V in Table 1, Eq. (4). This sub-model shares the *bulk curing-dilatation coefficient* in Table 1, Eq. (2) with the Reuss model, as it is constant during curing.

Note, that the sub-models are distinct in view of volumetric quantities, but both share the assumptions for the shear modulus in Table 1, Eq. (1). $G(Adolf)$ coincides basically with the relation given in [21] and applied in [22], [15]. The motivation for the ansatz is that below the gel point z_{gel}, the liquid resin of the matrix is allowing no more than hydrostatic pressure.

- Heterogeneous matrix: This meso-model denoted as het is used to obtain effective meso bulk properties from an analytical solution contrary to bounds for the mixture. In this way, the *composite spheres model* CHRISTENSEN [23], originally providing only an exact solution for the *effective bulk compression modulus* of a heterogeneous medium, is extended to thermal and chemical effects. Thus, the approach in [13] provides the *effective bulk heat-* and *effective curing-dilatation coefficients* α and β which are given in Table 1, Eq. (9) and Eq. (10) in addition to the *effective bulk compression modulus* K in Table 1, Eq. (8). For determination of both, α and β, the quantities L_{th} and L_{cur} are required. They are given in Appendix A, Eq. (A.1) and Eq. (A.2), respectively. We remark, that α and K resulting from the derivations in [13] are in agreement with that in [24] and [23], respectively. Moreover, we assume that the shear modulus is identical to the derivation in [23], [25], [26], [27] as it is not a volumetric quantity which might be affected by the purely volumetric modifications proposed in [13] (even the bulk compression modulus K as volumetric quantity is independent of the modifications in terms of heat- or curing-dilatation). For determination of the shear modulus $G_2(Christ)$ in Table 1, Eq. (8) the coefficients $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ are given in Appendix B, Eq. (A.3) to (A.6).

Remark 1

Due to the upper case in Table 1, Eq. (1), the heterogeneous and homogeneous matrix models already differ in view of the shear modulus for the uncured state $z[0] = 0$:

$$\begin{aligned} &1.\ hom\text{:} \quad G[0 \leq z < z_{\text{gel}}] = 0 \qquad &&\Longrightarrow \quad G[z=0] = 0, \\ &2.\ het\text{:} \quad G^m \leq G[0 \leq z \leq 1] \leq G^{sol} \qquad &&\Longrightarrow \quad G[z=0] = G^m. \end{aligned} \tag{1}$$

4 MESOSCOPIC MODELING

Based on the preliminaries of Section 3 the following exposition is directed to mesoscopic modeling, where we outline the approach in [14], where a general thermodynamic framework is specialized to the scenario of a mechanical-thermal-chemical process for polymeric materials.

4.1 Basic relations for the meso-RVE

The displacement gradient $\nabla \mathbf{u}$ is introduced at each material point P on the mesoscale Π in Figure 2.a. This defines the strain tensor of the geometric linear theory $\boldsymbol{\varepsilon}$ which is additively decomposed

• Mass phase fractions $\quad 1.\ \zeta^m[t] = \zeta^r[t] + \zeta^{ca}[t] = \dfrac{dm^r[t]}{dm_0} + \dfrac{dm^{ca}[t]}{dm_0}, \quad 2.\ \zeta^{sol}[t] = \dfrac{dm^{sol}[t]}{dm_0}$

• Degree of cure $\quad 1.\ \zeta^m = 1 - \dfrac{a^3}{b^3} \qquad 2.\ \zeta^{sol} = 1 - \zeta^m = \dfrac{a^3}{b^3} = z \qquad 3.\ \sum_{j=1}^{2} \zeta^j = 1,$
$4.\ \zeta^j \geq 0, \quad j = m, sol \quad 5.\ \underline{\zeta}_0 = [\zeta^m[0] = 1, \zeta^{sol}[0] = 0]^T$

• Homogeneous matrix (hom)

1. Shear modulus $G(Adolf) = \begin{cases} 0 & z < z_{\text{gel}}, \\ G^{sol}\left(\dfrac{z^2 - z_{\text{gel}}^2}{1 - z_{\text{gel}}^2}\right)^{8/3}, & z \geq z_{\text{gel}} \end{cases}$
2. Curing-dilatation coefficient $\beta(est) = -\beta^m + \beta^{sol}$

Voigt model $(homV)$

3. Compression modulus $K_V = K^m(1-z) + K^{sol}z$
4. Heat-dilatation coefficient $\alpha_V = \dfrac{(1-z)E^m\alpha^m + zE^{sol}\alpha^{sol}}{(1-z)E^m + zE^{sol}}$

Reuss model $(homR)$

5. Compression modulus $K_R = \dfrac{1}{\dfrac{1}{K^m}(1-z) + \dfrac{1}{K^{sol}}z}$
6. Heat-dilatation coefficient $\alpha_R = \alpha^m(1-z) + \alpha^{sol}z$

• Heterogeneous matrix (het)

7. Shear modulus $G_2(Christ) = -\dfrac{G^m\left(\mathfrak{B} - \sqrt{\mathfrak{B}^2 - 4\mathfrak{A}\,\mathfrak{C}}\right)}{2\mathfrak{A}}$
8. Compression modulus $K = K^m + \dfrac{z(K^{sol} - K^m)}{1 + \dfrac{(1-z)(K^{sol} - K^m)}{K^m + \dfrac{4G^m}{3}}}$
9. Heat-dilatation coefficient $\alpha = \dfrac{L_{th}\hat{p} + \dfrac{3K^m}{4G^m}L_{th}\hat{p} - 3K^m\alpha^m}{4G^m}$
10. Curing-dilatation coefficient $\beta = \dfrac{L_{cur}\hat{p} + \dfrac{3K^m}{4G^m}L_{cur}\hat{p} - 3K^m\beta^m}{z4G^m}$

• Material parameters $\underline{\kappa}_{hard} = [K^{sol}, K^m, G^{sol}, G^m, z_{\text{gel}}, \beta^m, \beta^{sol}, \alpha^m, \alpha^{sol}]$

Table 1: Summary and comparison of mesoscopic effective properties and bounds from [13]

into elastic, thermal and chemical parts, $\boldsymbol{\varepsilon}_{el},\ \boldsymbol{\varepsilon}_{th},\ \boldsymbol{\varepsilon}_{cur}$, in Table 2, Eq. (3.1). The volumetric $\boldsymbol{\varepsilon}^{vol}$ and deviatoric parts $\boldsymbol{\varepsilon}^{dev}$ of the strain tensor $\boldsymbol{\varepsilon}$ are defined in in Table 2, Eq. (1). For later use, we apply the *elastic strain tensor* $\boldsymbol{\varepsilon}_{el}$ in Table 2, Eq. (3.1) and define the *elastic volumetric strain* e_{el} in Table 2, Eq.

1. Volumetric and deviatoric strains	$\boldsymbol{\varepsilon}^{vol} = \frac{1}{3}\mathrm{tr}[\boldsymbol{\varepsilon}]\mathbf{1}, \quad \boldsymbol{\varepsilon}^{dev} = \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^{vol}$
2. Thermal and shrinking strains	$\boldsymbol{\varepsilon}_{th} = \int_{s=-\infty}^{t} \alpha[s] \frac{d\theta}{ds} ds\mathbf{1},\ \alpha[s] = \left\{ \begin{array}{l} het \\ homV, homR \end{array} \right.$
	$\boldsymbol{\varepsilon}_{cur} = \int_{s=-\infty}^{t} \beta[s] \frac{dz}{ds} ds\mathbf{1},\ \beta[s] = \left\{ \begin{array}{l} het \\ hom \end{array} \right.$
3. (Relative) elastic strain	1. $\boldsymbol{\varepsilon}_{el}[s] = \boldsymbol{\varepsilon}[s] - \boldsymbol{\varepsilon}_{th}[s] - \boldsymbol{\varepsilon}_{cur}[s]$, 2. $e_{el}[s] = \mathrm{tr}\boldsymbol{\varepsilon}_{el}[s]$,
	3. $\boldsymbol{\varepsilon}_t[s] = \boldsymbol{\varepsilon}[s] - \boldsymbol{\varepsilon}[t]$, 4. $e_t[s] = e_{el}[s] - e_{el}[t]$
4. Volumetric stress	$\boldsymbol{\sigma}_{vol} = \int_{s=-\infty}^{t} K[s,t] \frac{d}{ds} e_{el}[s]\ \mathbf{1}\ ds,\ K_0[s,t] = \left\{ \begin{array}{l} het \\ homV, homR \end{array} \right.$
5. Deviatoric stress	$\boldsymbol{\sigma}_{dev} = \int_{s=-\infty}^{t} 2G[s,t] \frac{d}{ds} \boldsymbol{\varepsilon}^{dev}[s] ds$
6. Shear modulus	$G[s,t] = G_0[s] + G_{visco}[s,t],\ G_0[s,t] = \left\{ \begin{array}{l} het \\ hom \end{array} \right.$
7. Curing rate	$\dot{z} = k_A \exp\left(\frac{-E_A}{R\theta}\right) z^a (1-z)^n$
8. Material parameters	$\underline{\kappa}_{kint} = [k_A, E_A, R, a, n, H_T]$
	$\underline{\kappa}_{hard} = [G^m, G^{sol}, \beta^m, \beta^{sol}, K^m, K^{sol}, \alpha^m, \alpha^{sol}]$

Table 2: Constitutive relations for thermo-visco-elasticity and curing for the matrix from [14]

(3.2), which both account for thermal-mechanical and chemical-mechanical coupling. Furthermore, we define the *relative strain tensor* $\boldsymbol{\varepsilon}_t[s]$ and the *relative elastic volumetric strain* e_t in Table 2, Eq. (3.3,4).

4.2 A prototype model for the matrix

The effective mesoscopic properties and bounds summarized in Section 3 (meso-models hom and het) are now applied to the prototype model in [15] to model the matrix as visco-elastic with a coupling to curing. The resulting constitutive relations for the thermal-mechanical-chemical coupled process accounting also for curing are summarized in Table 2, Eq. (1) to Eq. (8), where as in Section 3 two different meso-models are distinguished:

- Homogeneous matrix: According to this assumption the Voigt (upper) and Reuss (lower) bounds for meso bulk material properties are applied. Thus, the sub-models which are summarized in Table 1 $homV$ (Eq. (1)-(4)) and $homR$ (Eq. (1)-(2) and Eq. (5)-Eq. (6)) are employed.
- Heterogeneous matrix: According to this assumption the *effective* meso bulk properties denoted as model het in Table 1 (Eq. (7) -(10)) are employed.

4.2.1 Cure evolution

Following [19], the evolution of the degree of cure is expressed as a function of degree of cure itself in Table 2, Eq. (7), and drives the mass phase fractions $\zeta^m[s]$, $\zeta^{sol}[s]$ in Table 1. It exhibits the property $\dot{z} \geq 0$, where k_A, E_A, R are respectively the Arrhenius constant, the activation energy and the universal gas constant. Furthermore, a, n are power constants.

4.2.2 Stresses

The stress tensor $\boldsymbol{\sigma}$ is obtained by use of a special choice of the Helmholtz free energy which includes the chemically stored energy with respect to curing in [14]. Here, the decomposition $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{vol} + \boldsymbol{\sigma}_{dev}$, is applied, where $\boldsymbol{\sigma}_{vol}$ is the volumetric stress tensor and $\boldsymbol{\sigma}_{dev}$ is the deviatoric stress tensor which are stated in Table 2, Eq. (4) and (5), respectively, and which are be focused in the following two paragraphs.

Firstly, we consider the volumetric part of the stress tensor. Resulting from the decomposition in Table 2, Eq. (3) the stress tensor $\boldsymbol{\sigma}_{vol}$ in Table 2, Eq. (4) is written in terms of the time derivative of the *elastic volumetric strain* e_{el}

$$\frac{d}{ds}e_{el}[s] = \frac{d}{ds}\text{tr}\boldsymbol{\varepsilon}[s] - \frac{d}{ds}\text{tr}\boldsymbol{\varepsilon}_{th}[s] - \frac{d}{ds}\text{tr}\boldsymbol{\varepsilon}_{cur}[s] = \frac{d}{ds}\text{tr}\boldsymbol{\varepsilon}[s] - \frac{d}{ds}(3\alpha[s]\Delta\theta[s]) - \frac{d}{ds}(3\beta[s]z[s]), \quad (5)$$

where $\Delta\theta[s]$ is the temperature change. Although $e_{el}[s]$ takes the same form for both meso-models, however, a difference for the thermal and the curing strains $\boldsymbol{\varepsilon}_{th}$ and $\boldsymbol{\varepsilon}_{cur}$ occurs due to the choice of the meso bulk properties α and β:

- Homogeneous matrix: According to Table 2, Eq. (2), the *weighted* meso bulk curing-dilatation coefficient β for meso-model hom is given in Table 1, Eq. (2). Furthermore, the *weighted* meso bulk heat-dilatation coefficient α for meso-models $homV$ or $homR$ is stated in Table 1, Eq. (4) or Eq. (6), respectively. Due to the use of the quantities from above in the second and third term of the last part in Eq. (5) their time derivatives are required for $\boldsymbol{\sigma}_{vol}$. For details on them see [14].

- Heterogeneous matrix: According to Table 2, Eq. (2), α and β in the third part of Eq. (5) are the *effective* meso bulk properties used in meso-model het of Table 1, Eq. (9) and Eq. (10), respectively, which are also used to define the thermal and the curing strains $\boldsymbol{\varepsilon}_{th}$ and $\boldsymbol{\varepsilon}_{cur}$. The time derivatives of the second and third term for the last part of Eq. (5) are required for the determination of $\boldsymbol{\sigma}_{vol}$ and given in [14].

Moreover, the tensor $\boldsymbol{\sigma}_{vol}$ in Table 2, Eq. (4), is written in terms of the actual compression modulus K which is formulated in terms of the equilibrium compression modulus K_0 and also dependent on one of the underlying assumptions on the microscale. We assume a dependence on the degree of cure for the equilibrium bulk compression modulus K_0:

- Homogeneous matrix: As a consequence of the *weighted* meso bulk properties for $homV$ and $homR$ in Table 1, Eq. (3) or Eq. (5) are applied, respectively.

- Heterogeneous matrix: As a consequence of the *effective* meso bulk property for het in Table 1, Eq. (8) is applied.

Secondly, we consider the deviatoric stress tensor $\boldsymbol{\sigma}_{dev}$ in Table 2, Eq. (5). The shear modulus G for the tensor $\boldsymbol{\sigma}_{dev}$ in Table 2, Eq. (5), is formulated as a Prony series in Table 2, Eq. (6), where G_{visco} accounts for relaxation in dependence on the degree of cure (cf. [28], [22]) and G_0 is the equilibrium shear modulus. We formulate the latter in dependence on one of the underlying assumptions on the microscale and also dependent on the degree of cure:

- Homogeneous matrix: G_0 in Table 2, (6) is defined in Table 1, Eq. (1). For details see [21], [22].

- Heterogeneous matrix: Table 1, Eq. (7), is used for determination of G_0 in Table 2, Eq. (6) as an *effective* meso bulk property. For details see [13].

Finally, required material parameters are collected in the vectors $\underline{\kappa}_{kint}$ and $\underline{\kappa}_{hard}$, in Table 2, Eq. (8).

Remark 2

The degree of cure z occurs in the denominator of the effective bulk curing-dilatation coefficient β in Table 1, Eq. (10). Taking into account the initial state of the degree of cure $z[0] = 0$ dividing by z leads to a singularity of the term. However, if one considers β not on its own account but with regard to the curing strains $\boldsymbol{\varepsilon}_{cur}$ in Table 2, Eq. (2), it becomes clear that the degree of cure z cancels out:

$$\boldsymbol{\varepsilon}_{cur}[z] = \beta z \mathbf{1} = \frac{L_{cur}\hat{p} + \dfrac{3K^m}{4G^m} L_{cur}\hat{p} - 3K^m\beta^m}{z4G^m} z\mathbf{1} = \left(\left(3 + \frac{9K^m}{4G^m} \right) \frac{\hat{p}}{4G^m} L_{cur} - \frac{9K^m\beta^m}{4G^m} \right) \mathbf{1}. \tag{6}$$

4.3 Fibre modeling

For notational benefits, from on this subsection we provide the index $J = M, F$ in order to underline that the quantities stated in the following Section include the behavior of the total fibre-matrix-composite. The fibre $\bullet^F$ in Figure 2.a is considered as a thermo-linear-elastic solid. Consequently it can be modeled as a simplification with the equations for the matrix $\bullet^M$ in Table 2.

5 MESO TO MACRO TRANSITION

The following exposition is directed to the derivation of macroscopic effective properties. For the time being we are interested only in macroscopic *effective properties* for the thermo-linear-elastic fully cured composite. To this end, we denote the instant when the matrix reaches the fully cured state as $t_{z=1} = \min(t[z = 1, \dot{z} = 0)$ and assume from that instant onwards that the macroscopic strain tensor of the geometric linear theory $\bar{\boldsymbol{\varepsilon}}$ is additively decomposed into elastic and thermal parts:

$$\bar{\boldsymbol{\varepsilon}} = \operatorname{sym}\{\nabla \bar{\boldsymbol{u}}\} = \bar{\boldsymbol{\varepsilon}}^{el} + \bar{\boldsymbol{\varepsilon}}^{th}, \quad \text{for } t > t_{z=1}, \tag{7}$$

where $\bar{\boldsymbol{u}}$ is the displacement vector defined at any point $\bar{P}$ of the macrocontinuum $\bar{\Pi}$ in Figure 2.a. Thus, neglecting inelastic effects, we employ the following constitutive equations for the strains in Eq. (7)

$$1.\ \bar{\boldsymbol{\varepsilon}}^{el} = \mathbb{C}^{eff\,-1} \bar{\boldsymbol{\sigma}}, \quad 2.\ \bar{\boldsymbol{\varepsilon}}^{th} = \boldsymbol{A}^{eff} \Delta\bar{\theta}, \tag{8}$$

with the temperature change $\Delta\bar{\theta}$ that takes place uniformly throughout the meso-RVE Ω (which yields $\bar{\theta} = \theta^J$, $J = F, M$). Related matrices in Voigt notation of the *macroscopic effective elasticity tensor* $\mathbb{C}^{eff}$ and the *macroscopic effective heat-dilatation tensor* $\boldsymbol{A}^{eff}$ in Eq. (8) are defined as

$$\begin{aligned} &1.\ \underline{\mathbf{C}}^{eff} = [\underline{\hat{\sigma}}_1, ..., \underline{\hat{\sigma}}_n, ..., \underline{\hat{\sigma}}_6], \ \text{where}\ 2.\ \underline{\hat{\sigma}}_n = [\underline{\hat{\sigma}}_{1n}, ..., \underline{\hat{\sigma}}_{mn}, ..., \underline{\hat{\sigma}}_{6n}]^T, \ \text{for } m, n = 1, ..., 6, \\ &3.\ \underline{\mathbf{A}}^{eff} = [A_1, ..., A_l, ..., A_6]^T, \ \text{for } l = 1, ..., 6. \end{aligned} \tag{9}$$

The tensors $\mathbb{C}^{eff}$, $\boldsymbol{A}^{eff}$ both will be determined by application of a computational homogenization scheme, where the mesostructure variables are related to the homogenized macrocontinuum $\bar{\Pi}$ by use of the *Volume Averaging Theorem*, as formulated in [30]. Based on the decomposition $\Omega = \Omega^M \cup \Omega^F$ in Subsection 2.2 we redefine the overall strains $\bar{\boldsymbol{\varepsilon}}(t)$ and stresses $\bar{\boldsymbol{\sigma}}(t)$ at point $\bar{P}$ in dependence on the mesoscopic strains $\boldsymbol{\varepsilon}$ and stresses $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{vol} + \boldsymbol{\sigma}_{dev}$ (for the meso-RVE $\Omega = \Omega^M \cup \Omega^F$) in Table 2, Eq. (3.1) and Eq. (4-5), respectively:

$$1.\quad \bar{\boldsymbol{\varepsilon}}[t] = \textstyle\sum_J \dfrac{1}{|\Omega^J|} \displaystyle\int_{\Omega^J} \boldsymbol{\varepsilon}^J[t] dV^J, \quad 2.\ \bar{\boldsymbol{\sigma}}[t] = \sum_J \frac{1}{|\Omega^J|} \int_{\Omega^J} \boldsymbol{\sigma}^J[t] dV^J, \ \text{for } J = M,\ F. \tag{10}$$

The main idea is the determination of $\mathbb{C}^{eff}$ and $\boldsymbol{A}^{eff}$ due to prescribed thermal and deformation loadings applied to the unit cell Ω for $t > t_{z=1}$, respectively. In the following, we distinguish between prescribed purely thermal loading for the determination of $\boldsymbol{A}^{eff}$ and prescribed isothermal mechanical loading, i.e. prescribed strains $\hat{\boldsymbol{\varepsilon}}$, for the determination of $\mathbb{C}^{eff}$, where we adapt the approach in [31]. For the determination of $\mathbb{C}^{eff}$ we restrict ourselves to periodic displacement and anti-periodic stress BCs which assure the kinematic compatibility of two neighboring unit cells/meso-RVEs.

6 REPRESENTATIVE EXAMPLES

In this section two numerical examples are presented. In the first example, we illustrate the mechanical-thermo-chemical coupling for a meso-RVE in a three-dimensional finite-element simulation in Subsection 6.1. Here, we apply the material parameters from a parameter identification for a differential scanning calorimetry test (DSC) of an epoxy resin at two temperature rates in [14]. Some features on the coupling of temperature, curing and visco-elasticity of the proposed model in Table 2 are illustrated for a mesoscopic problem and its homogenized macroscopic response occurring in a curing process. On this basis, in the second example in Subsection 6.2 we determine the effective macroscopic anisotropic thermal expansion coefficient and the effective macroscopic anisotropic stiffness tensor for the fully cured fibre reinforced polymer.

6.1 Thermal-mechanical loading during curing of a meso-RVE

In this example, based on the comparative study in [13] for the effective properties and bounds given in Table 1 for $homV$, $homV$, het, some features on the coupling of temperature, curing and visco-elasticity of the proposed model in Table 2 are illustrated for a mesoscopic problem and its homogenized macroscopic response occurring in a curing process. As a simplification for the geometry of a real CFRP (carbon fibre reinforced composite) is shown in Figure 3.a. Here, stiff carbon fibres (blue) embedded in a soft epoxy resin matrix (yellow) are used to model a three dimensional RVE on the mesoscale with a fibre-volume-fraction of 19.6 %. The constitutive response of the stiff fibres is assumed as linear thermo-elastic, cf. Subsection 4.3, while the matrix material is assumed as visco-elastic, hence the meso-models in Table 2 accounting for a homogeneous (hom) and a heterogeneous matrix (het) are applied. Material parameters for the fibres $\underline{\kappa}_{carb}$ in Table 3 are taken from [9], while the remaining ones for the matrix are taken from Table 4. The thermal loading program which is handled as a homogeneously distributed temperature field on the mesoscale is divided into four phases, see θ in Figure 3.b. They are in accordance with the curing process of fibre reinforced composites.

- Phase 1, called HEAT phase: The temperature is increased with constant rate for $0 \leq t \leq 200$ s from the initial value $\theta = 25\ ^\circ C$ to a final value of $\theta = 180\ ^\circ C$.
- Phase 2, called CURE phase: The temperature $\theta = 180\ ^\circ C$ is kept fixed for $200 \leq t \leq 1800$ s.
- Phase 3, called COOL phase: The temperature is decreased from temperature $\theta = 180\ ^\circ C$ at constant rate to room temperature $\bar{\theta} = 25\ ^\circ C$ for $1800 \leq t \leq 2000$ s.
- Phase 4, called HOLD phase: The temperature $\theta = 25\ ^\circ C$ is kept fixed until 10^{15} s.

The displacement boundary conditions of the RVE are statically determined such that free expansion or contraction, respectively, due to thermal loading is possible.

$\underline{\kappa}_{carb}$:	K [MPa]	G [MPa]	α [1/K]
	209091.0	107812.0	$5.6{\cdot}10^{-6}$

Table 3: Values for carbon fibre material parameters (exerpt)

$\underline{\kappa}_{hard}$:	K^m [MPa]	K^{sol} [-]	G^m [MPa]	G^{sol} [-]	β^m [-]	β^{sol} [-]	α^m [1/K]	α^{sol} [1/K]
	150.0	3000.0	120.0	1200.0	0.0	-0.04	$1.5\cdot10^{-1}$	$7.5{\cdot}10^{-5}$

Table 4: Values for matrix material parameters (exerpt) in Tables 1 and 2

The thermal-mechanical-chemical coupled problem is solved with the finite element code ”Abaqus/ Standard” [32] by aid of a ”UMAT” subroutine for the meso-models in Table 2. The whole domain

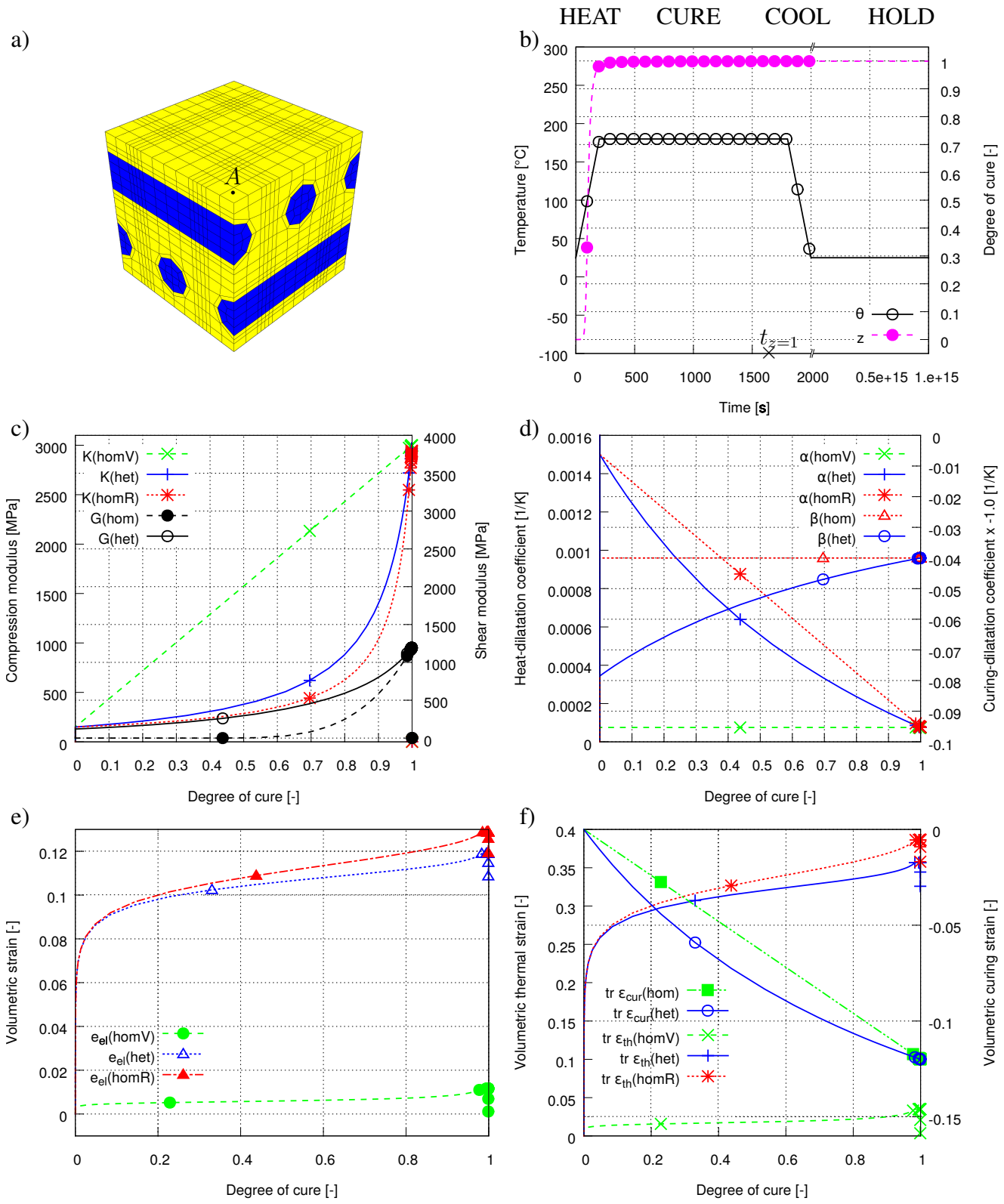


Figure 3: Thermal-mechanical loading during curing of a meso-RVE: a) Geometry with FE-mesh: carbon-fibers (blue) in resin-matrix (yellow) and location of point A considered in the following: b) thermal loading and degree of cure within four phases versus time, c) compression and shear moduli, d) heat- and curing-dilatation coefficients, e) elastic and f) thermal as well as curing volumetric strains versus the degree of cure shown in b).

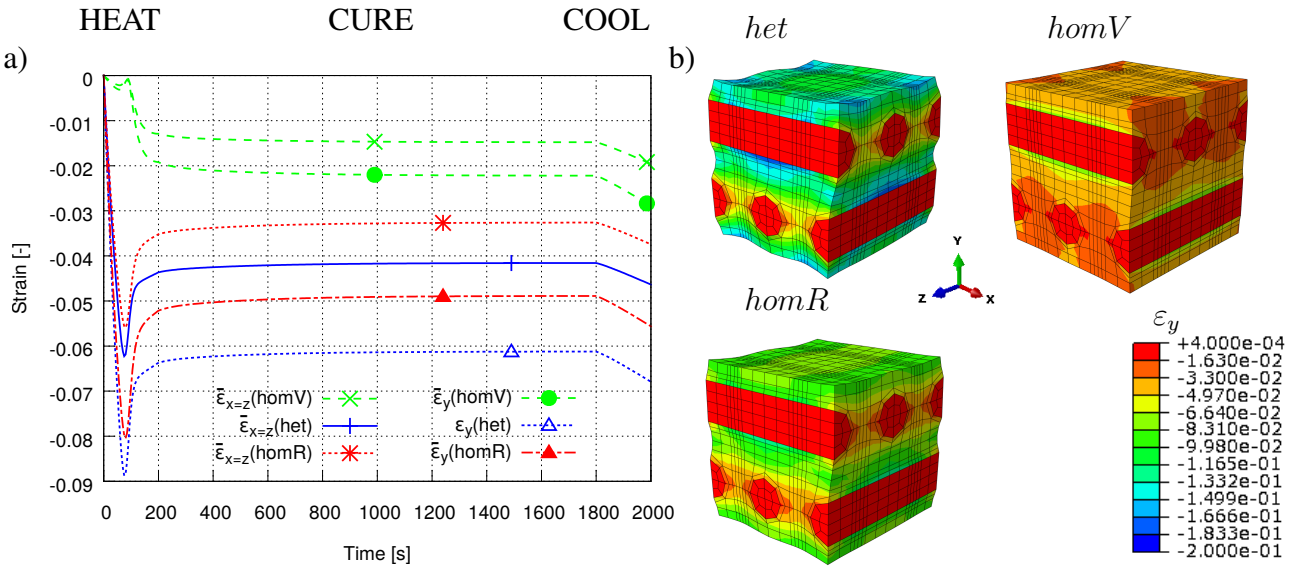


Figure 4: Thermal-mechanical loading during curing of a meso-RVE: a) macroscopic strain vs. time, b) and strain-conturplots resulting from loading program in Figure 3.b for three meso-models for the matrix $homV$, $homR$ and het after phase HOLD

shown in Figure 3.a is discretized using eight node linear hexahedral elements with reduced integration and hourglass control (C3D8R, [32]). Following [33], post processing on the "odb-file" is done using a "PYTHON" script to obtain homogenized quantities on the macro-scale.

Mesoscopic behavior of the matrix at point A

In Figure 3 we summarize some relevant quantities at Point A (in the matrix part) of Figure 3.a resulting from the mechanical-thermal-chemical coupling including the influence of curing and temperature on some material parameters. Figure 3.b shows the degree of cure z at point A versus time. We observe that after the phase CURE the material is fully cured ($z = 1$) which means that the time point $t_{z=1}$ is passed or at least reached. According to the meso-models for $homV$, $homR$ and het the equilibrium shear and bulk moduli Table 2, Eq.(4) and Eq.(6), versus the degree of cure are illustrated in Figure 3.c. Note, that in reality the epoxy behaves like a viscous liquid until gelation at z_{gel} occurs. It is only from that instant onwards that a finite shear modulus is detected. The heat- and curing-dilatation coefficient Table 2, Eq.(2), for the meso-models $homV$, $homR$ and het versus the degree of cure are illustrated in Figure 3.d. As kinematic quantities the volumetric elastic strain tr$\boldsymbol{\varepsilon}_{el}$, the volumetric thermal strain tr$\boldsymbol{\varepsilon}_{th}$ and the volumetric curing strain tr$\boldsymbol{\varepsilon}_{cur}$ are shown for the meso-models $homV$, $homR$ and het versus the degree of cure in Figure 3.e and Figure 3.f.

We observe that the effective meso bulk properties (compression modulus, heat-dilatation coefficient) resulting from meso-model het lie, on the one hand, within the bounds given by meso-models $homV$ and $homR$ and, on the other, within the physically meaningful bounds given by the values for matrix material parameters in Table 4. However, there are two exceptions, the shear modulus for meso-model hom and the curing-dilatation coefficient for meso-model het. The former lies not within the bounds, given by the microscopic material parameters, due to the nature of its definition in Table 1, Eq. (1) (see also Remark 1), the latter shows satisfying results for the shrinkage in Figure 3.d (see also Remark 2) which is in the same physically meaningful region as the shrinkage resulting from meso-model hom and not in contrast to the experimental results in [20].

We conclude, that in the HEAT and CURE phases curing becomes a dominating nonlinear effect, which in combination with the cure-dependent quantities, i.e. the compression and bulk moduli as

well as the dilatation coefficients, renders the strains as nonlinear. The elastic, thermal and curing strains in Figure 3.f do not retain their initial zero value at the fully cured state $t_{z=1}$. This is due to the change of the material properties in Figure 3.c and Figure 3.d. The final, deformed state of the cured RVE is obtained after the HOLD phase. Figure 4.b shows contour plots of the strain component ε_y for each of the meso-models $homV$, $homR$ and het. We observe that for each meso-model a different eigenstrain state remains. Meso-model $homV$ has the lowest amount of eigenstrain in the matrix, moderate eigenstrains are observed in the matrix where meso-model $homR$ is used and very inhomogeneous eigenstrains occur in the matrix accompanied with meso-model het.

Macroscopic behavior of the meso-RVE

Applying Eq. (10.1), homogenization leads to results on the less resolved macroscale. In Figure 4.a the components $\bar{\varepsilon}_{x,y,z}(homV), \bar{\varepsilon}_{x,y,z}(homR), \bar{\varepsilon}_{x,y,z}(het)$ of the macroscopic strain resulting from each meso-model are shown whereas the shear components remain zero at all and therefore are left out. In particular, shrinking due to the degree of cure is observed in phases HEAT and CURE. The macroscopic strains represent an orthotropic behavior due to the geometry of the meso-RVE. After the COOL phase there are remaining macroscopic (eigen-)strains. The amount is dependent on the chosen meso-model $homV$, $homR$ or het.

6.2 Effective properties of a meso-RVE after curing

For the meso-RVE from Section 6.1 a state with (meso- and macroscopic) eigenstrains after the phase HOLD (cf. Figure 3) at $t = 10^{15} > t_{z=1}$ remains which differs for each of the meso-models $homV$, $homR$ and het. Thus, we define the macroscopic strain state for each meso-model using Eq. (10.1): $\bar{\boldsymbol{\varepsilon}}^{\text{HOLD}}(k) = \bar{\boldsymbol{\varepsilon}}(k, t = 10^{15})$, for $k = homV,\ homR,\ het$. In the following, based on the loading program of Subsection 6.1, the effective properties for the fully cured meso-RVE are determined in a two-part example. In particular we determine the macroscopic thermal expansion tensor following an approach in [34] and the elastic stiffness tensor for $t > t_{z=1}$ following an approach in [31]. Therefore, the meso-RVE is subjected to two cases of second loadings, the purely thermal and the purely mechanical loading.

Macroscopic anisotropic thermal expansion tensor

The the meso-RVE with statically determined displacement boundary conditions is subjected to purely thermal second loading in this first part of the two-part example. For

- Phase 5, called THERM2 phase: The temperature is increased with constant rate in a time period of 10^{-6} s from $\theta = 25\ ^\circ C$ to a final value of $\theta = 26\ ^\circ C$ (temperature unit loading, $\theta = \bar{\theta} \Longrightarrow \Delta\bar{\theta} = 1$, as shown in Figure 5.a).

Insertion of Eq. (8) into Eq. (7) with the conditions of purely thermal loading of phase THERM2 in combination with the statically determined BCs yields:

$$\bar{\boldsymbol{\varepsilon}} - \bar{\boldsymbol{\varepsilon}}^{\text{HOLD}} = \bar{\boldsymbol{\varepsilon}}^{th} = \boldsymbol{A}^{eff} \Delta\bar{\theta}. \tag{11}$$

Next, we define the macroscopic strain state after the loading THERM2 by use of Eq. (10.1) as $\bar{\boldsymbol{\varepsilon}}^{\text{THERM2}}(k)$, for $k = homV,\ homR,\ het$. Exchanging the left term of Eq. (11) by $\bar{\boldsymbol{\varepsilon}}^{\text{THERM2}}(k)$ gives the coefficients for the anisotropic thermal expansion tensor in Eq. (9): $\bar{\boldsymbol{\varepsilon}}^{\text{THERM2}}(k) = \boldsymbol{A}^{eff}(k)$, for $k = homV,\ homR,\ het$. The effective thermal expansion tensor in matrix representation $\underline{\mathrm{A}}^{eff}(k)$ for the meso-models k is shown in Table 5. We conclude that $\boldsymbol{A}^{eff}(k)$ is almost independent of the eigenstrains resulting from models $k = homV,\ homR,\ het$. Additionally, we remark, that the assumption of linear thermo-elasticity on the macroscale (which leads to Eq.(7)) is justified for small temperature changes and time increments, i.e. phase THERM2.

Meso-model	A_1 [1/K]	A_2 [1/K]	A_3 [1/K]	A_4 [1/K]	A_5 [1/K]	A_6 [1/K]
$\underline{\mathrm{A}}^{eff}(homV)$	$1.178 \cdot 10^{-5}$	$1.640 \cdot 10^{-5}$	$1.179 \cdot 10^{-5}$	$4.636 \cdot 10^{-10}$	$7.064 \cdot 10^{-10}$	$1.023 \cdot 10^{-9}$
$\underline{\mathrm{A}}^{eff}(homR)$	$1.174 \cdot 10^{-5}$	$1.649 \cdot 10^{-5}$	$1.191 \cdot 10^{-5}$	$1.209 \cdot 10^{-9}$	$1.936 \cdot 10^{-9}$	$1.916 \cdot 10^{-9}$
$\underline{\mathrm{A}}^{eff}(het)$	$1.173 \cdot 10^{-5}$	$1.646 \cdot 10^{-5}$	$1.181 \cdot 10^{-5}$	$8.559 \cdot 10^{-10}$	$2.424 \cdot 10^{-9}$	$4.567 \cdot 10^{-9}$

Table 5: Effective properties of a meso-RVE after curing: Macroscopic effective thermal expansion matrices $\underline{\mathrm{A}}^{eff}(k)$ of Eq. (9) for $k = homV, homR, het$) resulting from second loading in phase THERM2 as shown in Figure 5.a

Macroscopic anisotropic elastic stiffness tensor

The meso-RVE with periodic displacement boundary conditions is subjected to purely mechanical second loading in this second part of the two-part example. For

- Phase 5, called MECH phase: The macroscopic strain is prescribed at constant rate in a time period of 10^{-6} s with six different prescribed macroscopic strains in Voigt notation $\hat{\underline{\varepsilon}}_x = [10^{-3}, 0, 0, 0, 0, 0]^T$,..., $\hat{\underline{\varepsilon}}_{yz} = [0, 0, 0, 0, 0, 10^{-3}]^T$ (mechanical loading shown in Figure 5.b).

Insertion of Eq. (8) into Eq. (7) with the conditions of purely mechanical loading of phase MECH yields

$$\bar{\boldsymbol{\varepsilon}} - \bar{\boldsymbol{\varepsilon}}^{\mathrm{HOLD}} = \bar{\boldsymbol{\varepsilon}}^{el} = \mathbb{C}^{eff\,-1}\bar{\boldsymbol{\sigma}}. \tag{12}$$

Additionally, from Eq. (10.1) in combination with the prescribed macroscopic strains (cf. Section 5) one gets

$$\bar{\boldsymbol{\varepsilon}} - \bar{\boldsymbol{\varepsilon}}^{\mathrm{HOLD}} = \hat{\boldsymbol{\varepsilon}}_n, \text{ for } n = x, ..., yz. \tag{13}$$

Rearrangement of Eq. (12) and insertion Eq. (13) as well as taking into account the definition for $\underline{\mathrm{C}}^{eff}$ in Eq. (9.1) results into the following matrix representation in Voigt notation

$$\underline{\mathrm{C}}^{eff}\hat{\underline{\varepsilon}}_n = [\hat{\underline{\sigma}}_1, ..., \hat{\underline{\sigma}}_n, ..., \hat{\underline{\sigma}}_6]\hat{\underline{\varepsilon}}_n = \bar{\underline{\sigma}}(\hat{\underline{\varepsilon}}_n), \text{ for } n = x, ..., yz, \tag{14}$$

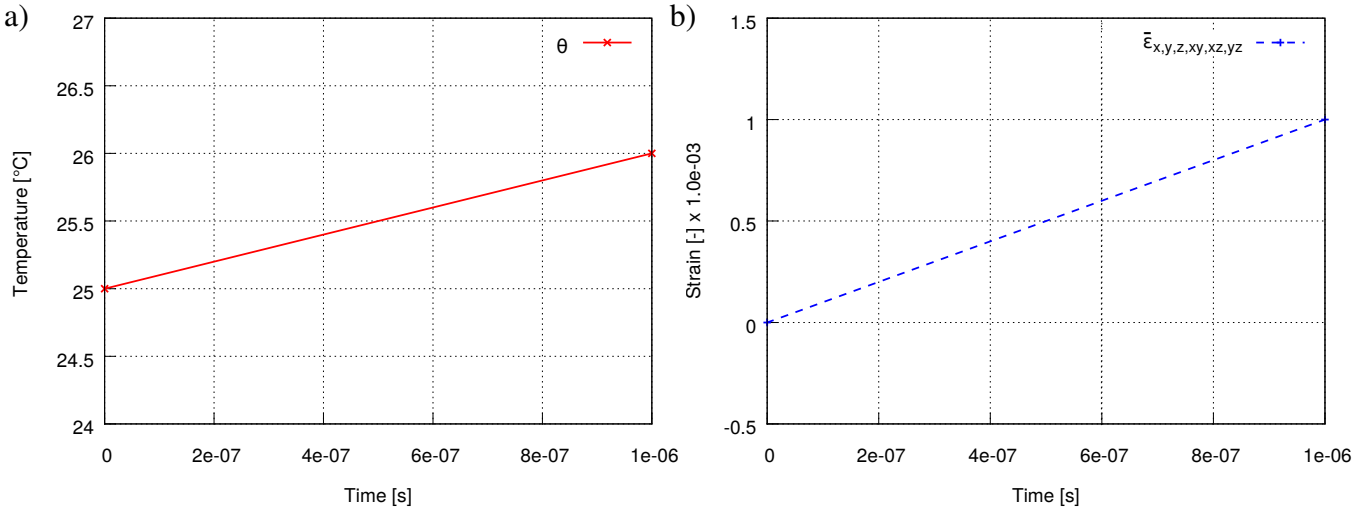


Figure 5: Effective properties of a meso-RVE after curing: Determination due to: a) thermal (THERM2) and b) mechanical (MECH) loading of the meso-RVE

Meso-model	$\underline{\hat{\sigma}}_1$ [MPa]	$\underline{\hat{\sigma}}_2$ [MPa]	$\underline{\hat{\sigma}}_3$ [MPa]	$\underline{\hat{\sigma}}_4$ [MPa]	$\underline{\hat{\sigma}}_5$ [MPa]	$\underline{\hat{\sigma}}_6$ [MPa]
$\underline{\mathbf{C}}^{eff}(homV)$	$6.082 \cdot 10^4$	$1.124 \cdot 10^3$	$9.076 \cdot 10^2$	$4.105e-02$	$6.104 \cdot 10^{-5}$	$-1.221 \cdot 10^{-4}$
	$1.124 \cdot 10^3$	$9.893 \cdot 10^3$	$1.124 \cdot 10^3$	$-1.328 \cdot 10^{-2}$	$1.709 \cdot 10^{-3}$	$1.495 \cdot 10^{-3}$
	$9.076 \cdot 10^2$	$1.124 \cdot 10^3$	$6.082 \cdot 10^4$	$4.883 \cdot 10^{-4}$	$9.155 \cdot 10^{-5}$	$1.831 \cdot 10^{-4}$
	$4.134 \cdot 10^{-2}$	$-1.491 \cdot 10^{-2}$	$2.507 \cdot 10^{-4}$	$4.385 \cdot 10^3$	$-1.026 \cdot 10^{-3}$	$2.915 \cdot 10^{-6}$
	$1.229 \cdot 10^{-4}$	$2.771 \cdot 10^{-6}$	$2.041 \cdot 10^{-5}$	$-9.907 \cdot 10^{-4}$	$4.647 \cdot 10^3$	$2.239 \cdot 10^{-4}$
	$-3.363 \cdot 10^{-5}$	$-5.244 \cdot 10^{-4}$	$-4.238 \cdot 10^{-4}$	$2.185 \cdot 10^{-6}$	$2.756 \cdot 10^{-4}$	$4.385 \cdot 10^3$
$\underline{\mathbf{C}}^{eff}(homR)$	$6.082 \cdot 10^4$	$1.124 \cdot 10^3$	$9.076 \cdot 10^2$	$4.192 \cdot 10^{-2}$	$9.308 \cdot 10^{-4}$	$7.477 \cdot 10^{-4}$
	$1.124 \cdot 10^3$	$9.893 \cdot 10^3$	$1.124 \cdot 10^3$	$-1.494 \cdot 10^{-2}$	$1.526 \cdot 10^{-4}$	$-1.688 \cdot 10^{-4}$
	$9.076 \cdot 10^2$	$1.124 \cdot 10^3$	$6.082 \cdot 10^4$	$3.052 \cdot 10^{-4}$	$4.120 \cdot 10^{-4}$	$0.000 \cdot 10^0$
	$4.134 \cdot 10^{-2}$	$-1.495 \cdot 10^{-2}$	$2.522 \cdot 10^{-4}$	$4.385 \cdot 10^3$	$-9.682 \cdot 10^{-4}$	$2.279 \cdot 10^{-6}$
	$1.231 \cdot 10^{-4}$	$1.397 \cdot 10^{-6}$	$2.007 \cdot 10^{-5}$	$-1.004 \cdot 10^{-3}$	$4.647 \cdot 10^3$	$2.642 \cdot 10^{-4}$
	$-3.489 \cdot 10^{-5}$	$-3.963 \cdot 10^{-4}$	$-4.342 \cdot 10^{-4}$	$2.204 \cdot 10^{-6}$	$3.073 \cdot 10^{-4}$	$4.385 \cdot 10^3$
$\underline{\mathbf{C}}^{eff}(het)$	$6.082 \cdot 10^4$	$1.124 \cdot 10^3$	$9.075 \cdot 10^2$	$4.021 \cdot 10^{-2}$	$-3.204 \cdot 10^{-4}$	$-1.068 \cdot 10^{-4}$
	$1.124 \cdot 10^3$	$9.893 \cdot 10^3$	$1.124 \cdot 10^3$	$-1.314 \cdot 10^{-2}$	$1.434 \cdot 10^{-3}$	$1.114 \cdot 10^{-3}$
	$9.076 \cdot 10^2$	$1.124 \cdot 10^3$	$6.082 \cdot 10^4$	$1.053 \cdot 10^{-3}$	$5.493 \cdot 10^{-4}$	$3.815 \cdot 10^{-4}$
	$4.135 \cdot 10^{-2}$	$-1.491 \cdot 10^{-2}$	$2.513 \cdot 10^{-4}$	$4.385 \cdot 10^3$	$-1.039 \cdot 10^{-3}$	$2.297 \cdot 10^{-6}$
	$1.231 \cdot 10^{-4}$	$1.829 \cdot 10^{-6}$	$2.034 \cdot 10^{-5}$	$-1.021 \cdot 10^{-3}$	$4.647 \cdot 10^3$	$2.418 \cdot 10^{-4}$
	$-3.515 \cdot 10^{-5}$	$-4.418 \cdot 10^{-4}$	$-4.605 \cdot 10^{-4}$	$2.498 \cdot 10^{-6}$	$2.809 \cdot 10^{-4}$	$4.385 \cdot 10^3$

Table 6: Effective properties of a meso-RVE after curing: Macroscopic effective stiffness matrices $\underline{\mathbf{C}}^{eff}$ of Eq. (9) for $k = homV, homR, het$) resulting from second loading in phase MECH as shown in Figure 5.b

where $\bar{\underline{\sigma}}(\hat{\underline{\varepsilon}}_n)$ are the stresses in Eq. (10.2) in Voigt notation resulting from the prescribed macroscopic strains in Voigt notation $\hat{\underline{\varepsilon}}_n$. For example for $n = x$ one gets

$$\underline{\mathbf{C}}^{eff}\hat{\underline{\varepsilon}}_x = \underline{\mathbf{C}}^{eff}[1,0,0,0,0,0]^T \cdot 10^{-3} = \hat{\underline{\sigma}}_1 \cdot 10^{-3} = [\hat{\underline{\sigma}}_{11}, ..., \hat{\underline{\sigma}}_{61}]^T \cdot 10^{-3} = \bar{\underline{\sigma}}(\hat{\underline{\varepsilon}}_x). \tag{15}$$

Taking into account the right-hand side of Eq. (15) we conclude that the n-th column of $\underline{\mathbf{C}}^{eff}$, denoted as $\hat{\underline{\sigma}}_n$, can be obtained due to the six prescribed macroscopic strains $\hat{\boldsymbol{\varepsilon}}_n$ in phase MECH, where $n = 1, ...6$. Consequently, proceeding in the same manner as in Eq. (15) for each of the meso-models $k = homV,\ homR,\ het$ gives the effective stiffness tensor in matrix representation $\underline{\mathbf{C}}^{eff}(k)$ which is shown in Table 6. We conclude that $\mathbb{C}^{eff}$ is almost independent of the eigenstrains resulting from models $k = homV,\ homR,\ het$. Moreover, we remark, that the assumption of linear thermo-elasticity on the macroscale (which leads to Eq. (7)) is justified for small strain and time increments, i.e. phase MECH.

7 CONCLUSIONS

In this paper, we applied a three-scale framework for fibre-reinforced polymer curing from [13]. The three-scale framework employs a meso-RVE consisting of the fibre and the matrix part and a micro-RVE for the latter consisting of the three components resin, curing agent and solidified material. The meso bulk properties in dependence on mass phase fractions resulting from the micro-RVE are taken from [13]. Thus, Voigt and Reuss bounds as well as effective properties are determined for the assumptions of a homogeneous and a heterogeneous matrix, respectively. In [14] those properties are incorporated into a mesoscopic constitutive model for temperature-dependent visco-elastic effects accompanied by curing which is summarized in this work. Thus, the incorporation of the meso

bulk properties, in particular the meso bulk compression and shear moduli as well as the heat- and shrinking-dilatation coefficients for the homogeneous as well as for the heterogeneous matrix, into a visco-elastic model with a coupling to curing [15] is outlined.

The simulation of the curing process for the meso-RVE subjected to mechanical-thermal-chemical loading demonstrates the capability of the model to simulate this process. The effect of anisotropic material behavior due to the incorporation of fibres into the polymeric matrix is constituted on the macroscale. Both, on the mesoscale, as well as on the macroscale the distinction between the homogeneous and the heterogeneous matrix assumption for the determination of the effective mesoscopic material properties influences the remaining eigenstrains.

Concerning further extensions, the meso bulk properties should also consider visco-elastic effective properties and a generalization for the *n-layered composite spheres model* in [27] could be taken into account. Taking the simulation of large-scale engineering finite-element structures into account, the meso to macro transition should be taken into account during the total curing process and parameter identification should be accomplished to obtain material properties or the visco-elastic model.

Acknowledgement

This work is based on investigations of the "SPP 1712 - Intrinsische Hybridverbunde fr Leichtbautragstrukturen", which is kindly supported by the Deutsche Forschungsgemeinschaft (DFG).

Appendix

In accordance with [13] for Table 1, equations (9), (10) and (8), respectively, we use the definitions

$$L_{th} = \frac{4\,G^m\,\left(3\,\alpha^m\;K^{sol}\,K^m + 4\,z\,\alpha^{sol}\,K^{sol}\,G^m - 4\,z\,\alpha^m\,K^m\,G^m + 4\,\alpha^m\;K^m\,G^m\right)}{\hat{p}\;\left(4\,z\,K^{sol}\,G^m - 4\,z\,K^m\,G^m + 4\;K^m\,G^m + 3\;K^{sol}\,K^m\right)}, \tag{A.1}$$

$$L_{cur} = \frac{4\,G^m\,\left(3\,\beta^m\;K^{sol}\,K^m + 4\,z\,\beta^{sol}\,K^{sol}\,G^m - 4\,z\,\beta^m\,K^m\,G^m + 4\,\beta^m\;K^m\,G^m\right)}{\hat{p}\;\left(4\,z\,K^{sol}\,G^m - 4\,z\,K^m\,G^m + 4\;K^m\,G^m + 3\;K^{sol}\,K^m\right)}, \tag{A.2}$$

and

$$\begin{aligned}\mathfrak{A} &= \eta_2\,z^{\frac{5}{3}}\,\left(\frac{252\,G^{sol}}{G^m} - 252\right) - \eta_2\,\eta_3\;\left(40\,\nu^m - 28\right) - z^{\frac{7}{3}}\,\left(4\,\eta_1\,\eta_3 + 2\,\eta_2\,\left(\frac{63\,G^{sol}}{G^m} - 63\right)\right)\\ &\quad -\eta_2\,z\,\left(\frac{50\,G^{sol}}{G^m} - 50\right)\,\left(8\,\nu^{m2} - 12\,\nu^m + 7\right) - \eta_1\,z^{\frac{10}{3}}\,\left(\frac{8\,G^{sol}}{G^m} - 8\right)\,\left(5\,\nu^m - 4\right)\end{aligned} \tag{A.3}$$

$$\begin{aligned}\mathfrak{B} &= z^{\frac{7}{3}}\,\left(4\,\eta_1\,\eta_3 + 2\,\eta_2\,\left(\frac{63\,G^{sol}}{G^m} - 63\right)\right) + \eta_2\,\eta_3\,\left(\frac{45\,\nu^m}{2} - \frac{21}{2}\right) - \eta_2\,z^{\frac{5}{3}}\,\left(\frac{252\,G^{sol}}{G^m} - 252\right)\\ &\quad +\eta_1\,z^{\frac{10}{3}}\,\left(\frac{2\,G^{sol}}{G^m} - 2\right)\,\left(5\,\nu^m - 1\right) - \eta_2\,\nu^m\,z\,\left(\frac{75\,G^{sol}}{G^m} - 75\right)\,\left(\nu^m - 3\right)\end{aligned} \tag{A.4}$$

$$\begin{aligned}\mathfrak{C} &= \eta_2\,z^{\frac{5}{3}}\,\left(\frac{252\,G^{sol}}{G^m} - 252\right) - \eta_2\,\eta_3\;\left(5\,\nu^m + 7\right) - z^{\frac{7}{3}}\,\left(4\,\eta_1\,\eta_3 + 2\,\eta_2\,\left(\frac{63\,G^{sol}}{G^m} - 63\right)\right)\\ &\quad +\eta_2\,z\,\left(\frac{25\,G^{sol}}{G^m} - 25\right)\,\left(\nu^{m2} - 7\right) + \eta_1\,z^{\frac{10}{3}}\,\left(\frac{4\,G^{sol}}{G^m} - 4\right)\,\left(5\,\nu^m - 7\right),\end{aligned} \tag{A.5}$$

where

$$\begin{aligned}1.\;\eta_1 &= 105\,\nu^{sol} - 105\,\nu^m - \left(\frac{G^{sol}}{G^m} - 1\right)\,\left(5\,\nu^{sol} + 7\right)\,\left(10\,\nu^m - 7\right),\\ 2.\;\eta_2 &= \left(\frac{G^{sol}}{G^m} - 1\right)\,\left(5\,\nu^{sol} + 7\right) - 35\,\nu^{sol} + 35,\\ 3.\;\eta_3 &= 15 - \left(\frac{G^{sol}}{G^m} - 1\right)\,\left(10\,\nu^m - 8\right) - 15\,\nu^m.\end{aligned} \tag{A.6}$$

REFERENCES

[1] E. RUIZ, F. TROCHU, Thermomechanical properties during cure of glass-polyester RTM composites: elastic and viscoelastic modeling, *J. Compos. Mater.*, **39**, 881-916, 2005

[2] J. LANGE, Viscoelastic properties and transitions during thermal and UV cure of a methacrylate resin, *Polym. Eng. Sc.*, **39**, 1651-1660, 1999

[3] D.B. ADOLF, R.S. CHAMBERS, J.M. CARUTHERS, Extensive validation of a thermodynamically consistent, nonlinear viscoelastic model for glassy polymers, *Polymer*, **45**, 4599-4621, 2004

[4] D.B. ADOLF, R.S. CHAMBERS, Verification of the capability for quantitative stress prediction during epoxy cure, *Polymer*, **38**, 5481-5490, 1997

[5] D.B. ADOLF, J.E. MARTIN, R.S. CHAMBERS, S.N. BURCHETT, T.R. GUESS, Stresses during thermoset cure, *J. Mater. Res.*, **13**, 530-550, 1998

[6] M. HOSSAIN, G. POSSART, P. STEINMANN, A small-strain model to simulate the curing of thermosets, *Comput. Mech.*, **43**(6), 769-779, 2009

[7] D.B. ADOLF, R.S. CHAMBERS, A thermodynamically consistent, nonlinear viscoelastic approach for modeling thermosets during cure, *J. Rheol.*, **51**, 23-50, 2007

[8] M. HOSSAIN, G. POSSART, P. STEINMANN, A finite strain framework for the simulation of polymer curing, *Comput. Mech.*, **44**, 621-630, 2009

[9] C. HEINRICH, M. ALDRIDGE, A.S. WINEMAN, J. KIEFFER, A. M. WAAS AND K. SHAHWAN, The influence of the representative volume element (RVE) size on the homogenized response of cured fibre composites. *Model. Simul. Materi. Sc.*, **20**, 75007, 2012

[10] P. HAUPT, A. LION, On finite linear viscoelasticity of incompressible isotropic materials, *Acta Mech.*, **159**, 87-124, 2002

[11] A. LION, On the large deformation behaviour of reinforced rubber at different temperatures, *J. Mech. Phys. Solids*, **45**, 1805-1834, 1997

[12] S. KLINGE, A. BARTELS, P. STEINMANN, Modeling of curing processes based on a multi-field potential. Single- and multiscale aspects, *Int. J. Solids Struct.*, **49**, 2320-2333, 2012

[13] R. MAHNKEN, C. DAMMANN, A three-scale framework for fibre-reinforced-polymer curing coupled to visco-elasticity Part I: Microscopic modeling and mesoscopic effective properties, submitted to *Int. J. Solids Struct.*, 2015

[14] R. MAHNKEN, C. DAMMANN, A three-scale framework for fibre-reinforced-polymer curing coupled to visco-elasticity Part II: Mesoscopic modeling and macroscopic effective properties, submitted to *Int. J. Solids Struct.*, 2016

[15] R. MAHNKEN, Thermodynamic consistent modeling of polymer curing coupled to viscoelasticity at large strains, *Int. J. Solids Struct.*, **50**(13), 2003-2021, 2013

[16] P. J. FLORY, Principles of Polymer Chemistry, *Cornell University Press*, 1953

[17] J.M.G. COWIE, V. ARRIGHI, Polymers: Chemistry and Physics of Modern Materials, *CRC Press*, New York, 2007

[18] J. K. STILLE: Step-Growth Polymerization, *Journal of Chemical Education*, **58**(11), 862-866, 2005

[19] A. LION, P. HÖFER, On the phenomenological representation of curing phenomena in continuum mechanics, *Arch. Mech.*, **59**, 59-89, 2007

[20] R. HILL; S. V. MUZUMDAR; L. J. LEE, Analysis of Volumetric Changes of Unsaturated Polyester Resins During Curing, *Polym. Eng. Sci.*, **35**(10), 852-859, 1995

[21] D.B. ADOLF, J.E. MARTIN, Calculation of stresses in crosslinking polymers, *J. Compos. Mater.*, **30**, 13-34, 1996

[22] M.H.H. MEUWISSEN, H.A. DE BOER, H. L.A.H. STEIJVERS, P.J.G. SCHREURS, M.G.D. GEERS, Residual stresses in microelectronics induced by thermoset packaging materials during cure, *Microelectro. Relia.*, **44**, 1985-1994, 2004

[23] R. M. CHRISTENSEN, *Mech.Compos. Mater.*, John Wiley and Sons, New York, 1979

[24] A. GUSEV, Effective coefficient of thermal expansion of n-layered composite sphere model: Exact solution and its finite element validation. *Int. J. Eng. Sci.*, **84**, 54-61, 2014

[25] R. M. CHRISTENSEN, K. H. LO, Solutions for effective shear properties in three phase sphere and cylinder models, *J. Mech. Phys. Solids*, **27**(4), 315-330, 1979

[26] R. M. CHRISTENSEN, A critical evaluation for a class of micromechancis models, *Mech. Phys. Solids*, **38**(3), 379-404, 1990

[27] E. HERVE, A. ZAOUI, N-layered inclusion based micromechanical modeling. *Int. J. Eng. Sc.*, **31**(1), 1-10, 1993

[28] J.E. MARTIN, D.B. ADOLF, Constitutive equation for cure-induced stresses in a viscoelastic material, *Macromolecules*, **23**, 5014-5019, 1990

[29] M.L. WILLIAMS, R.F. LANDEL, J.D. FERRY, The temperature dependence of relaxation mechanisms in amorphous polymers and other glass-forming liquids, *J. American Chem. Soc.*, **14**, 3701-3707, 1955

[30] R. HILL, On constitutive macro-variables for heterogeneous solids at finite strain, *Proc. R. Soc. London, A, Mathematical and Physical Sciences*, **326**(1565), 131-147, 1972

[31] M. KÄSTNER, *Skalenbergreifende Modellierung und Simulation des mechanischen Verhlatens von textilverstärktem Polypropylen unter Nutzung der XFEM*, Dissertation, TU Dresden, 2009

[32] ABAQUS, Dassault Systemes Simulia Corp.: *Theory manual - Version 6.13-1*, Providence, RI, USA, 2013

[33] E.J. BARBERO, Finite Element Analysis of Composite Materials using Abaqus. Composite Materials: Design and Analysis, *Crc. Pr. Inc.*, 2013

[34] C. DAMMANN, R. MAHNKEN, Mesoscopic modelling of the RTM process for homogenization, In: *Coupled Problems in Science and Enfgineering* **VI**, 713-724, Eds. B. Schrefler, E. Onate, M. Papadrakakis, 2015

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON APPLICATION OF THE MESHLESS FINITE DIFFERENCE METHOD TO NUMERICAL HOMOGENIZATION

Irena Jaworska[1]

[1] Institute for Computational Civil Engineering
Cracow University of Technology
Warszawska 24, 31-155 Cracow, Poland
e-mail: irena@L5.pk.edu.pl

Keywords: Numerical homogenization, Meshless methods, Multipoint method, Higher order approximation

Abstract. *The most commonly used engineering tool for numerical analysis of variety of the heterogeneous materials is the Finite Element Method (FEM). However, this paper focuses on the alternative approach based upon the Meshless Finite Difference Method (MFDM). The purpose of the work is to present the some features, as well as the selected results of the numerical homogenization problem in terms of the Meshless FDM.*

The MFDM solution approach and its higher order extensions, e.g. the Multipont meshless method, may be used at both – the macro and the micro levels in the two-scale analysis of heterogeneous materials. The Multipoint MFDM concept is based on raising the order of approximation of the unknown function by introducing additional degrees of freedom in the stencil nodes, taking into account e.g. the right hand side of the considered differential equation. It improves the FD solution without increasing the number of nodes in the mesh and may also be used for a posteriori error estimation of the results.

At the macro level of two-scale modeling, the heterogeneous material with the inclusions spaced periodically was assumed. The values of effective material constants were determined by the homogeneous equivalents calculated at the micro level for a single representative volume element (RVE). The various types of the inclusion locations in the RVE were tested. The scope of numerical analysis also includes the examination of the influence of some factors on solution quality based on the MFDM and the higher order Multipoint MFDM approaches.

All results obtained so far are encouraging. However, it is only the early stage of the research. Further research is planned.

1 INTRODUCTION

Most of the innovative materials, such as composite and structural ones, as well as numerous natural materials have heterogeneous structure. Therefore, prediction of the mechanical properties of those materials is the important part of many fields of engineering, among them the mechanical and civil ones. Various homogenization approaches have been developed last years [2, 3, 4, 10, 13] to obtain the overall properties of heterogeneous materials and to simplify the analysis of engineering problems.

One of the popular multiscale technique provides overall behavior of the heterogeneous materials from known properties of their constituents e.g. fiber or inclusion and matrix through an analysis of a periodic single representative volume element (RVE) at the micro level. At the macro level, on the other hand, the heterogeneous structure of the composite material is replaced by a homogeneous equivalent obtained from RVE.

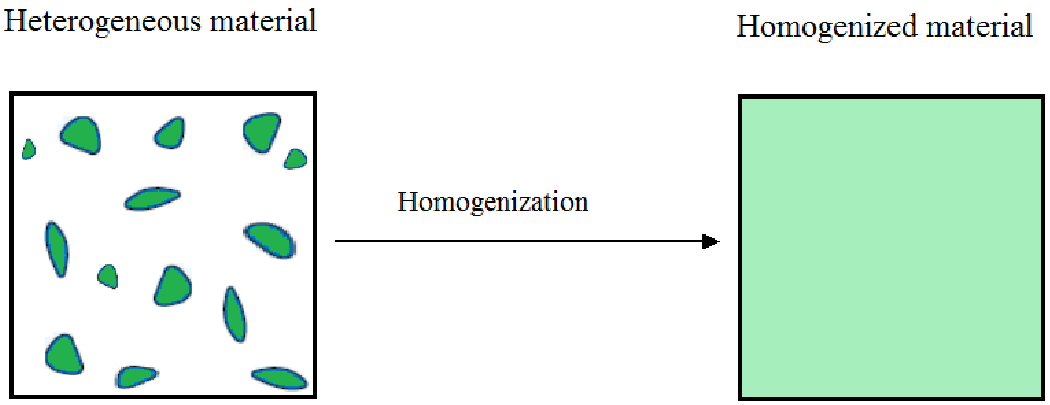


Figure 1: The homogenization approach

There are several methods used for the analysis of heterogeneous materials. The Finite Element Method (FEM) is the most commonly used engineering tool to analyze multiscale problem at both, the macro and the micro levels.

However, this paper focuses on the alternative approach based upon the Meshless Finite Difference Method (MFDM) [6, 12]. The purpose of the work is to present the formulation, as well as the some features and selected results of the numerical homogenization problem in terms of the Meshless FDM and its higher order modification – the Multipoint MFDM.

2 MESHLESS FDM AND MULTIPOINT MFDM APPROACHES

The meshless FDM solution approach [9, 12] and its higher order extensions [6, 11], e.g. the Multipont meshless method [5, 6], may be used at the macro and the micro levels in the two-scale analysis of heterogeneous materials [8, 11]. When compared with the finite element approach, the meshless methods may deal with the unstructured, totally irregular cloud of nodes rather than on elements. In the meshless FDM the local function approximation is built in terms of nodes only. In general, the location and density of nodes depend on the geometry and the physics of the given problem. Consequently, the MFDM allows to avoid the difficulties encountered in the traditional FE technique, such as time-consuming process of generation of the mesh of complex geometries, remeshing, and mesh distortions in large deformation problems.

The basic Meshless (Generalized) FDM scheme [12] consists of the several steps, which are listed below:

- selection of the appropriate boundary value problem formulation: local (strong), global (weak, e.g. variational) , or global-local (e.g. MLPG [1, 7]);
- domain discretization:
  - nodes generation (e.g. by applying the Liszka's type generator [9] based on the nodes density control),
  - domain partition by Voronoi tessellation and Delaunay triangulation,
  - domain topology determination (neighbourhoods);
- optimal stencil (MFD star) selection and generation;
- local approximation by moving weighted least squares (MWLS) method;
- generation of the difference formulas;
- discretization of the boundary conditions;
- generation and solution of the MFD simultaneous equations;
- postprocessing of the results.

The basic MFDM approach has been extended recently to the higher order Multipoint Meshless FDM [5, 6, 7].

The Multipoint MFDM concept is based on raising the order of approximation of the unknown function by introducing additional degrees of freedom in the stencil nodes, taking into account e.g. the right hand side of the considered differential equation (so-called the *specific* approach), or nodal derivatives (the *general* approach). In this way, in the multipoint MFD operators, the combination of values of searched function is equal to the combination of values of additional d.o.f. at all stencil nodes. It improves the FD solution quality without increasing the number of nodes in the mesh.

Let us consider the boundary value problem given in the domain Ω for the n-th order PDE in local (strong) formulation: $\mathcal{L}u = f$ with b.c. $\mathcal{G}u = g$, where $\mathcal{L}, \mathcal{G}$ are differential operators; or in equivalent global (weak) one formulated e.g. as a variational principle: $b(u, v) = l(v), \quad \forall\ v \in V$, where $b -$ is a bilinear functional dependent on the test function v and solution u, $V -$ is the space of test functions, $l -$ is a linear form dependent on v.

In the MFDM solution approach, the standard difference operator based on MFD star may be presented in the following form

$$\mathcal{L}u_i \approx Lu_i \quad \Rightarrow \quad Lu_i = f_i \quad \sum_j c_{ij} u_j = f_i \,. \tag{1}$$

In the multipoint formulation, the MFDM difference operator Lu is obtained by using the Taylor series expansion of unknown function u including the higher order derivatives, and using additional degrees of freedom at nodes [6]. Instead of the function value at the *central node only*, one may apply e.g. a combination of the right-hand side values of the considered equation (specific case, eq.(2)), or nodal derivatives (general case [5], eq. (3)) *at any node* of MFD star.

$$\mathcal{L}u_i \approx Lu_i \quad \Rightarrow \quad Lu_i = Mf_i \quad \sum_j c_{ij} u_j = \sum_j \alpha_{ij} f_j \tag{2}$$

$$\sum_j c_{ij} u_j = \sum_j \alpha_{ij} u_j^{(k)} \tag{3}$$

The following denotations have been assumed here: j – number of a node in a selected stencil, Mf_i – a combination of additional d.o.f. In general, L may be either referred to the left side of differential eqs or to the integrand in the global formulation of b.v. problem, and to the boundary conditions.

The main feature of the multipoint MFDM general case is the relation between the additional d.o.f. (k-th derivative values) and the basic ones (searched function u values) in the whole domain Ω. The required number of such relations for N-dimensional problem can be reduced to N, calculated only the first derivative relations. The higher order derivatives may be developed next using the formulae composition approach. In this way, the general formulation may be used for all types of b.v. problems including the nonlinear one.

3 APPLICATION OF THE MFDM AND ITS EXTENSION TO HOMOGENIZATION

A heterogeneous elastic material with the inclusions spaced periodically was assumed. The length l of the micro scale is assumed to much smaller than the characteristic length L of the macro domain. The macro scale constitutive behaviour is determined by analysis of an RVE subjected to three types (in 2D) of boundary conditions [2, 13]. These numerical tests enable computation of all entries of the homogenized tensor of material properties.

3.1 Formulation at the macro scale

The macro scale problem of the plane stress analysis consists of finding the displacement field $\boldsymbol{u}_0$, as well as the stress $\boldsymbol{\sigma}_0$ and the strain $\boldsymbol{\varepsilon}_0$.

The b.v. problem at the macro level is defined as follows:

$$\begin{cases} \operatorname{div}\boldsymbol{\sigma}_0 = -\boldsymbol{f} & \text{in } \Omega \\ \boldsymbol{\sigma}_0 = \boldsymbol{C}\boldsymbol{\varepsilon}_0 & \text{in } \Omega \\ \boldsymbol{\varepsilon}_0 = \operatorname{sym}\left(\nabla \boldsymbol{u}_0\right) & \\ + \text{boundary conditions} & \end{cases}, \tag{4}$$

where $\boldsymbol{C}$ denotes the matrix of elastic effective material parameters, $\boldsymbol{f}$ – is the field of body forces.

3.2 Formulation at the micro scale

The plane stress analysis is carried out at the micro scale defined over RVE. The boundary value problem on the RVE level may be defined as follows.

The equilibrium state equation is assumed

$$\operatorname{div}\boldsymbol{\sigma} = 0 \qquad \text{in } \Omega\,. \tag{5}$$

The symmetric stress $\boldsymbol{\sigma}$ is related to the strain $\boldsymbol{\varepsilon}$ by the Hook's low

$$\boldsymbol{\sigma} = \boldsymbol{c}\boldsymbol{\varepsilon} \qquad \text{in } \Omega \tag{6}$$

where $\boldsymbol{c}(x,y)$ denotes the tensor of elastic material parameters and $\boldsymbol{\varepsilon} = \operatorname{sym}\left(\nabla \boldsymbol{u}\right)$ is the strain.

The Dirichlet boundary conditions

$$\boldsymbol{u} = \boldsymbol{u}^0 \qquad \text{on } \partial\Omega_u \tag{7}$$

were assumed for the displacement $\boldsymbol{u}$ to model tension in two directions as well as shearing of the RVE.

Effective material properties are determined by using volume average of the stress $<\boldsymbol{\sigma}>$ and the strain $<\boldsymbol{\varepsilon}>$

$$\langle \boldsymbol{\sigma} \rangle = \boldsymbol{C} \langle \boldsymbol{\varepsilon} \rangle , \tag{8}$$

which provide nine equations with nine unknowns.

3.3 Numerical homogenization based on the MFDM approach

The numerical analysis has been based on the discretization of macro and micro structures in terms of the meshless FDM. To define the overall properties of the heterogeneous materials a variational-based approach has been assumed. The MFDM discretization may be based on any irregular clouds of nodes without any imposed structure, like in the FEM, or regularity rule, like in the standard FDM. The cloud of approximation nodes for the RVE may be generated either independently on the inclusions distribution or may be adjusted to them.

The next principles in the application of the MFDM and its higher order extensions to the numerical homogenization are assumed [7]:

- In case of FEM, both the trial u and the test v functions are approximated by means of the same basis functions (Bubnov-Galerkin approach). In case of MFDM, two different bases (Petrov-Galerkin approach) may be applied;
- The approximation of unknown function u and its derivatives is provided by the appropriate meshless FD (1) or multipoint MFD (3) operators;
- In case of the weak formulation the numerical integration is additionally required in the MFDM approach. It is performed by the Gauss qudrature on subdomains such as the Delaunay triangles (integration between nodes) or Voronoi polygons (integration around nodes);
- The values of the test function v and its derivatives at Gauss point P_k are calculated by MWLS approximation based on the MFD star, or interpolation based on integration subdomain

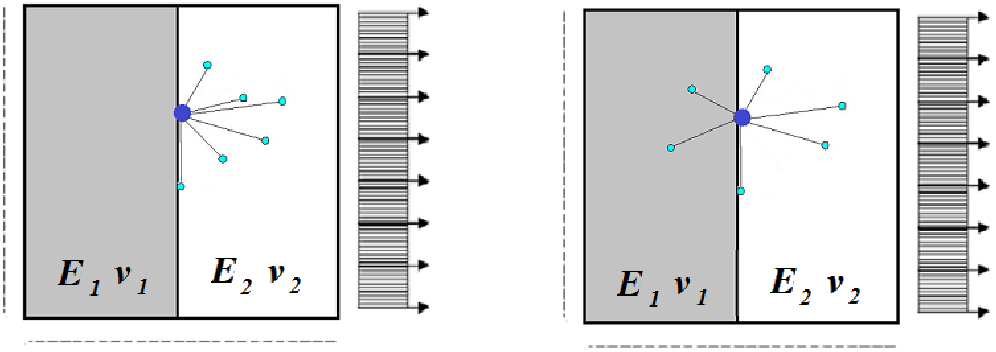
$$v(P_k) \approx \sum_j b_{j(k)} v_j , \qquad v_x\big|_{P_k} \approx \sum_j {}^x b_{j(k)} v_j , \qquad v_y\big|_{P_k} \approx \sum_j {}^y b_{j(k)} v_j . \tag{9}$$

The simultaneous equations are generated directly from the variational principle by aggregation (from each integration cell), and taking into account MFD or multipoint MFD operators [8]. After obtaining the nodal values of the displacement u_x and u_y, the effective values of the material parameters $\boldsymbol{C}$ may be evaluated by solving eq.(8).

4 SOME NUMERICAL RESULTS

The 2D linear elasticity problem for two-phase composite material is considered in order to verify the numerical homogenisation by meshless FDM and multipoint MFDM solution approaches. The micro scale model is set up based on the RVE technique.

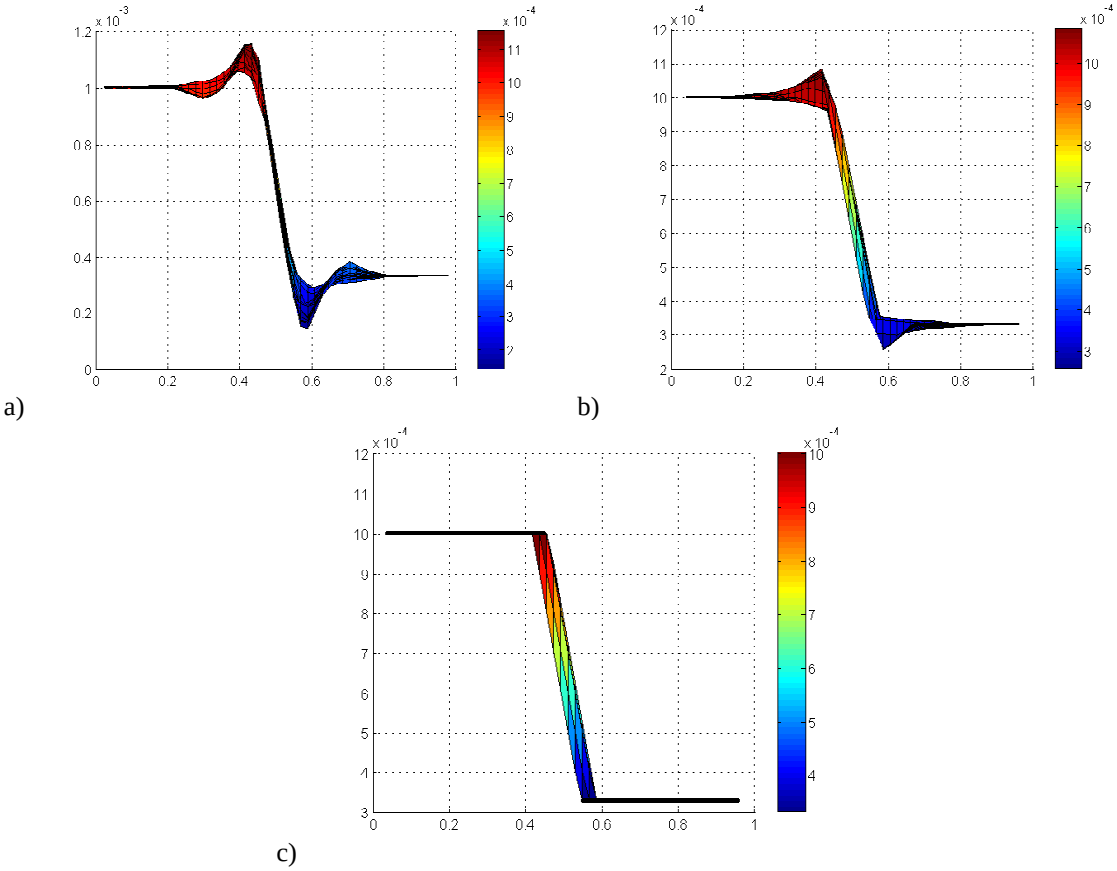
For the first simple benchmark test of MFD star configuration influence, the RVE with two equal part is assumed (Fig. 2). The elastic modules and Poisson's ratios for the left rectangular part are assumed as $E_1 = 1 \cdot 10^5$ MPa, $v_1 = 0.1$ and respectively for the right part: $E_2 = 3 \cdot 105$ MPa, $v_2 = 0.3$.



a) b)

Figure 2: The MFD star (stencil) configuration in RVE (test 1)

In the case, when the MFD star (stencil) configuration does not correspond to the distribution of the different material part (nodes of the stencil belong to the both type of material, Fig.2b), some oscillations near the boundary of inclusion are observed (Fig. 3a,b). The number of Gauss points in the e.g. Delaunay triangles, and stencil size may have an influence on the amplitude of the oscillations. However, when the MFD star is adjusted to the inclusions distribution (Fig. 2a) – the exact solution has been obtained (Fig. 3c).

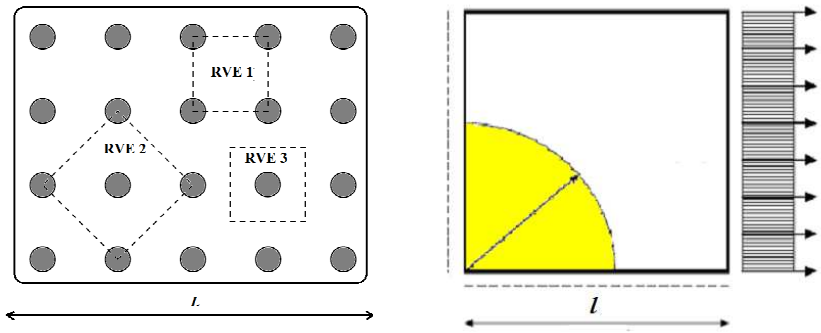


a) b)

c)

Figure 3: The strain $\boldsymbol{\varepsilon}_{xx}$ obtained by meshless FDM for test 1 in the case:
a) stencil does not correspond to the inclusion, 4 Gauss point integration; b) stencil does not correspond to the inclusion, 1 Gauss point integration; a) stencil is adjusted to the inclusion

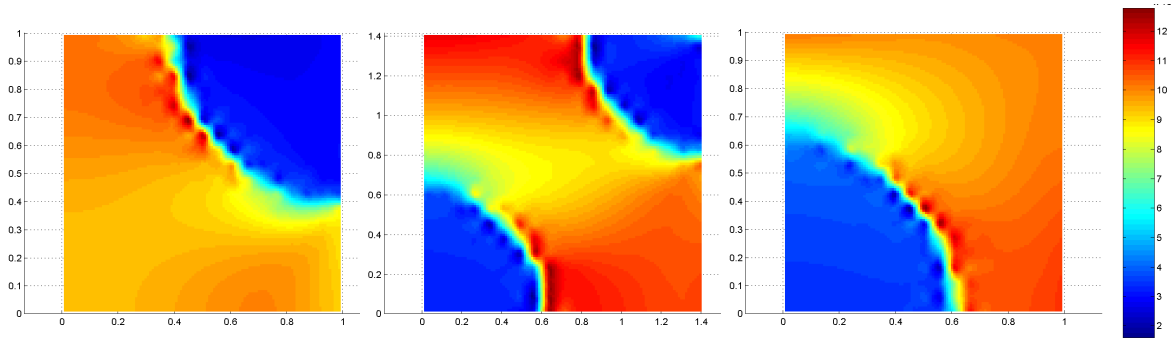
For the second test, the structure with uniformly distributed circular inclusions with the same radii is assumed. Therefore, heterogeneous material could be represented by one of the RVE models as presented in Fig. 4a. Due to symmetric RVE models, the quarter of the RVE may be considered only. The scheme of such e.g. RVE3 with appropriative boundary conditions is presented in Fig. 4b.



a) b)

Figure 4: Distribution of the inclusions on macro (a) and micro (b) scales

In all cases, the regular uniform meshes are applied in the MFDM solution approach. Numerical results of effective material constants obtained for all types of RVE model are close enough (Fig. 5).



a) b) c)

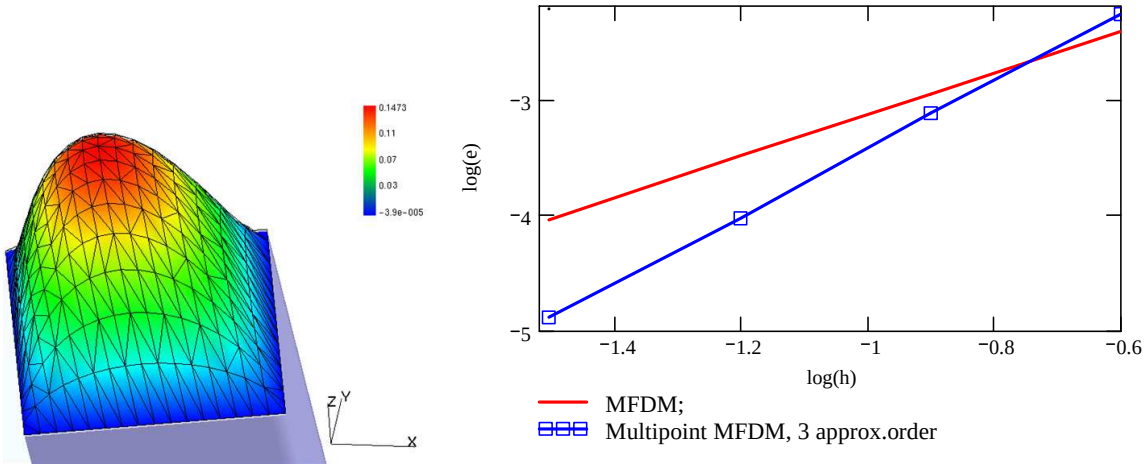
Figure 5: MFDM results of strain $\boldsymbol{\varepsilon}_{xx}$ obtained for RVE1, RVE2, and RVE3

The numerical analysis may be returns to the macro scale after determination of the effective material constants (beside the elastic modulus and Poisson's ratio, the value of the Kirchhoff parameter G may be evaluated also). These effective values homogenize the original heterogeneous material. Consequently, meshless FDM or higher order multipoint solution approach may be applied, e.g. for the de Saint-Venant torsional problem of the heterogeneous prismatic bar with the rectangular cross-section assumed as the next benchmark test. The local formulation in the form of the Poisson's equation with the essential boundary conditions was applied first

$$\begin{cases} \nabla^2 \Phi = -2G\theta, & \text{in} \quad \Omega \\ \Phi = 0, & \text{on} \quad \partial\Omega \end{cases}, \tag{10}$$

followed by variational form, where Φ – is the Prandtl stress function, $G\theta$ – torsional stiffness, Ω – domain of the bar cross-section.

The comparison of the MFDM and higher order multipoint MFDM convergence rates is presented in Fig. 6.



a) b)

Figure 6: Twisting of square cross-section prismatic bar: a) Prandtl stress function and b) its solution convergence rate for series of regular meshes. The exact error for the standard meshless FDM, and multipoint MFDM 3-rd approximation order

5 FINAL REMARKS

Formulation and some features of the meshless FDM and its higher order extension – the multipoint meshless FDM for the numerical homogenization problem of heterogeneous materials with periodic structure was presented. The research has been focused on the two-scale analysis based on the single representative volumetric element RVE. The stencil configuration influence was investigated. Several tests confirmed, that adjusting of the MFD star to the inclusion distribution may have an influence on the solution in terms of the meshless FDM. The effective values of the tensor of elastic material parameters were determined for the three RVE types. Results for different RVE types were compared with each other at the micro level. The analysis of the convergence results of the MFDM and multipoint MFDM at the macro level confirmed the better quality of the solution obtained by using the higher order extension of the meshless FDM.

Although obtained so far results are encouraging, it is only the early stage of the research. Therefore, further research is planned.

Acknowledgment

This research was supported by the National Science Center under grant 2011/01/B/ST6/07312.

REFERENCES

[1] S.N. Atluri, S. Shen, The Meshless Local Petrov-Galerkin (MLPG) Method, Tech Science Press, 2002

[2] W. Cecot, M. Oleksy, High order FEM for multigrid homogenization, Computers and Mathematics with Applications, 70, 1391–1400, 2015

[3] J. Fish, Practical Multiscaling, Wiley, 2013

[4] M.G.D. Geers, V.G.Kouznetsova, W.A.M. Brekelmans, Multi-scale first-order and second-order computational homogenizastion of microstructures towards continua. International Journal for Multiscale Computational Engineering, 1(4): 371–385, 2003.

[5] I. Jaworska, On the ill-conditioning in the new higher order multipoint method. Computers and Mathematics with Applications, 66(3): 238-249, 2013

[6] I.Jaworska, J. Orkisz, Higher order multipoint method – from Collatz to meshless FDM, Eng Anal Bound Elem, 50: 341–351, 2015

[7] I. Jaworska, J. Orkisz On the multipoint meshless FD method using the local Petrov–Galerkin approach. In: Eberhardsteiner J., Böhm H.J., Rammerstorfer F.G., editors. ECCOMAS-2012. Vienna; 6582–91, 2012

[8] I. Jaworska, S. Milewski, On two-scale analysis of heterogeneous materials by means of the meshless finite difference method, International Journal for Multiscale Computational Engineering, in print

[9] T. Liszka, J. Orkisz, Finite difference method for arbitrary irregular meshes in nonlinear problems of applied mechanics, IV SmiRt, San Francisco, 1977.

[10] C. Miehe, A. Koch, Computational micro-to-macro transitions of discretized microstructures undergoing small strains, Arch. Appl. Mech. 72, 300–317, 2002

[11] S. Milewski, Introduction to the numerical homogenization by means of the meshless finite difference method with the higher order approximation. Computer Methods in Mechanics CMM, Warsaw, Poland, 2011.

[12] J. Orkisz, Finite Difference Method (Part III). Handbook of Computational Solid Mechanics, ed. M.Kleiber, Springer-Verlag, Berlin, 336–431, 1998.

[13] T.I. Zohdi, P. Wriggers, An Introduction to Computational Micromechanics, in: Lecture Notes in Applied and Computational Mechanics, vol. 20, Springer, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A MESOSCALE STUDY OF DEFORMATION AND FAILURE OF ANGLE-PLY LAMINATES UNDER TENSILE LOADING BY MEANS OF NUMERICAL HOMOGENIZATION

Marek Romanowicz

Bialystok University of Technology, Department of Mechanical Engineering
ul. Wiejska 45C, 15 351 Bialystok, Poland
e-mail: m.romanowicz@pb.edu.pl

Keywords: numerical homogenization, polymer-matrix composites, finite element analysis.

Abstract. *A multi-ply unit cell model for predicting the tensile response of a ±45 angle-ply laminate based on the finite element method is presented. Two different interfiber failure mechanisms leading to matrix cracking are reproduced in the simulations by using appropriate constitutive equations. The effect of fiber-to-fiber interaction within the ply is incorporated in the model by applying a hexagonal array of fibers. The predictions from this model are compared with experimental data available in the literature, and are found to be in good agreement.* .

1 INTRODUCTION

Recently, researchers have shown an increased interest in studying the deformation and failure of composite laminates by using multi-ply unit cell models [1-8]. In the present paper, the mesoscale approach based on a rhombohedral unit cell proposed by Zhang et al. [6] is extended. This approach is based on an assumption that the structure of the whole laminate can be idealized at the ply scale as a periodic array of repeated unit cells. Some new issues, like fiber/matrix interface debonding, development of plastic strain in the matrix and hexagonal fiber packing are introduced in the mesoscale model of a ±45 angle-ply laminate. The objective of this paper is to show that an analysis of local stresses in the mesoscale model is capable of identifying the failure mechanism and predicting the tensile behavior of a ±45 angle-ply laminate.

2 MESOSCALE FINITE ELEMENT MODEL

As shown in Fig.1, two coordinate systems are used to analyze angle-ply laminates. The first one is a orthogonal coordinate system (x_1–x_2–x_3), which is used to determine the direction of loading. The second one is a skew coordinate system (x_1^s–x_2^s–x_3^s), which is used to define the unit cell geometry. This coordinate system is set up such that the x_1^s and x_2^s axes lie in the plane of the laminate along the fiber directions and the x_3^s axis is perpendicular to it. It should be noted that the fiber orientation angle θ is measured with respect to the x_1 axis of the orthogonal coordinate system. A unit cell of an angle-ply laminate consists of three rhombohedrons which represent three successive plies with various fiber orientations. The structure of an ±θ angle–ply laminate can be viewed at the ply scale in the skew coordinate system (x_1^s–x_2^s–x_3^s) as a periodical array of plies containing a periodic subarray of fibers. Thus, a unit cell of an angle-ply laminate is built with unit cells of plies. In this study, the fibers are distributed uniformly throughout the ply in the form of a hexagonal array.

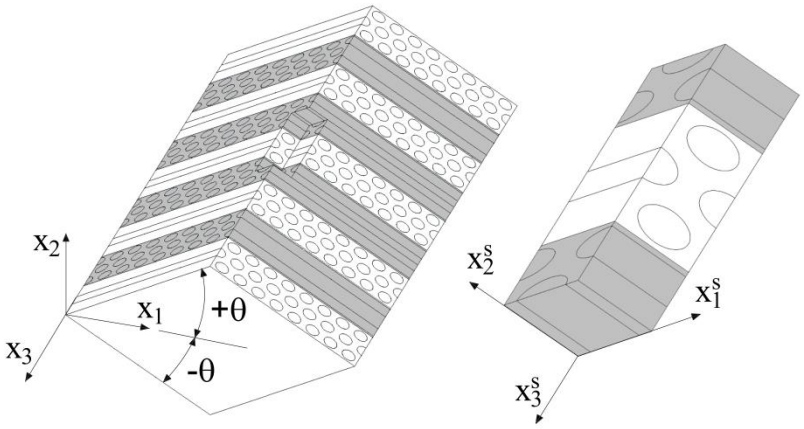


Figure. 1: Geometry of a rhombohedral unit cell.

In this paper, 3-D multi-ply mesoscale models of a ±45° angle-ply laminate subjected to tension along the x_2 direction are examined. The material properties of fibers and matrix (Table 1) correspond to those reported in the World Wide Failure Exercise [9] for Silenka E-glass 1200tex fibers and MY750/HY917 epoxy matrix. The dimensions of rhombohedral unit cells were chosen so as to accommodate the radius of fiber r_f = 8 μm, and to respect the fiber volume content V_f = 60%. The full rhombohedrons were meshed with three–dimensional eight–node elements (SOLID185) and solved by using Ansys finite element code [10].

Silenka E-glass fiber			MY750/HY917/DY063 epoxy matrix					
E_f [GPa]	ν_f	X_f [MPa]	E_m [GPa]	ν_m	k [MPa]	μ	Y_m [MPa]	S_m [MPa]
74	0.2	2150	4.5	0.35	43.35	0.1	100	87

fiber/matrix interface					
k_n [GPa/m]	k_t [GPa/m]	G_n^c [J/m^2]	G_t^c [J/m^2]	σ_n^c [MPa]	τ_t^c [MPa]
0.1x10^9	0.1x10^9	15	30	15	30

Table 1: Mechanical properties of the lamina constituents.

2.1 Numerical homogenization

Periodic deformation of the rhombohedral unit cell is controlled by macroscopic strains by using the following boundary conditions [11]

$$\begin{aligned} u_i^s(a_1, x_2^s, x_3^s) - u_i^s(-a_1, x_2^s, x_3^s) &= 2a_1\varepsilon_{i1}^s, \\ u_i^s(x_1^s, a_2, x_3^s) - u_i^s(x_2^s, -a_2, x_3^s) &= 2a_2\varepsilon_{i2}^s, \\ u_i^s(x_1^s, x_2^s, a_3) - u_i^s(x_1^s, x_2^s, -a_3) &= 2a_3\varepsilon_{i3}^s, \end{aligned} \tag{1}$$

where i = 1, 2, 3, and $2a_1$, $2a_2$, $2a_3$ are the dimensions of the unit cell in the skew coordinate system, ε_{ij}^s are the components of the applied strain in the skew coordinate system. In order to apply the periodic boundary conditions to the finite element model, the displacements in the skew coordinate system u_1^s, u_2^s, u_3^s must be rewritten in terms of the displacements in the orthogonal coordinate system u_1, u_2, u_3 by using the following relationships [11]

$$\begin{Bmatrix} u_1^s \\ u_2^s \\ u_3^s \end{Bmatrix} = \frac{1}{\sin 2\theta} \begin{bmatrix} \sin\theta & -\cos\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & \sin 2\theta \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \end{Bmatrix}. \tag{2}$$

The components of macroscopic stress corresponding to the applied strain can be calculated from

$$\bar{\sigma}_{ij} = \frac{1}{V}\int_V \sigma_{ij}\, \mathrm{d}V \tag{3}$$

where V is the volume of the rhombohedral unit cell.

2.2 Constitutive equations of matrix and interface

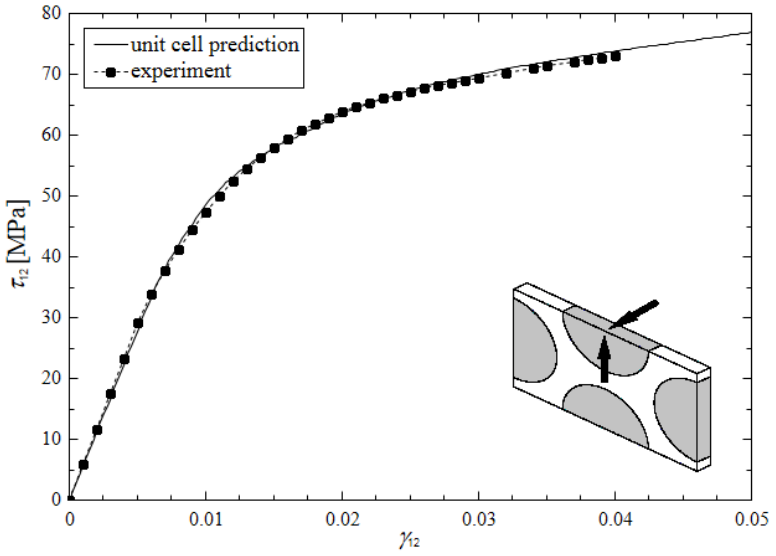
The epoxy matrix is modeled within the framework of the finite deformations as a elasto–plastic solid which hardens isotropically. The deformation of polymeric materials is highly sensitive to the hydrostatic pressure and plastic flow of these materials can exhibit plastic dilatancy. To address this requirement, the Drucker–Prager plasticity model [12], which incorporates the linear dependence on the hydrostatic stress, is used. In terms of the first invariant of stress I_1 and the second invariant of the deviatoric part of stress J_2, the yield function is given as

$$\mathrm{f} = (\mu\, I_1 / 3) + \mathrm{sqrt}(J_2) - k\,, \tag{4}$$

where μ is the pressure sensitivity factor, k is the flow stress of the material under pure shear. Experiments showed that the pressure–sensitivity factor μ ranges from 0.10 to 0.25 for polymers [13, 14]. The Drucker–Prager plasticity model with μ = 0.1 and k = 43.30 MPa was used

to study of the matrix cracking damage mechanism. An associative flow rule is used to compute the direction of plastic flow [10].

(a)



(b)

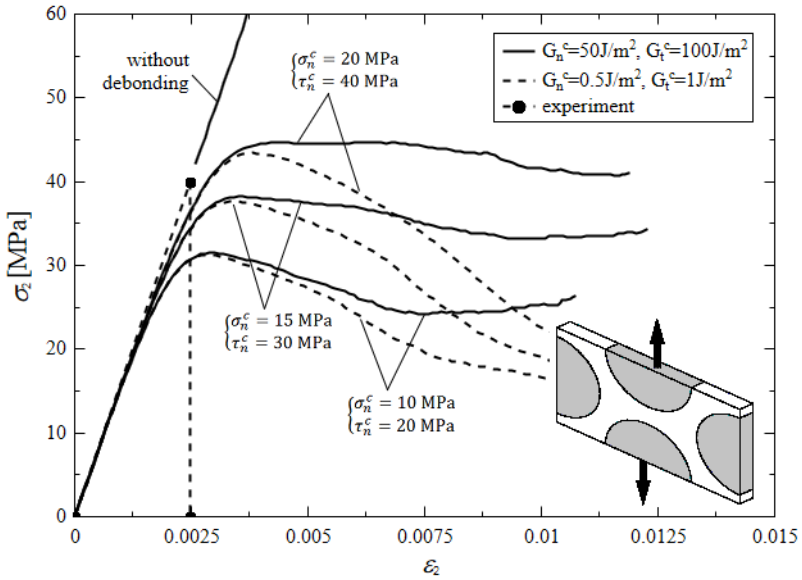


Figure 2: Calibration of the mesoscale model: (a) identification of the matrix plasticity model (experimental data from [9]), (b) identification of the fiber/matrix interface model (experimental data from [9]).

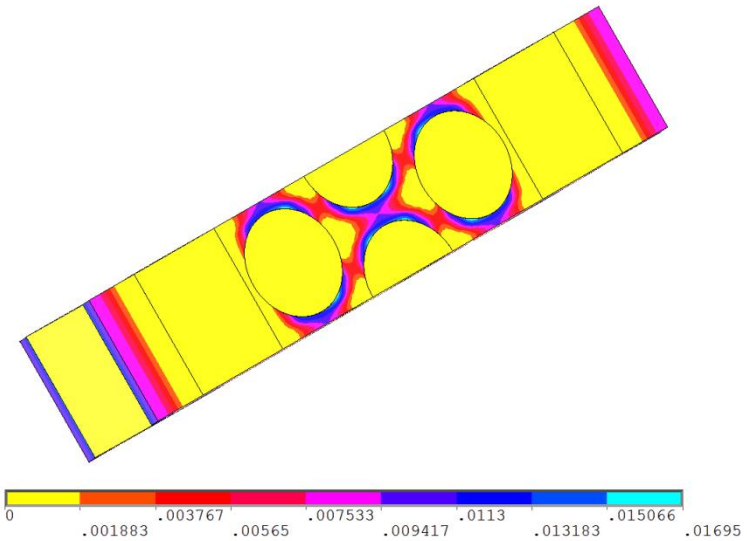
Experimental investigations of failure in transversely-loaded fiber-reinforced polymers have revealed that another damage mechanism is debonding occurring at the fiber/matrix interface [15,16]. For the fiber/matrix interface failure, a cohesive zone model is employed, in which the constitutive equations of the interface relate the normal σ_n and tangential τ_t cohesive tractions to the normal u_n and tangential u_t displacement jumps and a scalar damage variable d, through [17]

$$\sigma_n = k_n u_n (1-d), \quad \tau_t = k_t u_t (1-d), \tag{5}$$

where k_n , k_t are initial contact stiffnesses in the normal and tangential direction, respectively. The variable d represents the loss of stiffness and it is a function of both displacement jumps. The variable d takes values from 0 to 1. Relationships (5) demonstrate linear elastic loading region followed by linear softening region. When $d = 0$, the cohesive elements are closed and the tractions increase linearly up to their maximum values σ_n^c, τ_t^c in the normal and tangential direction, respectively. When $0 < d < 1$, these elements begin to open and the tractions decrease linearly. When $d = 1$, the tractions are zero and the cohesive elements are completely broken. To define the completion of fracture in the cohesive zone model, a linear energy criterion is used [17]

$$(G_n / G_n^c) + (G_t / G_t^c) = 1, \tag{6}$$

(a)



(b)

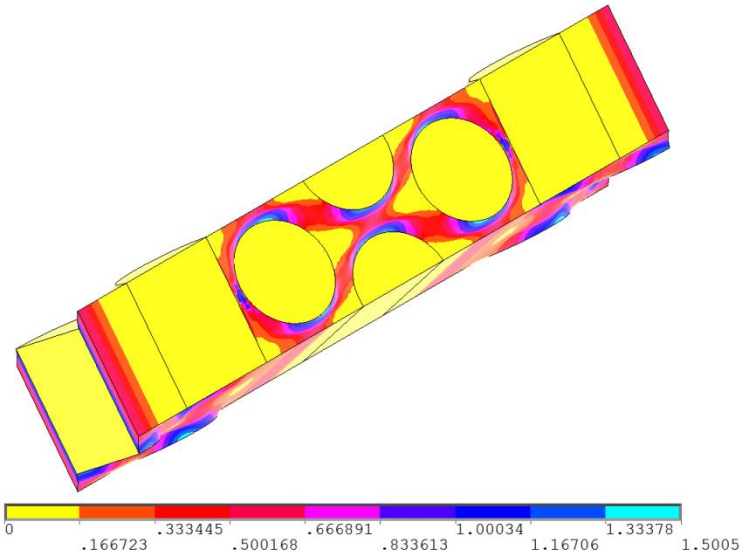


Figure 3: Contour plots of the effective plastic strain in the $\pm 45^o$ angle-ply laminate at the critical stage of deformation from the mesoscale model with (a) imperfect and (b) perfect fiber/matrix interface conditions.

where G_n, G_t denote energy release rates for mode I fracture and mode II fracture, respectively and G_n^c and G_t^c correspond to the interfacial fracture energies. After debonding is complet-

ed, the interface surface interaction is governed by standard contact constraints for normal and tangential directions. A cohesive layer consisted of contact elements with eight nodes (CONTA174, TARGE170) was introduced between the fibers and the matrix to reproduce the fiber–matrix debonding. The glass fibers are assumed to be linear elastic.

2.3 Calibration of the Drucker–Prager plasticity model

In this paper, the hardening curve of epoxy matrix is extracted from a hexagonal unit cell subjected to in–plane shear loading. The non–linearity of epoxy matrix is identified such that prediction obtained from the hexagonal unit cell matches the measured in–plane shear response of E-glass/MY750 epoxy unidirectional lamina reported in [9]. Fig.2a shows a comparison of the measured shear response and that from the hexagonal unit cell model. Agreement with experimental data is quite good.

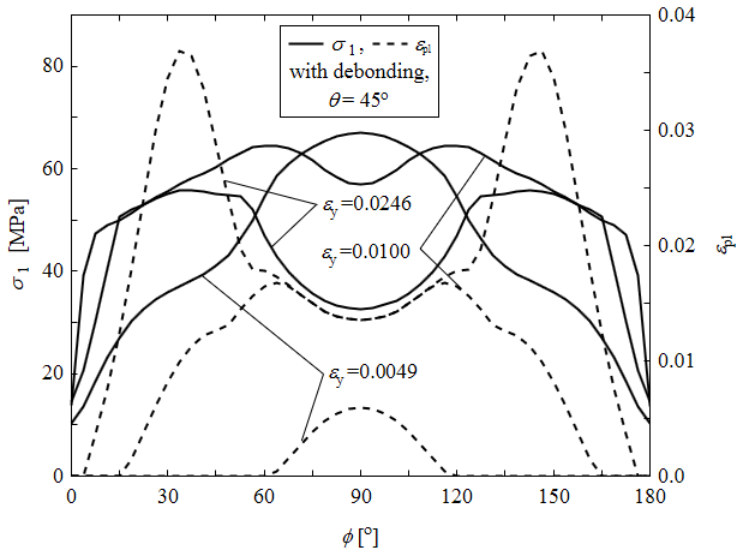


Figure 4: Angular distributions of the first principal stress and the effective plastic strain at the fiber/matrix interface at three successive stages of deformation from the mesoscale model with imperfect fiber/matrix interface.

2.4 Calibration of the interface model

In order to ascertain the cohesive properties for the interface under study, an analysis of debonding under transverse tension was carried out for various interfacial cohesive strengths and fracture energies. It was assumed that the ratio of the tensile interfacial strength to the shear interfacial strength, σ_n^c / τ_t^c and the ratio of the opening component to the sliding component of the interfacial fracture energy, G_n^c / G_t^c are 0.5. The macroscopic tensile stress–strain curves calculated for four different levels of interfacial strength (σ_n^c / τ_t^c = 10/20, 15/30, 20/40 and ∞) and two different levels of interfacial fracture energy (G_n^c / G_t^c = 0.5/1, 50/100) are shown in Fig.2b. It can be observed from this figure that an increase in the interfacial strength produces an increase in the transverse tensile strength of composite, and in turn, a decrease in the interfacial fracture energy leads to a larger decrease in macroscopic stress at the post critical stage. The strength of the interface under study was calibrated such that the peak of the transverse stress-strain curve obtained from the hexagonal unit cell matches the measured transverse tensile strength of the composite reported in [9]. Due to lack of experimental data on the softening behavior of the entire composite at the macroscale, the interfacial fracture energy was assumed to be of the same order of magnitude as that measured in single-

fiber tests reported in [18]. Finally, the interface with σ_n^c =15MPa, τ_t^c = 30MPa, G_n^c = 15J/m^2 and G_t^c = 30J/m^2 was chosen for the rest of the study.

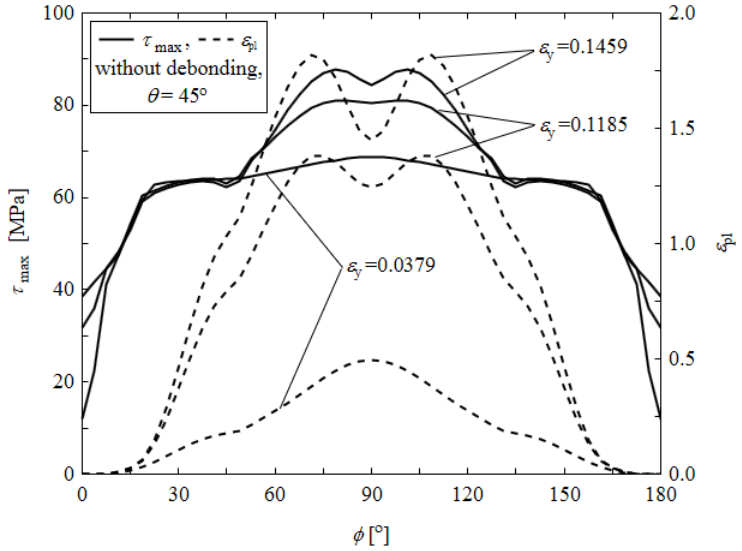


Figure 5: Angular distributions of the maximum shear stress and the effective plastic strain at the fiber/matrix interface at three successive stages of deformation from the mesoscale model with perfect fiber/matrix interface.

3 FAILURE MECHANISMS

Each ply in an angle-ply laminate subjected to tension is loaded by a combination of normal and shear stresses. The value of these stresses depends on the ply orientation angle θ. Both shear and tensile loading modes can induce initiation and growth of cracks in the matrix along the fiber direction. For the sake of simplicity, an interaction between the two fracture modes is omitted in the present paper. In order to identify the failure mechanism of angle-ply laminates under tensile loading, an analysis of the local stresses in the unit cell models for different ply orientations is performed by the finite element method. For this purpose, it is assumed that the initiation of cracks in the matrix occurs if the maximum shear stress in the matrix or the first principal stress in the matrix goes beyond the corresponding ultimate strengths of the matrix

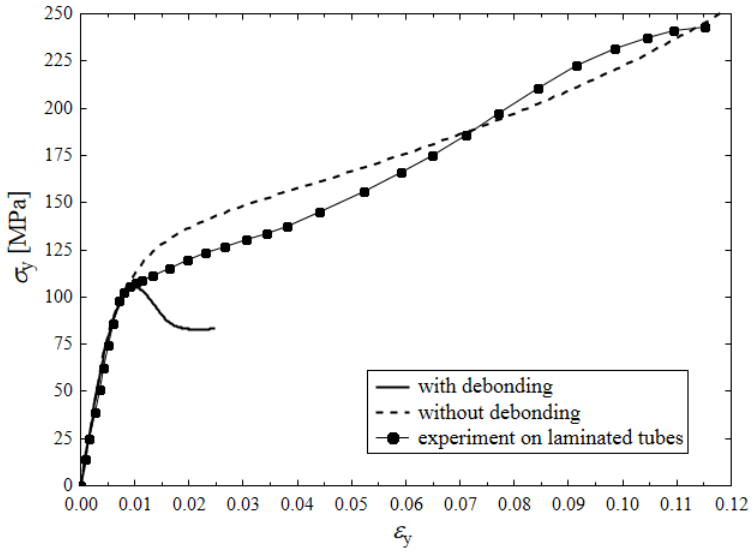
$$\tau_{max} \geq S_m, \quad \text{or} \quad \sigma_1 \geq Y_m. \tag{7}$$

Tension tests carried out on microscopic specimens indicate that the tensile strength of epoxy polymer Y_m is much higher than that obtained with standard macroscopic specimens. The use of microscopic data is essential for the current analysis. According to works by Fiedler et al., [19], Hobbiebrunken et al. [20], the tensile strength of the epoxy polymer obtained from microscopic tests can exceed 100 MPa, while macroscopic tests yield the value of 80 MPa. To the author's knowledge, there are no micro-shear tests for epoxy polymers. The shear strength of the epoxy polymer S_m obtained from macroscopic tests is 70 MPa [19, 20]. However, it is reasonable to assume that the true shear strength of the epoxy matrix is also higher. The shear strength of the epoxy matrix back-calculated from tension tests of the ±55 angle-ply laminate is 87 MPa [8].

4 RESULTS

Depending on the type of the fiber/matrix interface, two various mesoscale models of the ±45 angle-ply laminate were considered, namely models with the perfect and imperfect fiber/matrix interface conditions. Fig.3 shows distributions of the equivalent plastic strain in these models for critical loads. In the case of the imperfect interface, the critical load corresponds to a limit beyond which the fiber/matrix interface cracks are fully opened ($\varepsilon_y^c = 0.01$). In turn, in the case of the perfect interface, it corresponds to a limit beyond which the shear strength of the matrix is attained ($\varepsilon_y^c = 0.1185$). The most intense plastic deformation in the matrix is localized in shear bands which develop along planes containing the fiber direction. Thus, both models predict that cracking of the matrix takes place along the fiber direction within the shear bands.

(a)



(b)

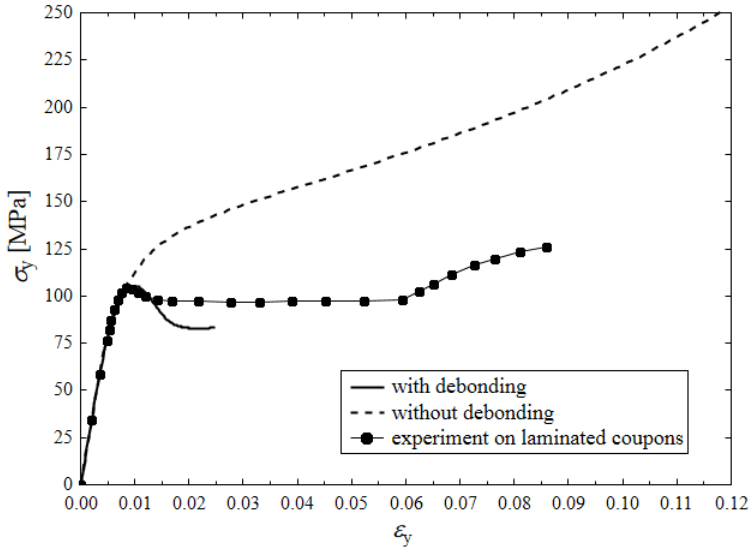


Figure 6: Comparison of the stress-strain curves of the ±45° angle-ply laminate obtained from different fiber/matrix interface conditions with experimental data: (a) tensile tests from [21] carried out on laminated tubes, (b) tensile tests from [22] carried out on laminated coupons.

In order to check the reliability of the opening mode of fracture, an analysis of the first principal stress σ_1 and the equivalent plastic strain ε_{pl} in the matrix is performed by using the mesoscale model with the imperfect fiber/matrix interface. The distributions of these quantities at the fiber/matrix interface are presented in Fig.4 at three successive stages of deformation corresponding to the beginning of inelastic behavior, the limit load and the softening regime. The polar angle at the fiber/matrix interface ϕ is measured from the normal to the interlaminar interface. It is interesting to note that the tensile stress in the matrix decreases at the softening regime and it is below the tensile strength (Y_m = 100 MPa). In contrast to the tensile stress, the plastic strain in the matrix increases constantly with increasing applied strain. This means that the condition for crack growth in the matrix under the opening mode of fracture cannot be satisfied.

In order to validate the shearing mode of fracture, an analysis of the maximum shear stress τ_{max} and the equivalent plastic strain ε_{pl} in the matrix is performed by using the mesoscale model with the perfect fiber/matrix interface. The distributions of these quantities at the fiber/matrix interface are presented in Fig.5 at three successive stages of deformation. It can be observed in Fig.5 that both the shear stress τ_{max} and the plastic strain ε_{pl} in the matrix increase with increasing applied strain. Thus, the shear stress in the matrix may go beyond the shear strength of the epoxy matrix (S_m = 87 MPa). It is also interesting to note that the mesoscale model with the perfect fiber/matrix interface exhibits a very ductile behavior. At the critical stage of deformation, the plastic strain in the matrix exceeds 140% .

Predictions of the stress-strain behavior for the perfect and imperfect fiber/matrix interface conditions are shown in Figs.6a and 6b. Two different experimental curves reported in the literature [21-22] for the ±45 angle-ply laminate are included for quick reference in these figures. The two experimental curves of the ±45 angle-ply laminate show a very non-linear behavior prior to ultimate failure. This behavior is manifested by a hardening region at high strains. The observed increasing stiffness can be attributed to a rotation of fibers due to shear stress. As the ply shear stress increases, the matrix is plasticized and the orientation of the fibers became more aligned with the loading axis, which leads to an increase in load carried by fibers. The fibers are capable of carrying more load only if the displacement compatibility between matrix and fibers exists. Thus, the experimental curves of the ±45 angle-ply laminate obtained from both tests suggest that fiber-matrix debonding has a negligible influence on the deformation of this laminate under tensile loading. It can be observed in Fig.6a that the stress-strain curve obtained from the mesoscale model with the perfect fiber/matrix interface match quite well with experimental data. In turn, the experimental curve of the ±45 angle-ply laminate obtained from tests on laminated coupons exhibits a maximum and a plateau before a hardening region. It can be observed in Fig.6b that the maximum of the stress-strain curve obtained from the mesoscale model with the imperfect fiber/matrix interface overlaps with the maximum from the tests. Thus, this model may be used to predict the beginning of the plateau occurring during tensile tests carried out on laminated coupons.

5 CONCLUSIONS

The efficiency of two types of mesoscale models without and with fiber/matrix debonding for simulating the deformation and failure mechanisms in a ±45 angle-ply laminate under tensile loading was evaluated. The results from both models suggest that failure of the laminate is dominated by the matrix shearing. The first type was found to reproduce well the tensile response of the laminate obtained from laminated tubes, and in turn, the second type was able to predict the beginning of the plateau occurring during tensile tests carried out on laminated coupons.

ACKNOWLEDGMENT

The financial support of the National Science Centre of Poland under contract DEC–2011/03/D/ST8/04817 is thankfully acknowledged.

REFERENCES

[1] M. Herráez, D. Mora, F. Naya, C.S. Lopes, C. González, J. LLorca, Transverse cracking of cross-ply laminates: A computational micromechanics perspective. *Compos. Sci. Technol.*, **110**, 196–204, 2015.

[2] G. Soni, R. Singh, M. Mitra, B.G. Falzon, Modelling matrix damage and fibre–matrix interfacial decohesion in composite laminates via a multi-fibre multi-layer representative volume element (M2RVE). *Inter. J. Solids Struct.*,**51**, 449–461, 2014.

[3] W.H. Ng, A.G. Salvi, A.M. Waas, Characterization of the in-situ non-linear shear response of laminated fiber-reinforced composites. *Compos. Sci. Technol.*, **70**, 1126–1134, 2010.

[4] T, Matsuda, D. Okumura, N. Ohno, M. Kawai, Three-dimensional microscopic interlaminar analysis of cross-ply laminates based on a homogenization theory. *Int. J. Solids Struct.*, **44**, 8274–8284, 2007.

[5] Y. Zhang, Z. Xia, F. Ellyin, Viscoelastic and damage analysis of fibrous polymer laminates by micro/meso-mechanical modeling. *J. Compos. Mater.*,**39**, 2001-2022, 2005.

[6] Y. Zhang, Z. Xia, F. Ellyin, A 3D meso-scale amalysis of angle-ply laminates. *Mech. Adv. Mater. Struct.*, **20**, 801-810, 2013.

[7] F. Ellyin, Y. Zhang, Z. Xia, Meso-scale analysis of angle-ply laminates. *Proc. Eng.*, **10**, 63–68, 2011.

[8] M. Romanowicz, A mesoscale study of failure mechanisms in angle-ply laminates under tensile loading. *Composites Part B*, **90**, 45-57, 2016.

[9] P.D. Soden, M.J. Hinton, A.S. Kaddour, Lamina properties, lay-up configurations and loading conditions for a range of fibre-reinforced composite laminates. *Compos. Sci. Technol.*, **58**, 1011–1022, 1998.

[10] Ansys. *Theory reference for the mechanical APDL and mechanical applications*, Release 14.1. Canonsburg PA: Ansys, 2012.

[11] Z. Xia, Y. Zhang and F. Ellyin Unified periodical boundary conditions for representative volume elements of composites and applications. *Int. J. Solids Struct*., **40**, 1907–1921, 2003.

[12] D.C. Drucker, W. Prager. Soil mechanics and plastic analysis for limit design. *Q. Appl. Math.*, **10**, 157–165, 1952.

[13] A.J. Kinloch, R.J. Young, *Fracture behavior of polymers*. Netherlands: Elsevier, 1983.

[14] R. Quinson, J. Perez, M. Rink, A. Pavan, Yield criteria for amorphous glassy polymers. *J. Mater. Sci.*, **32**, 1371–1379, 1997.

[15] B. Fiedler, C. De Jong, T. Hobbiebrunken, M. Hojo, K. Schulte, Micro/macro-mechanical approach of first ply failure in CFRP. *J. Mater. Sci.*, **41**, 6760–6767, 2006.

[16] D.C. Foster, G.P. Tandon, M. Zoghi, Evaluation of failure behavior of transversely loaded unidirectional model composites. *Exp. Mech.* **46**, 217–243, 2006.

[17] G. Alfano, M.A. Crisfield, Finite element interface models for the delamination analysis of laminated composites: mechanical and computational issues. *Int. J. Numer. Method. Eng.*, **50**, 1701–36, 2001.

[18] J. Varna, L.A. Berglund, M.L. Ericson, Transverse single-fibre test for interfacial debonding in composites: 2. Modelling. *Composites Part A*, **28**, 317–26, 1997.

[19] B. Fiedler, M. Hojo, S. Ochiai, K. Schulte, M. Ando, Failure behavior of an epoxy matrix under different kindsof static loading. *Compos. Sci. Technol.*, **61**, 1615–1624, 2001.

[20] T. Hobbiebrunken, B. Fiedler, M. Hojo, M. Tanaka, Experimental determination of the true epoxy resin strength using micro-scaled specimens. *Compos Part A*, **38**, 814–818, 2007.

[21] F.A.R. Al-Salehi, S.T.S. Al-Hassani, M.J. Hinton, An Experimental investigation into the strength of angle ply GRP Tubes under High Rate of Loading. *J. Compos. Mater.* **23**, 288-305, 1989.

[22] A. Rotem, Z. Hashin, Failure modes of angle ply laminates. *J. Compos. Mater.*, **9**, 191-206, 1975.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MODELING OF PHASE CHANGES IN MICRO-DOMAIN INDUCED BY AN ULTRASHORT LASER PULSE

Ewa Majchrzak[1], Lukasz Turchan[2], and Jolanta Dziatkiewicz[2]

[1] Silesian University of Technology
44-100 Gliwice, Konarskiego 18A, Poland
e-mail: ewa.majchrzak@polsl.pl

[2] Silesian University of Technology
44-100 Gliwice, Konarskiego 18A, Poland
lukasz.turchan@polsl.pl, jolanta.dziatkiewicz@polsl.pl

Keywords: microscale heat transfer, two-temperature model, phase transitions, finite difference method.

Abstract. *An axisymmetric micro-domain subjected to the ultrashort laser pulse is considered. To describe the process of heat conduction occurring in the analyzed domain, the two-temperature hyperbolic model together with the isothermal solid-liquid and liquid-vapor phase changes is applied. This model, consisting of four equations describing the electrons and lattice temperatures and also the electrons and lattice heat fluxes, is transformed to the model consisting of only two equations describing the electrons and lattice temperatures. Derived equations together with the appropriate boundary and initial conditions, are solved using the finite difference method supplemented by additional numerical procedures which allow to take into account the phase changes. In the final part of the paper the results of computations are shown and the conclusions are formulated.*

1 INTRODUCTION

Analysis of thermal processes occurring in the micro-domains subjected to an ultrashort laser pulse is of vital importance in microtechnology applications [1]. It should be noted that taking into account the extreme temperature gradients, extremely short duration of the process and the domain dimensions expressed in nanometers, the macroscopic heat conduction equation basing on the Fourier law cannot be applied [2-5]. So, to analyze the process, various alternative mathematical models can be used, for example the two-temperature hyperbolic model [6-10]. At high power of laser the phase transitions can occur this means melting and evaporation. In this paper the axisymmetric two-temperature hyperbolic model together with the isothermal solid–liquid and liquid-vapor phase changes is discussed. This model, presented in the chapter 2, consists of four equations describing the electrons and lattice temperatures and also the electrons and lattice heat fluxes. In chapter 3, using appropriate mathematical transformations, the model consisting of two equations describing only the electrons and lattice temperatures is proposed. Chapter 4 contains the description of numerical algorithm based on the finite difference method, while in the chapter 5 the results of computations are shown. In the final part of the paper the conclusions are formulated.

2 GOVERNING EQUATIONS

Axisymmetric two-temperature model describing the temporal and spatial evolution of the lattice and electrons temperatures in the irradiated metal together with the isothermal solid–liquid and liquid-vapor phase changes is of the form [6-8]

$$C_e(T_e)\frac{\partial T_e(r,z,t)}{\partial t} = -\nabla\cdot\mathbf{q}_e(r,z,t) - G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] + Q(r,z,t) \tag{1}$$

$$C_l\frac{\partial T_l(r,z,t)}{\partial t} = -\nabla\cdot\mathbf{q}_l(r,z,t) + G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] + Q_m(r,z,t) + Q_{ev}(r,z,t) \tag{2}$$

where T_e (r, z, t), T_l (r, z, t), $\mathbf{q}_e$ (r, z, t), $\mathbf{q}_l$ (r, z, t) are the temperatures and heat fluxes of the electrons and lattice, respectively, C_e (T_e), C_l are the volumetric specific heats, $G(T_e, T_l)$ is the electron-phonon coupling factor which characterizes the energy exchange between electrons and phonons, $Q(r, z, t)$ is the source function associated with the irradiation, while $Q_m(r, z, t)$ and $Q_{ev}(r, z, t)$ are the source functions associated with the melting and evaporation, respectively.

In a place of classical Fourier law the following formulas are introduced

$$\mathbf{q}_e(r,z,t) + \tau_e\frac{\partial\mathbf{q}_e(r,z,t)}{\partial t} = -\lambda_e(T_e,T_l)\nabla T_e(r,z,t) \tag{3}$$

$$\mathbf{q}_l(r,z,t) + \tau_l\frac{\partial\mathbf{q}_l(r,z,t)}{\partial t} = -\lambda_l\nabla T_l(r,z,t) \tag{4}$$

where λ_e (T_e , T_l), λ_l are the thermal conductivities of the electrons and lattice, respectively, τ_e is the relaxation time of free electrons in metals (the mean time for electrons to change their states), τ_l is the relaxation time in phonon collisions and $\nabla(\cdot)$ denotes the gradient.

The laser irradiation is described by a source term introduced in equation (1) [11, 12]

$$Q(r,z,t) = \sqrt{\frac{4\ln 2}{\pi}}(1-R)\frac{I_0}{\delta t_p}\exp\left[-\frac{r^2}{r_D^2}-\frac{z}{\delta}-4\ln 2\frac{\left(t-2t_p\right)^2}{t_p^2}\right] \tag{5}$$

where I_0 is the laser intensity, t_p is the characteristic time of laser pulse, δ is the optical penetration depth, R is the reflectivity of the irradiated surface and r_D is the laser beam radius.

The internal heat sources resulting from the phase changes (melting and evaporation) take a form [13-15]

$$Q_m(r,z,t) = -L_m \frac{\partial S_m(r,z,t)}{\partial t}$$
$$Q_{ev}(r,z,t) = -L_{ev} \frac{\partial S_{ev}(r,z,t)}{\partial t} \tag{6}$$

where L_m is the volumetric heat of fusion, L_{ev} is the volumetric heat of evaporation, S_m and S_{ev} are the volumetric molten and gaseous state fractions in the surroundings of the point considered. Both S_m and S_{ev} are equal to zero at the beginning of heating process and increase from 0 to 1 when the local temperature achieves the melting T_m and boiling T_{ev} temperatures, respectively.

The above presented mathematical model is supplemented by the boundary conditions (no-flux conditions), it means

$$(r,\ z) \in \Gamma: \quad \begin{aligned} q_{be}(r,\ z,\ t) &= -\lambda_e(T_e,T_l)\ \mathbf{n}\cdot\nabla T_e(r,z,t) = 0 \\ q_{bl}(r,\ z,\ t) &= -\lambda_l\ \mathbf{n}\cdot\nabla T_l(r,z,t) = 0 \end{aligned} \tag{7}$$

where $\mathbf{n}$ is the outward unit normal vector.

The initial temperature distribution is also known

$$t = 0: \qquad T_e(r,\ z,\ t) = T_l(r,\ z,\ t) = T_0 \tag{8}$$

where T_0 is constant.

For high laser intensity the following dependencies are used [6, 8, 16]

$$C_e(T_e) = \begin{cases} AT_e, & T_e < T_F/\pi^2 \\ AT_F/\pi^2 + \dfrac{N k_B - AT_F/\pi^2}{2T_F/\pi^2}\left(T_e - T_F/\pi^2\right), & T_F/\pi^2 \le T_e < 3T_F/\pi^2 \\ N k_B + \dfrac{N k_B/2}{T_F - 3T_F/\pi^2}\left(T_e - 3T_F/\pi^2\right), & 3T_F/\pi^2 \le T_e < T_F \\ 3N k_B/2, & T_e \ge T_F \end{cases} \tag{9}$$

and

$$\lambda_e(T_e,T_l) = \chi \frac{\left[\left(T_e/T_F\right)^2 + 0.16\right]^{5/4}\left[\left(T_e/T_F\right)^2 + 0.44\right]\left(T_e/T_F\right)}{\left[\left(T_e/T_F\right)^2 + 0.092\right]^{1/2}\left[\left(T_e/T_F\right)^2 + \eta\left(T_l/T_F\right)\right]} \tag{10}$$

while

$$G(T_e,T_l) = G_{rt}\left[\frac{A_e}{B_l}\left(T_e + T_l\right) + 1\right] \tag{11}$$

where T_F is the Fermi temperature, N is the density of electrons, k_B is the Boltzmann constant, $A = \pi^2\ N\ k_B / (2\ T_F)$, χ, η, A_e, B_l are the constants and G_{rt} is the coupling factor at room temperature. Other parameters: λ_l, C_l, τ_e, τ_l are assumed to be constant.

3 MATHEMATICAL MANIPULATIONS

In this chapter, another form of the two-temperature model will be presented. Introducing (3) into (1) and (4) into (2), respectively, one has

$$C_e(T_e)\frac{\partial T_e(r,z,t)}{\partial t} = \nabla\cdot\left[\tau_e\frac{\partial \mathbf{q}_e(r,z,t)}{\partial t} + \lambda_e(T_e,T_l)\,\nabla T_e(r,z,t)\right] - G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] + Q(r,z,t) \tag{12}$$

and

$$C_l\frac{\partial T_l(r,z,t)}{\partial t} = \nabla\cdot\left[\tau_l\frac{\partial \mathbf{q}_l(r,z,t)}{\partial t} + \lambda_l\,\nabla T_l(r,z,t)\right] + G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] + Q_m(r,z,t) + Q_{ev}(r,z,t) \tag{13}$$

Assuming that τ_e and τ_l are the constant values, the equations (12), (13) can be written in the form

$$C_e(T_e)\frac{\partial T_e(r,z,t)}{\partial t} = \tau_e\frac{\partial}{\partial t}\left[\nabla\cdot\mathbf{q}_e(r,z,t)\right] + \nabla\left[\lambda_e(T_e,T_l)\,\nabla T_e(r,z,t)\right] - G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] + Q(r,z,t) \tag{14}$$

and

$$C_l\frac{\partial T_l(r,z,t)}{\partial t} = \tau_l\frac{\partial}{\partial t}\left[\nabla\cdot\mathbf{q}_l(r,z,t)\right] + \nabla\left[\lambda_l\,\nabla T_l(r,z,t)\right] + G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] + Q_m(r,z,t) + Q_{ev}(r,z,t) \tag{15}$$

From equations (1), (2) results that

$$-\nabla\cdot\mathbf{q}_e(r,z,t) = C_e(T_e)\frac{\partial T_e(r,z,t)}{\partial t} + G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] - Q(r,z,t) \tag{16}$$

and

$$-\nabla\cdot\mathbf{q}_l(r,z,t) = C_l\frac{\partial T_l(r,z,t)}{\partial t} - G(T_e,T_l)\left[T_e(r,z,t)-T_l(r,z,t)\right] - Q_m(r,z,t) - Q_{ev}(r,z,t) \tag{17}$$

Introducing (16) into (14) and (17) into (15), respectively, the equations (14), (15) have a form

$$C_e(T_e)\frac{\partial T_e(r,z,t)}{\partial t} + \tau_e\frac{\partial C_e(T_e)}{\partial t}\frac{\partial T_e(r,z,t)}{\partial t} + \tau_e C_e(T_e)\frac{\partial^2 T_e(r,z,t)}{\partial t^2} = \nabla\left[\lambda_e(T_e,T_l)\,\nabla T_e(r,z,t)\right] - \left[G(T_e,T_l) + \tau_e\frac{\partial G(T_e,T_l)}{\partial t}\right]\left[T_e(r,z,t)-T_l(r,z,t)\right] - \tau_e G(T_e,T_l)\left[\frac{\partial T_e(r,z,t)}{\partial t} - \frac{\partial T_l(r,z,t)}{\partial t}\right] + Q(r,z,t) + \tau_e\frac{\partial Q(r,z,t)}{\partial t} \tag{18}$$

and

$$C_l \frac{\partial T_l(r,z,t)}{\partial t} + \tau_l C_l \frac{\partial^2 T_l(r,z,t)}{\partial t^2} = \nabla\left[\lambda_l \nabla T_l(r,z,t)\right] +$$
$$\left[G(T_e,T_l) + \tau_l \frac{\partial G(T_e,T_l)}{\partial t}\right]\left[T_e(r,z,t) - T_l(r,z,t)\right] + \tau_l G(T_e,T_l)\left[\frac{\partial T_e(r,z,t)}{\partial t} - \frac{\partial T_l(r,z,t)}{\partial t}\right] + \quad (19)$$
$$Q_m(r,z,t) + \tau_l \frac{\partial Q_m(r,z,t)}{\partial t} + Q_{ev}(r,z,t) + \tau_l \frac{\partial Q_{ev}(r,z,t)}{\partial t}$$

Because

$$\frac{\partial C_e(T_e)}{\partial t} = \frac{\mathrm{d}C_e(T_e)}{\mathrm{d}T_e}\frac{\partial T_e}{\partial t}$$
$$\frac{\partial G(T_e,T_l)}{\partial t} = \frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e}{\partial t} + \frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l}{\partial t} \quad (20)$$

therefore

$$C_e(T_e)\left[\frac{\partial T_e(r,z,t)}{\partial t} + \tau_e \frac{\partial^2 T_e(r,z,t)}{\partial t^2}\right] + \tau_e G(T_e,T_l)\frac{\partial T_e(r,z,t)}{\partial t} =$$
$$\nabla\left[\lambda_e(T_e,T_l)\,\nabla T_e(r,z,t)\right] - \tau_e \frac{\mathrm{d}C_e(T_e)}{\mathrm{d}T_e}\left[\frac{\partial T_e(r,z,t)}{\partial t}\right]^2 -$$
$$\left\{G(T_e,T_l) + \tau_e\left[\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t} + \frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l(r,z,t)}{\partial t}\right]\right\}\left[T_e(r,z,t) - T_l(r,z,t)\right] + \quad (21)$$
$$\tau_e G(T_e,T_l)\frac{\partial T_l(r,z,t)}{\partial t} + Q(r,z,t) + \tau_e \frac{\partial Q(r,z,t)}{\partial t}$$

and

$$C_l\left[\frac{\partial T_l(r,z,t)}{\partial t} + \tau_l \frac{\partial^2 T_l(r,z,t)}{\partial t^2}\right] + \tau_l G(T_e,T_l)\frac{\partial T_l(r,z,t)}{\partial t} = \nabla\left[\lambda_l \nabla T_l(r,z,t)\right] +$$
$$\left\{G(T_e,T_l) + \tau_l\left[\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t} + \frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l(r,z,t)}{\partial t}\right]\right\}\left[T_e(r,z,t) - T_l(r,z,t)\right] + \quad (22)$$
$$\tau_l G(T_e,T_l)\frac{\partial T_e(r,z,t)}{\partial t} + Q_m(r,z,t) + \tau_l \frac{\partial Q_m(r,z,t)}{\partial t} + Q_{ev}(r,z,t) + \tau_l \frac{\partial Q_{ev}(r,z,t)}{\partial t}$$

Finally

$$\begin{aligned}
&\left[C_e(T_e)+\tau_e G(T_e,T_l)\right]\frac{\partial T_e(r,z,t)}{\partial t}+\tau_e C_e(T_e)\frac{\partial^2 T_e(r,z,t)}{\partial t^2}= \\
&\nabla\left[\lambda_e(T_e,T_l)\,\nabla T_e(r,z,t)\right]-\tau_e\frac{\mathrm{d}C_e(T_e)}{\mathrm{d}T_e}\left[\frac{\partial T_e(r,z,t)}{\partial t}\right]^2- \\
&\left\{G(T_e,T_l)+\tau_e\left[\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t}+\frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l(r,z,t)}{\partial t}\right]\right\}\left[T_e(r,z,t)-T_l(r,z,t)\right]+ \\
&\tau_e G(T_e,T_l)\frac{\partial T_l(r,z,t)}{\partial t}+Q(r,z,t)+\tau_e\frac{\partial Q(r,z,t)}{\partial t}
\end{aligned} \tag{23}$$

and

$$\begin{aligned}
&C_l\left[\frac{\partial T_l(r,z,t)}{\partial t}+\tau_l\frac{\partial^2 T_l(r,z,t)}{\partial t^2}\right]+\tau_l G(T_e,T_l)\frac{\partial T_l(r,z,t)}{\partial t}=\nabla\left[\lambda_l\,\nabla T_l(r,z,t)\right]+ \\
&\left\{G(T_e,T_l)+\tau_l\left[\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t}+\frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l(r,z,t)}{\partial t}\right]\right\}\left[T_e(r,z,t)-T_l(r,z,t)\right]+ \\
&\tau_l G(T_e,T_l)\frac{\partial T_e(r,z,t)}{\partial t}+Q_m(r,z,t)+\tau_l\frac{\partial Q_m(r,z,t)}{\partial t}+Q_{ev}(r,z,t)+\tau_l\frac{\partial Q_{ev}(r,z,t)}{\partial t}
\end{aligned} \tag{24}$$

Summarizing, the equations (23) and (24) supplemented by boundary conditions (7) and initial condition (8) create an equivalent two-temperature model described in the chapter 2. It should be noted that in this model the second time derivatives appear and therefore the additional initial condition should be introduced

$$t=0:\quad \left.\frac{\partial T(r,z,t)}{\partial t}\right|_{t=0}=0 \tag{25}$$

The phase transitions are modeled using the authorial version of the algorithm called 'a temperature recovery method' [17]. When the local temperature at the node (i, j) achieves the value of melting point T_m then the source function Q_m (r_i , z_j , t) starts, at the same time, because the melting process proceeds at the constant temperature the derivatives of lattice temperature with respect to time (equation (24)) are equal to 0. So, one has (c.f. equation (24))

$$\begin{aligned}
&0=\nabla\left[\lambda_l\,\nabla T_l(r,z,t)\right]+\left[G(T_e,T_l)+\tau_l\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t}\right]\left[T_e(r,z,t)-T_m\right]+ \\
&\tau_l G(T_e,T_l)\frac{\partial T_e(r,z,t)}{\partial t}-L_m\frac{\partial S_m(r,z,t)}{\partial t}-\tau_l L_m\frac{\partial^2 S_m(r,z,t)}{\partial t^2}
\end{aligned} \tag{26}$$

it means

$$\begin{aligned}
&L_m\frac{\partial S_m(r,z,t)}{\partial t}+\tau_l L_m\frac{\partial^2 S_m(r,z,t)}{\partial t^2}=\nabla\left[\lambda_l\,\nabla T_l(r,z,t)\right]+ \\
&\left[G(T_e,T_l)+\tau_l\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t}\right]\left[T_e(r,z,t)-T_m\right]+\tau_l G(T_e,T_l)\frac{\partial T_e(r,z,t)}{\partial t}
\end{aligned} \tag{27}$$

The similar equation describes the evaporation process

$$L_{ev}\frac{\partial S_{ev}(r,z,t)}{\partial t}+\tau_l L_{ev}\frac{\partial^2 S_{ev}(r,z,t)}{\partial t^2}=\nabla\left[\lambda_l \nabla T_l(r,z,t)\right]+$$
$$\left[G(T_e,T_l)+\tau_l\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e(r,z,t)}{\partial t}\right]\left[T_e(r,z,t)-T_{ev}\right]+\tau_l G(T_e,T_l)\frac{\partial T_e(r,z,t)}{\partial t} \tag{28}$$

4 METHOD OF SOLUTION

To solve the problem formulated in the chapter 3, the finite difference method is used [18, 19]. The geometrical mesh with dimensions $n\times n$ is introduced and the temperatures for time $t^f = f\,\Delta t$ ($f \geq 2$, Δt is the constant time step) at the node (i, j) are denoted as $T^f_{e\,i,j}=T_e(r_j,z_i,t^f)$ and $T^f_{l\,i,j}=T_l(r_j,z_i,t^f)$, respectively.

The following approximation of equation (23) is proposed

$$\left(C^{f-1}_{ei,j}+\tau_e G^{f-1}_{i,j}\right)\frac{T^f_{ei,j}-T^{f-1}_{ei,j}}{\Delta t}+\tau_e C^{f-1}_{ei,j}\frac{T^f_{ei,j}-2T^{f-1}_{ei,j}+T^{f-2}_{ei,j}}{(\Delta t)^2}=$$
$$\nabla\left(\lambda_e\nabla T_e\right)^{f-1}_{i,j}-\tau_e\left(\frac{\mathrm{d}C_e}{\mathrm{d}T_e}\right)^{f-1}_{i,j}\frac{\left(T^{f-1}_{ei,j}-T^{f-2}_{ei,j}\right)^2}{(\Delta t)^2}-$$
$$\left\{G^{f-1}_{i,j}+\tau_e\left[\left(\frac{\partial G}{\partial T_e}\right)^{f-1}_{i,j}\frac{T^{f-1}_{ei,j}-T^{f-2}_{ei,j}}{\Delta t}+\left(\frac{\partial G}{\partial T_l}\right)^{f-1}_{i,j}\frac{T^{f-1}_{li,j}-T^{f-2}_{li,j}}{\Delta t}\right]\right\}\left(T^{f-1}_{ei,j}-T^{f-1}_{li,j}\right)+ \tag{29}$$
$$\tau_e\left[\left(\frac{\partial G}{\partial T_e}\right)^{f-1}_{i,j}\frac{T^{f-1}_{ei,j}-T^{f-2}_{ei,j}}{\Delta t}+\left(\frac{\partial G}{\partial T_l}\right)^{f-1}_{i,j}\frac{T^{f-1}_{li,j}-T^{f-2}_{li,j}}{\Delta t}\right]\left(T^{f-1}_{ei,j}-T^{f-1}_{li,j}\right)+$$
$$\tau_e G^{f-1}_{i,j}\frac{T^{f-1}_{li,j}-T^{f-2}_{li,j}}{\Delta t}+Q^{f-1}_{i,j}+\tau_e\left(\frac{\partial Q}{\partial t}\right)^{f-1}_{i,j}$$

where

$$\nabla\left(\lambda_e\nabla T_e\right)^{f-1}_{i,j}=\Phi_1\frac{T^{f-1}_{ei,j-1}-T^{f-1}_{ei,j}}{R^{f-1}_{i,j1}}+\Phi_2\frac{T^{f-1}_{ei,j+1}-T^{f-1}_{ei,j}}{R^{f-1}_{i,j2}}+$$
$$\Phi_3\frac{T^{f-1}_{ei-1,j}-T^{f-1}_{ei,j}}{R^{f-1}_{i,j3}}+\Phi_4\frac{T^{f-1}_{ei+1,j}-T^{f-1}_{ei,j}}{R^{f-1}_{i,j4}} \tag{30}$$

while

$$R^{f-1}_{i,j1}=\frac{h}{2}\left(\frac{1}{\lambda^{f-1}_{ei,j}}+\frac{1}{\lambda^{f-1}_{ei,j-1}}\right),\quad R^{f-1}_{i,j2}=\frac{h}{2}\left(\frac{1}{\lambda^{f-1}_{ei,j}}+\frac{1}{\lambda^{f-1}_{ei,j+1}}\right)$$
$$R^{f-1}_{i,j3}=\frac{h}{2}\left(\frac{1}{\lambda^{f-1}_{ei,j}}+\frac{1}{\lambda^{f-1}_{ei-1,j}}\right),\quad R^{f-1}_{i,j4}=\frac{h}{2}\left(\frac{1}{\lambda^{f-1}_{ei,j}}+\frac{1}{\lambda^{f-1}_{ei+1,j}}\right) \tag{31}$$

are the thermal resistances between adjacent nodes [18, 19] and

$$\Phi_1 = \left(r_{i,j} - 0.5h\right)/\left(h\, r_{i,j}\right), \quad \Phi_2 = \left(r_{i,j} + 0.5h\right)/\left(h\, r_{i,j}\right), \quad \Phi_3 = \Phi_4 = 1/h \tag{32}$$

are the shape functions of differential mesh for the axially-symmetrical task.
After mathematical manipulations one has

$$\begin{aligned} T_{ei,j}^{f} = {} & \frac{\left(\Delta t\right)^2 \nabla\left(\lambda_e \nabla T_e\right)_{i,j}^{f-1}}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} + \frac{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + 2\tau_e C_{ei,j}^{f-1}}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} T_{ei,j}^{f-1} - \\ & \frac{\tau_e C_{ei,j}^{f-1} T_{ei,j}^{f-2}}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} - \frac{\tau_e \left(T_{ei,j}^{f-1} - T_{ei,j}^{f-2}\right)^2}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} \left(\frac{\mathrm{d}C_e}{\mathrm{d}T_e}\right)_{i,j}^{f-1} - \\ & \frac{\left(\Delta t\right)^2 \left(T_{ei,j}^{f-1} - T_{li,j}^{f-1}\right)}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} \left\{ G_{i,j}^{f-1} + \tau_e \left[\left(\frac{\partial G}{\partial T_e}\right)_{i,j}^{f-1} \frac{T_{ei,j}^{f-1} - T_{ei,j}^{f-2}}{\Delta t} + \left(\frac{\partial G}{\partial T_l}\right)_{i,j}^{f-1} \frac{T_{li,j}^{f-1} - T_{li,j}^{f-2}}{\Delta t} \right] \right\} + \\ & \frac{\tau_e G_{i,j}^{f-1} \Delta t \left(T_{li,j}^{f-1} - T_{li,j}^{f-2}\right)}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} + \frac{\left(\Delta t\right)^2}{\Delta t\left(C_{ei,j}^{f-1} + \tau_e G_{i,j}^{f-1}\right) + \tau_e C_{ei,j}^{f-1}} \left[Q_{i,j}^{f-1} + \tau_e \left(\frac{\partial Q}{\partial t}\right)_{i,j}^{f-1} \right] \end{aligned} \tag{33}$$

In a similar way the equation (24) can be approximated and for the temperatures below the melting point one obtains

$$\begin{aligned} T_{li,j}^{f} = {} & \frac{\left(\Delta t\right)^2 \nabla\left(\lambda_l \nabla T_l\right)_{i,j}^{f-1}}{\Delta t\left(C_l + \tau_l G_{i,j}^{f-1}\right) + \tau_l C_l} + \frac{\Delta t\left(C_l + \tau_l G_{i,j}^{f-1}\right) + 2\tau_l C_l}{\Delta t\left(C_l + \tau_l G_{i,j}^{f-1}\right) + \tau_l C_l} T_{li,j}^{f-1} - \\ & \frac{\tau_l C_l}{\Delta t\left(C_l + \tau_l G_{i,j}^{f-1}\right) + \tau_l C_l} T_{li,j}^{f-2} + \frac{\tau_l G_{i,j}^{f-1} \Delta t}{\Delta t\left(C_l + \tau_l G_{i,j}^{f-1}\right) + \tau_l C_l} \left(T_{ei,j}^{f-1} - T_{ei,j}^{f-2}\right) + \\ & \frac{\left(\Delta t\right)^2 \left(T_{ei,j}^{f-1} - T_{li,j}^{f-1}\right)}{\Delta t\left(C_l + \tau_l G_{i,j}^{f-1}\right) + \tau_l C_l} \left\{ G_{i,j}^{f-1} + \tau_l \left[\left(\frac{\partial G}{\partial T_e}\right)_{i,j}^{f-1} \frac{T_{ei,j}^{f-1} - T_{ei,j}^{f-2}}{\Delta t} + \left(\frac{\partial G}{\partial T_l}\right)_{i,j}^{f-1} \frac{T_{li,j}^{f-1} - T_{li,j}^{f-2}}{\Delta t} \right] \right\} \end{aligned} \tag{34}$$

In turn, the finite difference approximation of equation (27) is of the form

$$\begin{aligned} & L_m \frac{S_{mi,j}^{f} - S_{mi,j}^{f-1}}{\Delta t} + \tau_l L_m \frac{S_{mi,j}^{f} - 2S_{mi,j}^{f-1} + S_{mi,j}^{f-2}}{\left(\Delta t\right)^2} = \nabla\left[\lambda_l \nabla T_l(r,z,t)\right]_{i,j}^{f-1} + \\ & \left[G_{i,j}^{f-1} + \tau_l \left(\frac{\partial G}{\partial T_e}\right)_{i,j}^{f-1} \frac{T_{ei,j}^{f-1} - T_{ei,j}^{f-2}}{\Delta t} \right] \left(T_{ei,j}^{f-1} - T_m\right) + \tau_l G_{i,j}^{f-1} \frac{T_{ei,j}^{f-1} - T_{ei,j}^{f-2}}{\Delta t} \end{aligned} \tag{35}$$

and next

$$\begin{aligned} S_{mi,j}^{f} = {} & \frac{\Delta t + 2\tau_l}{\Delta t + \tau_l} S_{mi,j}^{f-1} - \frac{\tau_l}{\Delta t + \tau_l} S_{mi,j}^{f-2} + \frac{\left(\Delta t\right)^2}{L_m\left(\Delta t + \tau_l\right)} \nabla\left[\lambda_l \nabla T_l(r,z,t)\right]_{i,j}^{f-1} + \\ & \frac{\left(\Delta t\right)^2}{L_m\left(\Delta t + \tau_l\right)} \left[G_{i,j}^{f-1} + \tau_l \left(\frac{\partial G}{\partial T_e}\right)_{i,j}^{f-1} \frac{T_{ei,j}^{f-1} - T_{ei,j}^{f-2}}{\Delta t} \right] \left(T_{ei,j}^{f-1} - T_m\right) + \frac{\tau_l \Delta t G_{i,j}^{f-1}}{L_m\left(\Delta t + \tau_l\right)} \left(T_{ei,j}^{f-1} - T_{ei,j}^{f-2}\right) \end{aligned} \tag{36}$$

In a similar way one obtains (c.f. equation (28))

$$S_{evi,j}^{f} = \frac{\Delta t + 2\tau_l}{\Delta t + \tau_l} S_{evi,j}^{f-1} - \frac{\tau_l}{\Delta t + \tau_l} S_{evi,j}^{f-2} + \frac{(\Delta t)^2}{L_{ev}(\Delta t + \tau_l)} \nabla\left[\lambda_l \nabla T_l(r,z,t)\right]_{i,j}^{f-1} + \frac{(\Delta t)^2}{L_{ev}(\Delta t + \tau_l)} \left[G_{i,j}^{f-1} + \tau_l \left(\frac{\partial G}{\partial T_e} \right)_{i,j}^{f-1} \frac{T_{ei,j}^{f-1} - T_{ei,j}^{f-2}}{\Delta t} \right] \left(T_{ei,j}^{f-1} - T_{ev} \right) + \frac{\tau_l \Delta t G_{i,j}^{f-1}}{L_{ev}(\Delta t + \tau_l)} \left(T_{ei,j}^{f-1} - T_{ei,j}^{f-2} \right) \tag{37}$$

5 RESULTS OF COMPUTATIONS

Numerical simulation of a thermal process taking place in the cylindrical domain with dimensions $Z = 100 \cdot 10^{-9}$ [m] and $R = 100 \cdot 10^{-9}$ [m] subjected to the short-pulse laser heating has been done. The initial temperature is equal to $T_p = 300$ K. Thermophysical parameters of material are equal to (gold): thermal conductivity of lattice $\lambda_l = 315$ W/(mK), volumetric specific heat of lattice $C_l = 2.5$ MJ/(m^3 K), electrons relaxation time $\tau_e = 0.04$ ps, phonons relaxation time $\tau_l = 0.8$ ps [20], reflectivity $R = 0.93$, optical penetration depth $\delta = 15.3$ nm. The Fermi temperature is equal to $T_F = 64\,200$ K and the density of electrons is equal to $N = 5.9 \cdot 10^{28}$ 1/m (c.f. equation (9)) [6]. The constants in equations (10), (11) are the following: $\chi = 353$ W/(mK), $\eta = 0.16$, $A_e = 1.2 \cdot 10^7$ 1/(K^2 s), $B_l = 1.23 \cdot 10^{11}$ 1/(Ks) and $G_{rt} = 2.2 \cdot 10^{16}$ W/(m^{3}K) [6]. The melting point is equal to $T_m = 1336$ K, the boiling point: $T_e = 3127$ K, the volumetric heat of fusion: $L_m = 1.23 \cdot 10^3$ MJ/m^3 and the volumetric heat of evaporation: $L_{ev} = 3.277 \cdot 10^4$ MJ/m^3.

The laser beam radius is equal to $r_D = R/8$. The calculations were performed for five values of the laser intensity: $I_0 = 2 \cdot 10^5$ J/m^2, $I_0 = 4 \cdot 10^5$ J/m^2, $I_0 = 6 \cdot 10^5$ J/m^2 , $I_0 = 8 \cdot 10^5$ J/m^2 and $I_0 = 10^6$ J/m^2 under the assumption that the characteristic time of laser pulse is equal to $t_p = 100$ ps (c.f. formula (5)). The problem is solved using the finite difference method. Number of nodes is equal to 50×50=2500 (h=2 nm) and time step is equal to 0.0005 ps.

In Figures 1 and 2 the heating curves at the point (2nm, 2nm) are shown. As can be seen, for the lowest laser intensity only the heating process is observed, while for the higher laser intensity the melting process appears ('stop' corresponding to the melting process). In the case of the largest laser intensity (Figure 2) the isothermal liquid-vapour phase change starts. Figures 3 and 4 illustrate the courses of electrons and lattice temperatures at the same point.

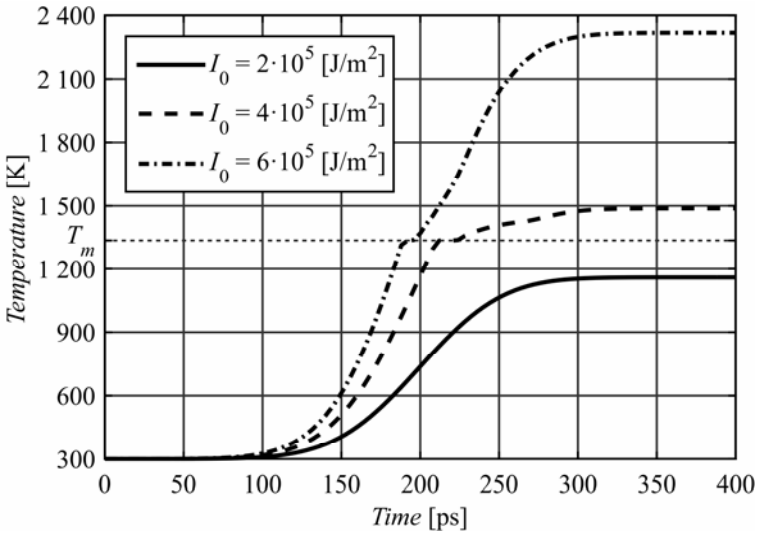


Figure 1: Lattice temperature history at the point (2 nm, 2nm).

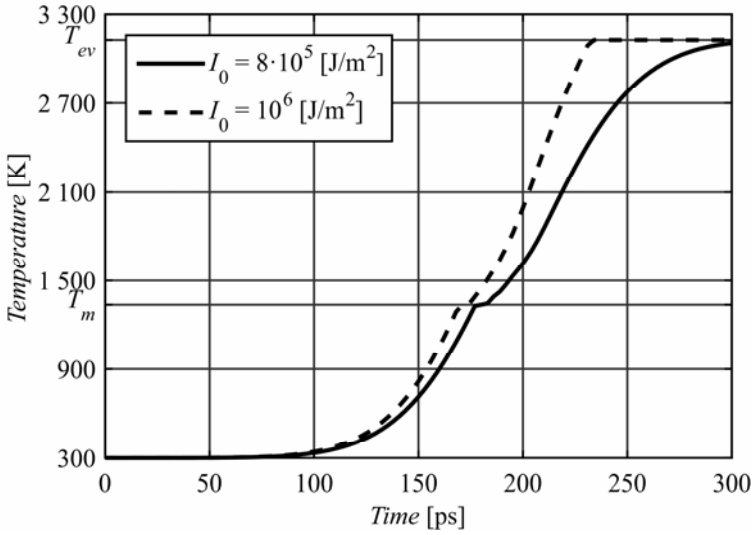


Figure 2: Lattice temperature history at the point (2 nm, 2nm).

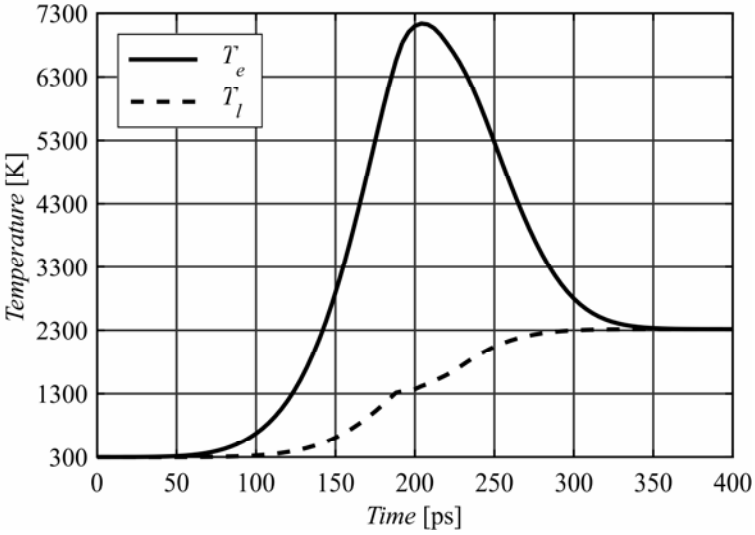


Figure 3: Temperature history at the point (2nm, 2nm) for $I_0 = 6 \cdot 10^5$ J/m^2.

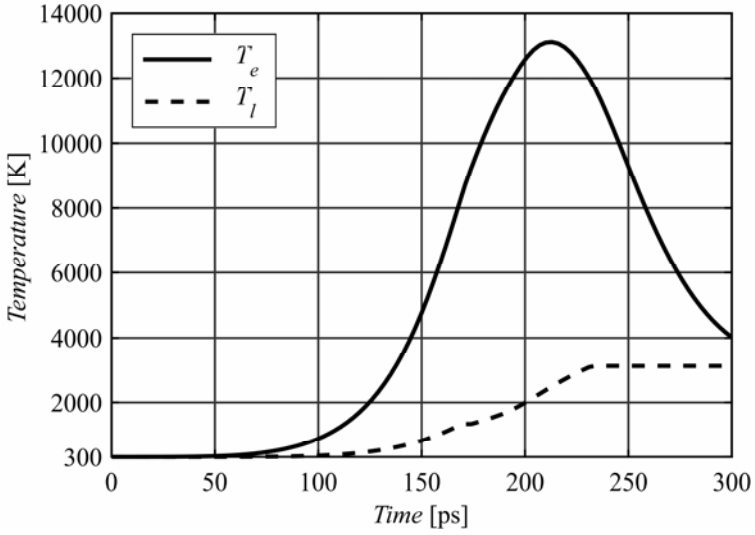


Figure 4: Temperature history at point (2nm, 2nm) for $I_0 = 10^6$ J/m^2.

In Figure 5 the electrons and phonons temperature distribution after 150 ps for laser intensity $I_0 = 8 \cdot 10^5$ J/m^2 is shown.

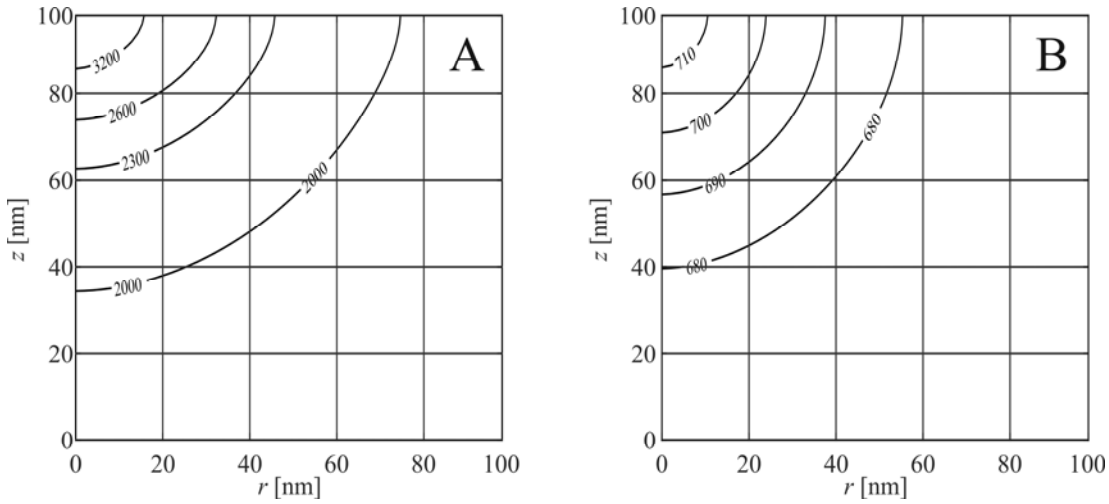


Figure 5: Temperature distribution after 150 ps, A) electrons, B) lattice.

6 FINAL REMARKS

The thermal processes occurring in the axisymmetric micro-domain subjected to the ultrashort laser pulse are analyzed. The problem is described by two equations for electrons and lattice temperatures supplemented by appropriate boundary and initial conditions. At the stage of numerical computations the finite difference method together with the procedures modeling the phase transitions is used.

In the future, the algorithm will be supplemented by a procedure modeling the ablation process.

ACKNOWLEDGEMENT

This work is supported by the project BK-255/RMT4/2016.

REFERENCES

[1] R.A. Escobar, S.S. Ghai, M.S. Jhon, C.H. Amon, Multi-length and time scale thermal transport using the lattice Boltzmann method with application to electronic cooling. *International Journal of Heat and Mass Transfer*, **49**, 97-107, 2006.

[2] G. Chen, D. Borca-Tasciuc, R.G. Yang, Nanoscale Heat Transfer. *Encyclopedia of NanoScience and Nanotechnology*, 17, 429–459, 2004.

[3] Z.M. Zhang, *Nano/microscale heat transfer, 1st Edition*. McGraw Hill, 2007.

[4] A.N. Smith, P.M. Norris, Microscale heat transfer. *A. Bejan, D. Kraus (Eds.), Heat Transfer Handbook, John Wiley & Sons, New Jersey*, **18**, 1309-1359, 2003.

[5] D.Y. Tzou, *Macro- to Microscale Heat Transfer: The lagging behavior, 2nd Edition*, Taylor and Francis, 1997.

[6] J.K. Chen, D.Y. Tzou, J.E. Beraun, A semiclassical two-temperature model for ultrafast laser heating. *International Journal of Heat and Mass Transfer*, **49**, 307-316, 2006.

[7] J.K Chen, W.P. Latham, J.E. Beraun, The role of electron–phonon coupling in ultrafast laser heating. *Journal of Laser Applications*, **17(1)**, 63-68, 2005.

[8] J. Huang, K. Baheti, J.K. Chen, Y. Zhang, An axisymmetric model for solid-liquid-vapor phase change in thin metal films induced by an ultrashort laser pulse. *Frontiers in Heat and Mass Transfer*, **2**, 013005, 2011.

[9] E. Majchrzak, J. Dziatkiewicz, Ł. Turchan, Sensitivity analysis and inverse problems in microscale heat transfer. *Defect and Diffusion Forum*, **362**, 209-223, 2015.

[10] J. Dziatkiewicz, W. Kuś. E. Majchrzak, T. Burczyński, Ł. Turchan, Bioinspired identification of parameters in microscale heat transfer. *International Journal for Multiscale Computational Engineering*, **12(1)**, 79-89, 2014.

[11] C.P. Grigoropoulos, A. Chimmalgi, D.J. Hwang, Nano-structuring using pulsed laser radiation. *Laser Ablation and its Applications, Springer Series in Optical Sciences*, **129**, 473-504, 2007.

[12] E. Majchrzak, B. Mochnacki, Modeling of melting and resolidification in domain of metal film subjected to a laser pulse. *Archives of Foundry Engineering*, **16(1)**, 41-44, 2016.

[13] A. Tarzia, A. Domingo, A Bibliography on Moving-Free Boundary Problems for the Heat-Diffusion Equation. The Stefan and Related Problems. *MAT, Series A: Conferencias, seminarios y trabajos de matemática*, **2**, 1–297, 2000.

[14] B. Mochnacki, R. Szopa, Generalized micro/macro model of crystallization and its numerical realization. *International Journal for Multiscale Computational Engineering*, **8(3)**, 259-266, 2010.

[15] B. Mochnacki, E. Majchrzak, Identification of macro and micro parameters in solidification model. *Bulletin of the Polish Academy of Sciences, Technical Sciences*, **55(1)**, 107-113, 2007.

[16] E. Majchrzak, J. Dziatkiewicz, Analysis of ultashort laser pulse interactions with metal films using a two-temperature model. *Journal of Applied Mathematics and Computational Mechanics*, **14(2)**, 31-39, 2015.

[17] E. Majchrzak, B. Mochnacki, J.S. Suchy, Computer simulation of heat transfer between the particles and metal matrix during the solidification of a cast composite. *International Journal of Cast Metals Research*, **12(4)**, 241-249, 1999.

[18] E. Majchrzak, Ł. Turchan, J. Dziatkiewicz, Modeling of skin tissue heating using the generalized dual-phase lag equation. *Archives of Mechanics*, **67(6)**, 417-437, 2015.

[19] B. Mochnacki, M. Ciesielski, Numerical model of thermal processes in domain of thin film subjected to a cyclic external heat flux. *THERMEC 2011, PTS 1-4 Book Series: Materials Science Forum*, **706–709**, 1460–1465, 2012.

[20] J.K. Chen, J.E. Beraun, Numerical study of ultrashort laser pulse interactions with metal films. *Numerical Heat Transfer, Part A*, **40**, 1-40, 2001.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

VALIDATION OF A DISCRETE CRACK MODEL FOR LIGHTWEIGHT AGGREGATE CONCRETE BEAMS

D. Dias-da-Costa[1,2], R. Graça-e-Costa[3], and R.N.F. Carmo[4]

[1] School of Civil Engineering, The University of Sydney, Sydney, NSW 2006, Australia
[2] ISISE, University of Coimbra, Rua Luís Reis Santos, 3030-788 Coimbra, Portugal
e-mail: daniel.diasdacosta@sydney.edu.au

[3] CEPAC, Universidade do Algarve
Campus de Gambelas, 8005-139 Faro, Portugal
e-mail: rcosta@ualg.pt

[4] ICIST, Polytechnic Institute of Coimbra
Rua Pedro Nunes, 3030-199 Coimbra, Portugal
e-mail: carmo@isec.pt

Keywords: Discrete cracks, Lightweight concrete, Crack opening, Flexural behaviour.

Abstract. *The development of crack models capable of simulating the discrete nature of fracture is of interest to many different areas of research. For example, the structural analysis innovative designs now made possible by new ultra-high performance concrete mixtures would certainly benefit from such improved predictive capabilities. Currently, there are numerous numerical approaches available in the literature, for instance based on nodal or element enrichment techniques, or even on remeshing strategies. Typically, the validation of such approaches was achieved using benchmark tests that contained few cracks and where the overall displacements were compared until failure. Having this into account, this paper describes a detailed validation of a discrete crack model based on embedded discontinuities for predicting the behaviour of lightweight aggregate concrete. The model itself includes the rigid body movements associated with the opening of cracks and relies on a robust non-iterative algorithm to overcome convergence difficulties typically found with numerous cracks and material non-linearities. Validation was achieved using experimental data from tests performed on lightweight concrete beams (LWAC) under flexural load, where displacements, curvatures and cracks width were properly monitored. This data include, not only overall displacements, but also the complex crack patterns produced during the tests. The model was shown to predict well the overall crack patterns and openings, and was used to perform extrapolations on crack widths for different reinforcement ratios.*

1 INTRODUCTION

There are many situations where the accurate prediction of the fractured behaviour of a structure requires models with the capability of simulating discrete cracks. This is true whenever the interaction between cracks and other structural features – e.g. reinforcement or strengthening materials – is critical to identify premature failures and to characterise the structural behaviour under serviceability conditions. Currently, there are many different approaches available in the literature, which can simulate material separation. Most approaches are based on nodal [1-4] or element enrichment techniques [5-11], or even on remeshing strategies [12-14]. These approaches were typically validated using benchmark tests with few cracks and by comparing overall displacements. Starting from this observation, this paper describes a validation study concerning a discrete crack model for embedding discontinuities and that can be efficiently used for predicting the behaviour of lightweight aggregate concrete (LWAC). The reason underlying the choice of LWAC is related to the fact of this cementitious material being relatively new in the field of civil engineering and currently facing many new applications. As example, it can be mentioned the strengthening of existing structures and the new structures being built with reduced overall gravity load, which impacts directly the design of the structural members and, indirectly, because of the decreased magnitude of seismic actions [15-17].

Until recently, most research in the field of LWAC was focused on the material behaviour from an experimental perspective [18-22], lacking numerical models that can be efficiently used for predicting the structural behaviour. Simultaneously, existing standards still need proper validation, in which case a robust numerical model will be needed to support further studies. Developed within this context, the model presented in this paper assumes the opening of the cracks to occur as if it were a rigid body movement and relies on a robust non-iterative algorithm to overcome potential convergence difficulties found in the presence of numerous cracks and material non-linearities. In the following sections, the model is described in detail and validated using experimental data from tests performed on light-weight concrete beams under flexural loads, focusing the overall displacements and the complex crack patterns produced during the tests.

2 DISCRETE CRACK APPROACH

2.1 Variational framework

The variational principle for a body Ω containing a surface of discontinuity, Γ_d, splitting it into two subdomains, Ω^- and Ω^+, and subjected to quasi-static body forces $\mathbf{b}$ and stresses $\mathbf{t}$ distributed over the external boundary, Γ_t, is given by [23]:

$$\int_{\Omega\setminus\Gamma_d} (\nabla^s \delta\mathbf{u}) : \sigma(\varepsilon) d\Omega + \int_{\Gamma_d} \delta[\![\mathbf{u}]\!] \cdot \mathbf{t}^+ d\Gamma = \int_{\Omega\setminus\Gamma_d} \delta\mathbf{u}\cdot\mathbf{b} d\Omega + \int_{\Gamma_t} \delta\mathbf{u}\cdot\mathbf{t} d\Gamma \,, \tag{0}$$

where σ is the stress tensor and $\mathbf{t}^+$ is the stress at the crack, $[\![\mathbf{u}]\!]$ is the opening of the crack, and $\mathbf{u}$ is the total displacement field.

The total displacement can be considered to be the sum of two contributions, the regular displacement field, $\hat{\mathbf{u}}$, and the enhanced displacement field, $\tilde{\mathbf{u}}$, according to which:

$$\delta\mathbf{u} = \delta\hat{\mathbf{u}} + H_{\Gamma_d}\delta\tilde{\mathbf{u}} \,, \tag{0}$$

where H_{Γ_d} is the standard Heaviside function, with value '1' on Ω^+ (and '0' otherwise).

For small displacements:

$$\nabla^s \delta\mathbf{u} = \nabla^s \delta\hat{\mathbf{u}} + H_{\Gamma_d}\left(\nabla^s \delta\tilde{\mathbf{u}}\right) + \delta_{\Gamma_d}\left(\left[\!\left[\mathbf{u}\right]\!\right] \otimes \mathbf{n}^+\right)^S, \tag{0}$$

where H_{Γ_d} is the Dirac's delta function along the surface of discontinuity and $\mathbf{n}^+$ the vector orthogonal to the discontinuity and pointing inwards Ω^+.
If the opening of the discontinuity is transmitted as if it were a rigid body motion, then $\nabla^s \delta\tilde{\mathbf{u}} = \mathbf{0}$. By replacing Eqs. (2) and (3) in Eq. (1), and by taking $\delta\tilde{\mathbf{u}} = \mathbf{0}$ and then $\delta\hat{\mathbf{u}} = \mathbf{0}$, the following governing equations is derived:

$$\begin{aligned} &\int_{\Omega\setminus\Gamma_d} (\nabla^s \delta\hat{\mathbf{u}}) : \sigma(\hat{\varepsilon}) d\Omega = \int_{\Omega\setminus\Gamma_d} \delta\hat{\mathbf{u}} \cdot \mathbf{b} d\Omega + \int_{\Gamma_t} \delta\hat{\mathbf{u}} \cdot \mathbf{t} d\Gamma \\ &\int_{\Gamma_d} \delta\left[\!\left[\mathbf{u}\right]\!\right] \cdot \mathbf{t}^+ d\Gamma = \int_{\Omega^+} \delta\tilde{\mathbf{u}} \cdot \mathbf{b} d\Omega + \int_{\Gamma_{t^+}} \delta\tilde{\mathbf{u}} \cdot \mathbf{t} d\Gamma \end{aligned}. \tag{0}$$

2.2 Discretisation

Assuming that the body is discretised into a set of finite elements, the displacement field within each element containing a crack is interpolated by:

$$\begin{aligned} &\mathbf{u}^e = \mathbf{N}^e\left(\hat{\mathbf{a}} + H^e_{\Gamma_d}\mathbf{M}^{ek}_w \mathbf{w}^e\right) \text{ in } \Omega^e \setminus \Gamma^e_d \\ &\left[\!\left[\mathbf{u}\right]\!\right]^e = \mathbf{M}^e_w \mathbf{w}^e = \mathbf{N}^e_w \mathbf{w}^e \text{ at } \Gamma^e_d \end{aligned}, \tag{0}$$

where $\mathbf{N}^e$ is a matrix with the shape functions of the element, $\hat{\mathbf{a}}$ are the nodal degrees of freedom associated with $\hat{\mathbf{u}}^e$, $\mathbf{w}^e$ are the degrees of freedom for the opening of the crack at both extremities, $\mathbf{N}^e_w$ contains linear interpolation functions and matrix $\mathbf{M}^e_w$ is defined as:

$$\mathbf{M}^e_w = \begin{bmatrix} 1 - \dfrac{(x_2 - x_2^i)\sin\alpha^e}{l^e_d} & \dfrac{(x_2 - x_2^i)\cos\alpha^e}{l^e_d} & \dfrac{(x_2 - x_2^i)\sin\alpha^e}{l^e_d} & -\dfrac{(x_2 - x_2^i)\cos\alpha^e}{l^e_d} \\ \dfrac{(x_1 - x_1^i)\sin\alpha^e}{l^e_d} & 1 - \dfrac{(x_1 - x_1^i)\cos\alpha^e}{l^e_d} & -\dfrac{(x_1 - x_1^i)\sin\alpha^e}{l^e_d} & \dfrac{(x_1 - x_1^i)\cos\alpha^e}{l^e_d} \end{bmatrix}, \tag{0}$$

where (x_1, x_2) are the coordinates of any point inside the element, (x_1^i, x_2^i) are the coordinates of the first tip of the discontinuity, l^e_d and α^e are, respectively, the length and angle of the discontinuity. $\mathbf{M}^{ek}_w$ is obtained by calculating Eq. (6) at each regular node of the finite element and stacking the result into rows.
Following standard finite element procedures, the previous equations can be used to discretise Eq. (4) into the following system:

$$\begin{bmatrix} \mathbf{K}^e_{\hat{\mathbf{a}}\hat{\mathbf{a}}} & -\mathbf{K}^e_{\mathbf{aw}} \\ -\mathbf{K}^e_{\mathbf{aw}} & \mathbf{K}^e_{\mathbf{ww}} + \mathbf{K}^e_d + \mathbf{K}^e_p \end{bmatrix} \begin{Bmatrix} \mathrm{d}\mathbf{a}^e \\ \mathrm{d}\mathbf{w}^e \end{Bmatrix} = \begin{Bmatrix} \mathrm{d}\hat{\mathbf{f}}^e \\ \mathrm{d}\mathbf{f}_w{}^e - \left(\mathbf{H}^e_{\Gamma_d}\mathbf{M}^{ek}_w\right)^T \mathrm{d}\hat{\mathbf{f}}^e \end{Bmatrix}, \tag{0}$$

where $\mathbf{K}^e_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$ is the stiffness of a regular finite element, $\mathbf{K}^e_d$ is the stiffness of the discontinuity, $\mathbf{K}^e_{\mathbf{aw}} = \mathbf{K}^e_{\hat{\mathbf{a}}\hat{\mathbf{a}}}\mathbf{H}^e_{\Gamma_d}\mathbf{M}^{ek}_w$, $\mathbf{K}^e_{\mathbf{wa}} = \mathbf{K}^{eT}_{\mathbf{aw}}$, and $\mathbf{K}^e_{\mathbf{ww}} = \left(\mathbf{H}^e_{\Gamma_d}\mathbf{M}^{ek}_w\right)^T \mathbf{K}^e_{\hat{\mathbf{a}}\hat{\mathbf{a}}}\mathbf{H}^e_{\Gamma_d}\mathbf{M}^{ek}_w$. Finally, $\mathbf{K}^e_p$ assures appropriate shear jump transmission along the discontinuity (more details in [11, 24]).

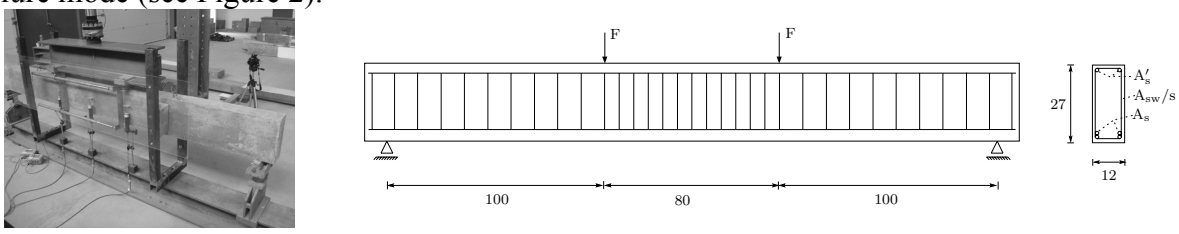
2.3 Crack propagation and solution procedure

The Rankine criterion is used to identify the onset of cracking, with new cracks being introduced through the centre of the cracked element. In the case of crack propagation, the new crack segment is inserted from the crack tip after evaluating the stress field using a procedure that smooths the stress within an average support length of two or three times the typical element size (see [3]). The discontinuity always crosses the entire finite element, with the new degrees of freedom being treated as global unknowns.

The overall system of equations is solved using the Non-Iterative Energy Based Method (NIEM) [25]. According to this method, an incremental analysis is performed until reaching critical bifurcation points. In this situation, a transition to a total approach is performed such that damage is enforced according to information retrieved during the analysis. An energy criterion follows the path leading to the highest dissipation of energy. A comprehensive discussion about this algorithm is found in [25, 26].

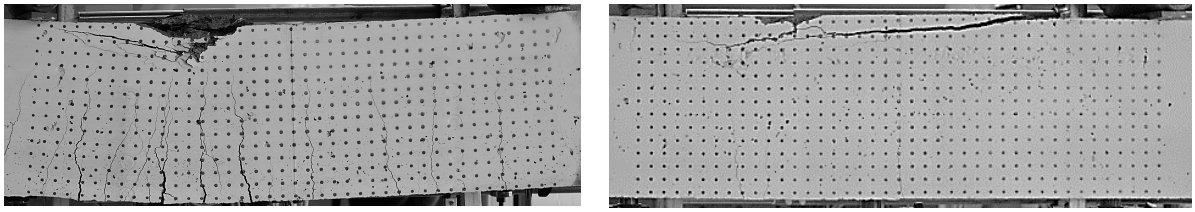
3 EXPERIMENTAL TESTS

The structural scheme adopted for the experimental tests is shown in Figure 1. A total of four beams were grouped into two sets – designated by 1T and 3T – according to the longitudinal tensile reinforcement ratios, respectively 1.12% and 2.96%. These ratios were selected to mobilise two limit failure responses: one more ductile and corresponding to an under-reinforced failure mode, and the other more brittle and corresponding to an over-reinforced failure mode (see Figure 2).



(a) (b)

Figure 1: (a) Loading and measurement apparatus; (b) structural scheme (dimensions in 'mm').



(a) (b)

Figure 2: Failure mode of specimens: (a) 1T; and (b) 3T.

The steel reinforcement adopted for each beam is shown in the following table.

Specimen	A_s (cm^2)	Effective depth (cm)	A'_s (cm^2)	A_{sw}/s (cm^2/m) near supports	A_{sw}/s (cm^2/m) middle span
1T	3.14	23.4	0.57	0.47	0.47
3T	8.04	22.6	0.57	0.47	1.34

Table 1: Steel reinforcement for each tested beam.

All beams were produced using LWAC with an average density of 1870 kg/m^3, compressive and tensile strengths of 57 MPa and 4 MPa, respectively, and a Young's modulus of 25.5 GPa [27, 28]. The longitudinal and transverse reinforcement consisted of hot rolled and ribbed S500NR-SD class bars, with an average yield stress of 545 and an average tensile strength of 645 MPa [29]. Each beam was instrumented using traditional measurement devices to monitor load and the vertical displacements at the three points shown in Figure 1(a). The crack openings were measured for one beam during the test [30-32].

Figure 3 shows the load vs. displacement curves for all tested specimens and a representation of all stages monitored. In this chart, the displacements were measured at the mid-span. As a general remark, four phases can be identified in the overall structural response of the beams. The first or initial phase corresponds to the uncracked stage and is characterised by the highest value in terms of flexural stiffness. This phase finishes at the onset of cracking and the second phase then progresses until the steel reinforcement starts to yield. After this, the third phase is characterised by a significant reduction of flexural stiffness, which ends at the point of maximum load. The fourth and last phase depends on the ductility of the specimen and is the post-peak softening response. This last stage can be quite insignificant in the case of the beams with the highest longitudinal reinforcement ratio.

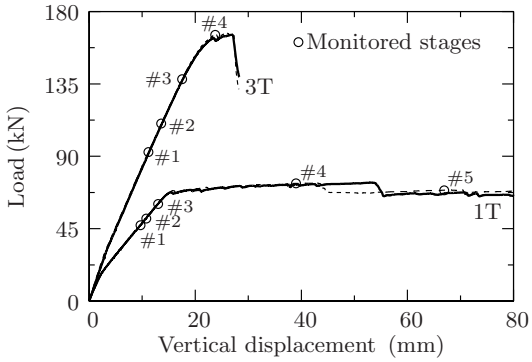


Figure 3: Load vs. mid-span displacement curves and monitored stages.

4 NUMERICAL MODELLING AND VALIDATION

The beams described in the previous section were simulated under plane stress conditions. The concrete was modelled using bilinear elements, whereas linear truss elements were used for steel reinforcements (see Figure 4). In what regards constitutive models, the reinforcements were connected to the concrete elements using interface elements equipped with the Model Code 2010 bond model [33]. The concrete elements were assumed to be linear elastic and perfectly plastic under compression, whereas the discrete crack approach model described in Section 2 was used for the elements where the tensile strength of the material was reached. All embedded cracks followed a mode-I traction separation law with exponential softening and fracture energy equal to 0.10 N/mm.

The numerical model is explored and validated in this section in its ability to simulate crack opening and patterns, and the overall behaviour, namely displacements and curvatures.

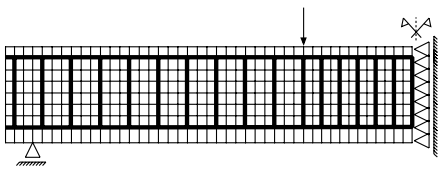


Figure 4: Finite element mesh.

4.1 Displacements and curvatures

The load vs. displacement curves obtained with all numerical models are shown in Figure 5. Comparison between numerical and experimental data shows that the numerical model can predict the overall behaviour for the first three phases. Regarding the post-peak branch, however, there is an inherent inability to predict any of the softening found in the beams with the highest reinforcement ratio. This limitation can be explained by the plastic constitutive model used for concrete crushing, which cannot simulate the real behaviour beyond that. As expected, the transverse steel reinforcement ratio has nearly no impact on both experimental and numerical results (see Table 1 and Figure 5) since no shear cracks developed.

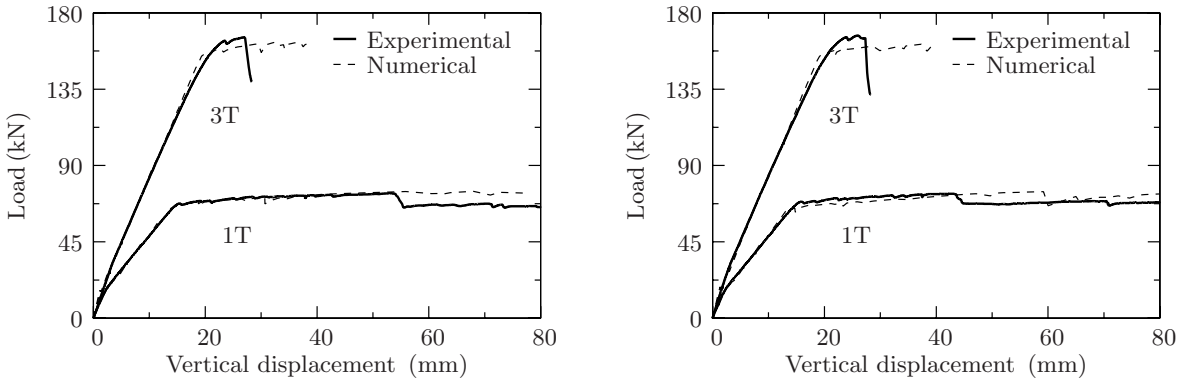


Figure 5: Load vs. vertical mid-span displacement for the: (a) lowest; and (b) highest transverse steel reinforcement ratios.

Table 2 summarises the main features and measures extracted for the models with highest transverse reinforcement ratio. Comparison between numerical and experimental results, shows the good agreement not only in overall displacements, but also in what regards cracking and yielding loads, maximum load, as well as in the curvature at mid-span (see also Figure 6).

		Specimen 1T		Specimen 3T	
		Exp.	Num.	Exp.	Num.
Cracking load (kN)		12.0	13.7	15.7	15.3
Yielding load (kN)		68.0	66.1	157.9	155.3
Max. load F_{max} (kN)		73.4	75.3	166.6	160.6
Mid-span vertical disp. (mm)	0.6 F_{max}	8.4	8.0	12.2	12.2
	0.7 F_{max}	10.2	10.6	14.0	14.2
	0.8 F_{max}	11.8	11.7	16.9	16.3
Curvature ($x10^{-3}m^{-1}$)	0.6 F_{max}	9.4	8.4	12.4	14.4
	0.7 F_{max}	11.2	13.4	14.3	16.2
	0.8 F_{max}	12.8	14.6	17.2	19.0

Table 2: Summary of the most relevant measures retrieved from numerical and experimental data.

4.2 Crack propagation

This section focuses in particular the surface crack pattern and crack openings. In terms of overall pattern, Figure 7 shows a comparison between numerical and experimental data for the lowest longitudinal reinforcement ratio. This specimen was selected for illustration purposes since it is the one showing the most visible cracks (in terms of its number and extension). To support this analysis, the region of interest is represented inside a dashed box in

Figure 7 (a), which can be directly compared with the monitored area shown in Figure 7 (b). Although real crack patterns can be quite random, the main features are well captured by the numerical model. For example, the model predicts well both spacing and number of active cracks. This is certainly related with the interaction between longitudinal reinforcement and cracks, as well as the relative slip between reinforcements and concrete. It should also be highlighted that the numerical model also predicts the horizontal cracks associated with concrete crushing (see Figure 7).

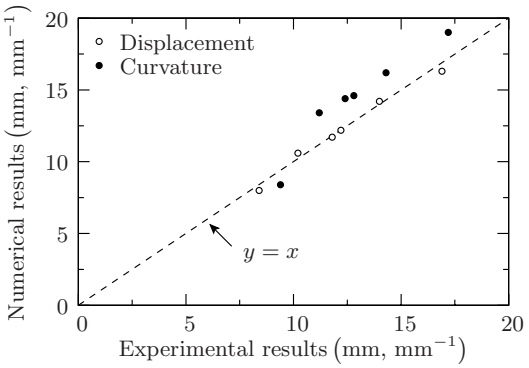


Figure 6: Comparison between displacements and curvatures.

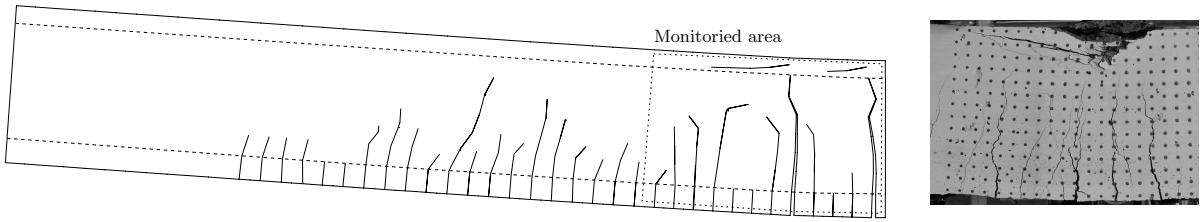


Figure 7: (a) Failure mode and crack pattern for beam 1T: (a) numerical; and (b) experimental results.

Figure 8 shows the sum of the opening of all the cracks located inside the monitored area at the level of the longitudinal reinforcement, for both types of beams and for the different stages identified in Figure 3. As a general remark, it can be mentioned that the crack openings are in good agreement with the experimental results, regardless of the amount of steel reinforcement and the failure mechanisms being quite different between models. As expected, the crack openings for the beam with higher longitudinal reinforcement ratio are significantly smaller (see also Figure 2).

5 EXTRAPOLATIONS BASED ON THE NUMERICAL MODEL

This section presents a study on the role of the longitudinal reinforcement ratio for LWAC beams. Models 1T and 3T are now re-analysed with models 0.5T, 2T and 5T, respectively having a reinforcement ratios of 0.58%, 2.06% and 4.63%. Models 0.5T and 5T are close to the minimum and maximum tensile reinforcement ratios allowed by the Eurocode 2 [34]. The objective is to assess the design code in what regards crack openings for LWAC under flexural loads. Figure 9(a) shows the load vs. vertical displacements for all numerical models and the impact different longitudinal reinforcement ratios have on the overall response.

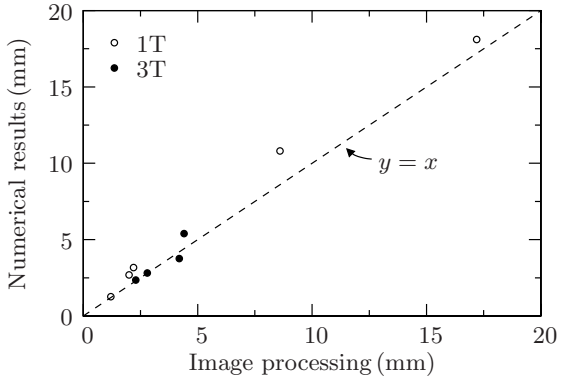


Figure 8: Comparison between experimental and numerical crack openings for both beams with the highest transverse reinforcement ratio.

The characteristic crack width can be defined as the crack opening corresponding to the 95% percentile value of the distribution. Since the numerical model has a discrete representation of cracks, a statistical model can be directly applied to quantify the characteristic opening at the level of the tensile reinforcements [35]. Figure 9(b) shows the characteristic opening obtained with the numerical models and the estimate from Eurocode 2 [34], the latter below the expected values for all the range of tested reinforcements. Although further studies are still needed, this could mean that the Eurocode 2 [3434] is not conservative for predicting characteristic crack openings in LWAC beams under flexural loads.

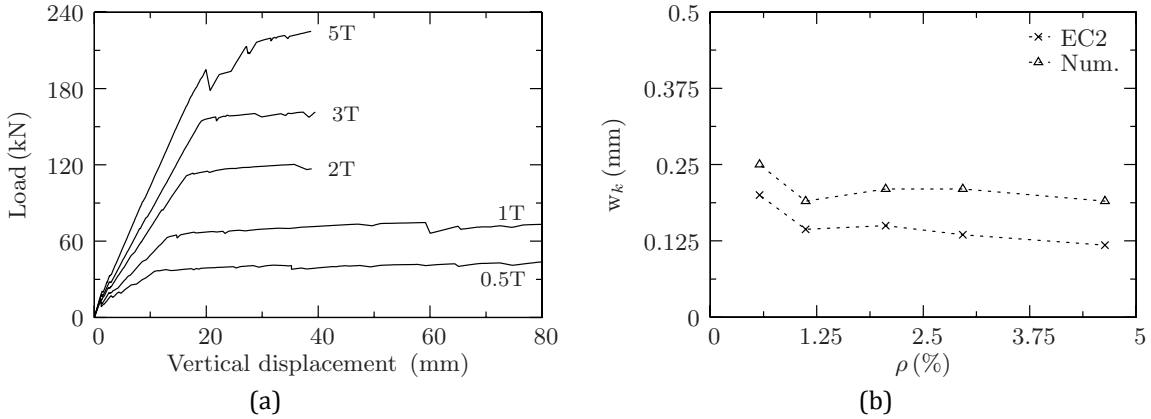


Figure 9: (a) Load vs. vertical mid-span displacement for all models; (b) Comparison between characteristic crack widths for 0.6 F_{max}.

6 CONCLUSIONS

This study validated a numerical model based on the embedment of strong discontinuities for the simulation of LWAC beams. The model relies on the assumption that the opening of the cracks occurs as if it were a rigid body motion, which simplifies the derived equations and implementation procedure. The model was shown to compare well with experimental data of LWAC beams. The numerical results are in good agreement, not only concerning displacements and curvatures, but also regarding the cracking and yielding loads, and maximum load. In addition, the crack pattern obtained from the simulations captures the main features observed in the laboratory, including the number and spacing of active cracks, as well as the overall opening of the cracks. This highlights the capability of the model to properly simulate the fracture and its interaction with the steel reinforcement, and the slip between reinforce-

ments and concrete. It was also shown that for the tested range of reinforcements, Eurocode 2 might underestimate the characteristic crack openings for LWAC beams under flexure. This conclusion, however, will need further studies.

ACKNOWLEDGEMENTS

The first author would like to acknowledge the support from the Australian Research Council through its Discovery Early Career Researcher Award (DE150101703) and from the Faculty of Engineering & Information Technologies, The University of Sydney, under the Faculty Research Cluster Program. Acknowledgment is also extended to FCT (Portuguese Foundation for Science and Technology), through ISISE, under project UID/ECI/04029/2013.

REFERENCES

[1] Belytschko, T. and T. Black, Elastic crack growth in finite elements with minimal remeshing. *International Journal for Numerical Methods in Engineering*, **45**(5), 601-620, 1999.

[2] Moës, N., J. Dolbow, and T. Belytschko, A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering*, **46**(1), 131-150, 1999.

[3] Wells, G.N. and L.J. Sluys, A new method for modelling cohesive cracks using finite elements. *International Journal for Numerical Methods in Engineering*, **50**(12), 2667-2682, 2001.

[4] Areias, P.M.A. and T. Belytschko, Analysis of three-dimensional crack initiation and propagation using the extended finite element method. *International Journal for Numerical Methods in Engineering*, **63**(5), 760-788, 2005.

[5] Dvorkin, E.N., A.M. Cuitiño, and G. Gioia, Finite elements with displacement interpolated embedded localization lines insensitive to mesh size and distortions. *International Journal for Numerical Methods in Engineering*, **30**(3), 541-564, 1990.

[6] Oliver, J., Modelling strong discontinuities in solid mechanics via strain softening constitutive equations. Part 2: Numerical simulation. *International Journal for Numerical Methods in Engineering*, **39**(21), 3601-3623, 1996.

[7] Jirásek, M. and T. Zimmermann, Embedded crack model: Part I. basic formulation. *International Journal for Numerical Methods in Engineering*, **50**(6), 1269-1290, 2001.

[8] Wells, G.N. and L.J. Sluys, Three-dimensional embedded discontinuity model for brittle fracture. *International Journal of Solids and Structures*, **38**(5), 897-913, 2001.

[9] Linder, C. and F. Armero, Finite elements with embedded strong discontinuities for the modeling of failure in solids. *International Journal for Numerical Methods in Engineering*, **72**(12), 1391-1433, 2007.

[10] Alfaiate, J., G.N. Wells, and L.J. Sluys, On the use of embedded discontinuity elements with crack path continuity for mode-I and mixed-mode fracture. *Engineering Fracture Mechanics*, **69**(6), 661-686, 2002.

[11] Dias-da-Costa, D., et al., An embedded formulation with conforming finite elements to capture strong discontinuities. *International Journal For Numerical Methods In Engineering*, **93**(2), 224-244, 2013.

[12] Ingraffea, A.R., et al., Fracture mechanics of bond in reinforced concrete. *Journal of Structural Engineering*, **110**(4), 871-890, 1984.

[13] Bocca, P., A. Carpinteri, and S. Valente, Mixed mode fracture of concrete. *International Journal of Solids and Structures*, **27**(9), 1139-1153, 1991.

[14] Areias, P.M.A., et al., Arbitrary bi-dimensional finite strain cohesive crack propagation. *Computational Mechanics*, **45**(1), 61-75, 2009.

[15] Chandra, S. and L. Berntsson, *Lightweight aggregate concrete*. 2002: William Andrew.

[16] EuroLightCon, *BE96-3942/R2: LWAC material properties. State-of-the-Art*. 1998: European Union–Brite EuRam III.

[17] EuroLightCon, *BE96-3942/R22: The economic potential of lightweight aggregate concrete in c.i.p. concrete bridges*. 2000: European Union–Brite EuRam III.

[18] Costa, H., *Lightweight aggregate structural concrete: precast and strengthening of existing structures.*, in *Civil Engineering Department* 2012, University of Coimbra: Coimbra.

[19] Sajedi, F. and P. Shafigh, High-Strength Lightweight Concrete Using Leca, Silica Fume, and Limestone. *Arabian Journal for Science and Engineering*, **37**(7), 1885-1893, 2012.

[20] Cui, H.Z., et al., Effect of lightweight aggregates on the mechanical properties and brittleness of lightweight aggregate concrete. *Construction and Building Materials*, **35**, 149-158, 2012.

[21] Bogas, J.A., A. Gomes, and M.F.C. Pereira, Self-compacting lightweight concrete produced with expanded clay aggregate. *Construction and Building Materials*, **35**, 1013-1022, 2012.

[22] Felicetti, R., P.G. Gambarova, and P. Bamonte, Thermal and mechanical properties of light-weight concrete exposed to high temperature. *Fire and Materials*, **37**(3), 200-216, 2013.

[23] Malvern, L.E., *Introduction to the mechanics of a continuous medium*. 1969, Englewood Cliffs, New Jersey: Prentice-Hall International. 713.

[24] Dias-da-Costa, D., et al., A comparative study on the modelling of discontinuous fracture by means of enriched nodal and element techniques and interface elements. *International Journal of Fracture*, **161**(1), 97-119, 2010.

[25] Graça-e-Costa, R., et al., A non-iterative approach for the modelling of quasi-brittle materials. *International Journal of Fracture*, **178**(1-2), 281-298, 2012.

[26] Graça-e-Costa, R., et al., Generalisation of non-iterative methods for the modelling of structures under non-proportional loading. *International Journal of Fracture*, **182**(1), 21-38, 2013.

[27] Carmo, R.N.F., et al., Influence of both concrete strength and transverse confinement on bending behavior of reinforced LWAC beams. *Engineering Structures*, **48**, 329-341, 2013.

[28] Carmo, R.N.F. and D. Dias-da-Costa, Tensile and flexural behaviour of LWAC members under short-term service loads. *Engineering Structures*, **92**, 142-155, 2015.

[29] *EN 10002-1. 2001. Tensile testing of metallic materials - Part 1: Method of Test at Ambient Temperature*, 2001, European Committee for Standardisation: Brussels, Belgium.

[30] Dias-da-Costa, D., J. Valença, and R.N.F. Carmo, Curvature assessment of reinforced concrete beams using photogrammetric techniques. *Materials and Structures*, 1-16, 2013.

[31] Valença, J., et al., Automatic crack monitoring using photogrammetry and image processing. *Measurement*, **46**(1), 433-441, 2013.

[32] Valença, J., D. Dias-da-Costa, and E.N.B.S. Júlio, Characterisation of concrete cracking during laboratorial tests using image processing. *Construction and Building Materials*, **28**(1), 607-615, 2012.

[33] fib, *Model Code for Concrete Structures 2010*. 2013: Wiley-VCH Verlag GmbH & Co. KGaA.

[34] CEN, *EN 1992-1-1: Eurocode 2: Design of concrete structures - Part 1-1: General rules and rules for buildings*, 2004, European Committee for Standardization (CEN).

[35] Borosnyól, A. and G.L. Balázs, Models for flexural cracking in concrete: The state of the art. *Structural Concrete*, **6**(2), 53-62, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

REINFORCEMENT CORROSION IN CONCRETE DUE TO CARBONATION AND CHLORIDE INGRESS UP AND BEYOND INDUCTION PERIOD

K. Hájková[1], L. Jendele[2], V. Šmilauer[1], T. Sajdlová[2], J. Červenka[2]

[1] Czech Technical University in Prague, Faculty of Civil Engineering
Thákurova 7; 166 29, Prague 6; CZ
e-mail: {karolina.hajkova, vit.smilauer}@fsv.cvut.cz

[2] Červenka Consulting s.r.o.
Na Hřebenkách 55, 150 00, Prague, CZ
{libor.jendele, tereza.sajdlova, jan.cervenka}@cervenka.cz

Keywords: Concrete, corrosion, propagation phase, carbonation, chloride ingress

Abstract. *Carbonation and chloride ingress are considered to be the most severe mechanisms for steel corrosion in reinforced concrete structures. Here, the service life is traditionally divided into two main phases; the initial and the propagation ones. The initial phase was explored recently and the results show high influence of cracks on accelerating of carbonation and chloride ingress in concrete structures. Our model focuses now on the propagation period and predicts x_{corr} for radial corrosion depth, including cracking and spalling of concrete cover. The presented models were implemented in ATENA software and demonstrate their application on a prestressed box-girder bridge. The simulation shows reinforcement corrosion due to carbonation and chloride ingress, and its impact on a bridge behavior during ULS analysis.*

1 INTRODUCTION

Reinforcement corrosion due to carbonation and chloride ingress are the most damaging mechanisms in reinforced concrete structures. The service life t_l of reinforced concrete structures is generally divided into two time phases; the initiation (induction) period t_i and the propagation period t_p. The initiation period for damaging mechanisms was described and validated in the previous paper [1] and preliminary results show strong influence of cracks to transport properties and acceleration of damaging mechanisms. For traditional cement-based materials, cracks 0.3 mm decrease induction time approximately 6 times for carbonation and approximately 9 times for chloride ingress of sea water. Preventing macrocracks and designing proper concrete is essential for durable concrete [1].

Our model focuses on the propagation period t_p where corrosion of reinforcing steel takes place. During this period, reinforcement decreases and accompanied with growing corrosion products. Corrosion of reinforcement is described according to the general diagram in Figure 1. A uniform corrosion (the most widespread form of corrosion) is characteristic for carbonation and a pitting corrosion (creation of small pits) for chlorides [2]. The cracking of concrete cover during propagation period $t_{p,cr}$ corresponds to the depth of corrosion $x_{corr,cr}$ and spalling of concrete cover $t_{p,sp}$ corresponds to the depth of penetration $x_{corr,sp}$ [3].

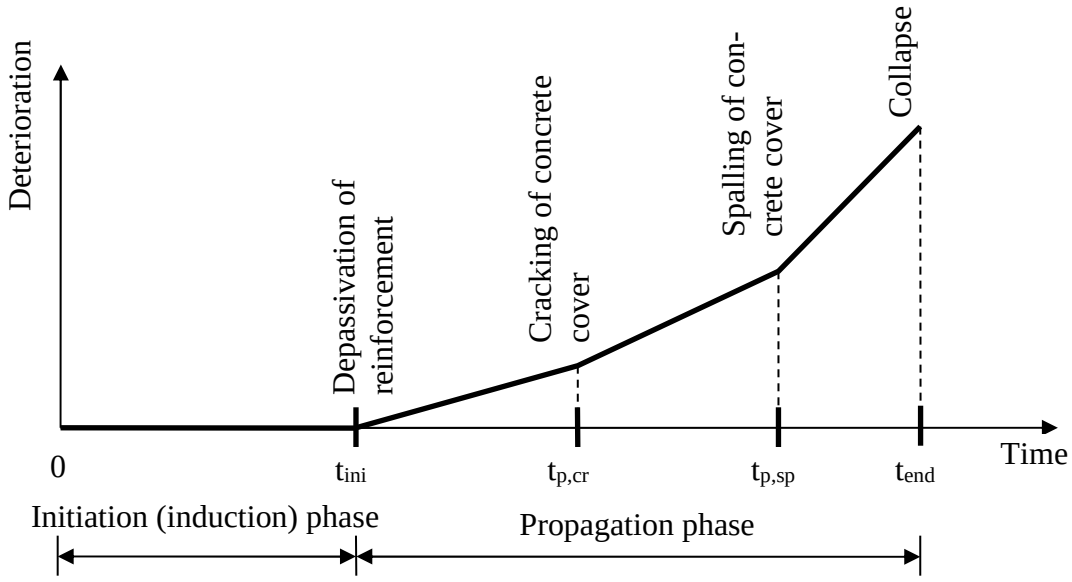


Figure 1: Initiation and propagation phase [3]

2 MODELS FOR PROPAGATION PHASE

2.1 Carbonation during propagation phase

The corrosion rate for the carbonation depends on the corrosion current density i_{corr} [$\mu A/cm^2$], which ranges between 0.1-10 (passive corrosion-high corrosion) and depends on the quality and the relative humidity of the concrete [5]. This model predicts amount of corroded steel during the whole propagation period t_p. The corrosion rate is based on Faraday's law [4], determined as follows:

$$\dot{x}_{\mathrm{corr}}(t) = 0.0116 i_{\mathrm{corr}}(t) \tag{1}$$

where $\dot{x}_{\mathrm{corr}}$ is the average corrosion rate in the radial direction [μm/year], i_{corr} is corrosion current density [μA/cm^2] and t is calculated time after the end of induction period [years].

By integration of Eq. (1), it is obtained the corroded depth for 1D propagation:

$$x_{\mathrm{corr}}(t) = \int_{t_{ini}}^{t} 0.0116 i_{\mathrm{corr}}(t) R_{\mathrm{corr}} \, \mathrm{d}t \tag{2}$$

where x_{corr} is the total amount of corroded steel in radial direction [mm] and R_{corr} is parameter, depends on the type of corrosion [-]. For uniform corrosion (carbonation) $R_{corr} = 1$, for pitting corrosion (chlorides) $R_{corr} = \langle 2; 4 \rangle$ according to [9] or $R_{corr} = \langle 4; 5.5 \rangle$ according to [10].

Effective bar diameter for both types of corrosion is obtained from:

$$d(t) = d_{ini} - \psi 2 x_{corr}(t) \tag{3}$$

where $d(t)$ is evolution of bar diameter in time t, d_{ini} is initial bar diameter [mm], ψ is uncertainty factor of the model [-], mean value $\psi = 1$ and x_{corr} is the total amount of corroded steel according to (2).

2.2 Chloride ingress during propagation phase

The corrosion rate for chlorides is more complicated because it is affected by concentration of chlorides in the concrete. Calculation of corrosion current density was formulated by Liu and Weyer's model [6]:

$$i_{corr} = 0.926 \cdot \exp\left[7.98 + 0.7771 \ln(1.69 C_t) - \frac{3006}{T} - 0.000116 R_C + 2.24 t^{-0.215} \right] \tag{4}$$

where i_{corr} is corrosion current density [μA/cm^2], C_t is total chloride content [kg/m^3 of concrete] on reinforcement which is determined from 1D nonstationary transport, T is temperature at the depth of reinforcement [K] and R_c is ohmic resistance of the cover concrete [Ω] [7] and t is time after initiation [years]:

$$R_C = \exp[8.03 - 0.549 \ln(1 + 1.69 C_t)] \tag{5}$$

The average corrosion rate in radial direction is determined further when plugging (6),(7) to (1). The total amount of corroded steel in radial direction stems from (2) and the effective bar diameter from (3).

2.3 Cracking of concrete cover

Cracking of concrete cover for both carbonation and chlorides can be estimated from DuraCrete model which provides realistic results [3]. The critical penetration depth of corroded steel $x_{corr,cr}$ is formulated as:

$$x_{corr,cr} = a_1 + a_2 \frac{C}{d_{ini}} + a_3 f_{t,ch} \tag{8}$$

where parameter a_1 is equal 7.44e-5 [m], parameter a_2 is equal 7.30e-6 [m], a_3 is [-1.74e-5 m/MPa], C is cover thickness of concrete [m], d_{ini} initial bar diameter [m], $f_{t,ch}$ is characteristic splitting tensile strength of concrete [MPa].

2.4 Spalling of concrete cover

The critical penetration depth of corroded steel $x_{corr,sp}$ for both carbonation and chlorides is calculated from [3] as:

$$x_{corr,sp} = \frac{w^d - w_0}{b} + x_{corr,cr} \tag{9}$$

where parameter b depends on the position of the bar (for top reinforcement 8.6 µm/µm and bottom 10.4 µm/ µm), w^d is critical crack width for spalling (characteristic value 1 mm), w_0 is width of initial crack (known from previous ATENA computation) and $x_{corr,cr}$ depth of corroded steel at the time of cracking [m].

After spalling of concrete cover, corrosion of reinforcement takes place in direct contact with the environment. To determine the rate of corrosion of reinforcement after spalling, Table 1 gives rates of reinforcement corrosion [8].

Corrosivity zone (ISO 9223)		Typical environment	Corrosion rate for first year (µm/y)	
Category	Description		Mild steel	Zinc
C1	Very low	Dry indoors	≤1,3	≤0,1
C2	Low	Arid/Urban inland	>1,3 a ≤25	>0,1 a ≤0,7
C3	Medium	Coastal and industrial	>25 a ≤50	>0,7 a ≤2,1
C4	High	Calm sea-shore	>50 a ≤80	>2,1 a ≤4,2
C5	Very High	Surf sea-shore	>80 a ≤200	>4,2 a ≤8,4
CX	Extreme	Ocean/Off-shore	>200 a ≤700	>8,4 a ≤25

Table 1: Corrosion rates of steel under atmospheric exposition

3 ANALYSIS OF A PRESTRESSED CONCRETE BRIDGE

The present model can be used for a wide range of structures from civil engineering. This is documented on an assessment of a prestressed box-girder concrete bridge of Mr. Pavel Wonka over the river Elbe in Pardubice, Czech Republic. It is beyond the scope of this paper to present detailed results from this analysis and hence, only the most important global results and results related to the presented durability analysis of the bridge are given. More details are available in [12], including material parameters, description of the bridge geometry and prestressing tendons etc.

The bridge was designed and erected between 1956 and 1959. The structure is depicted in Figure 2. The bridge consists of three arches, having spans 50 + 70 + 50 m. Average depth of cross sections is up to 3.5 m.

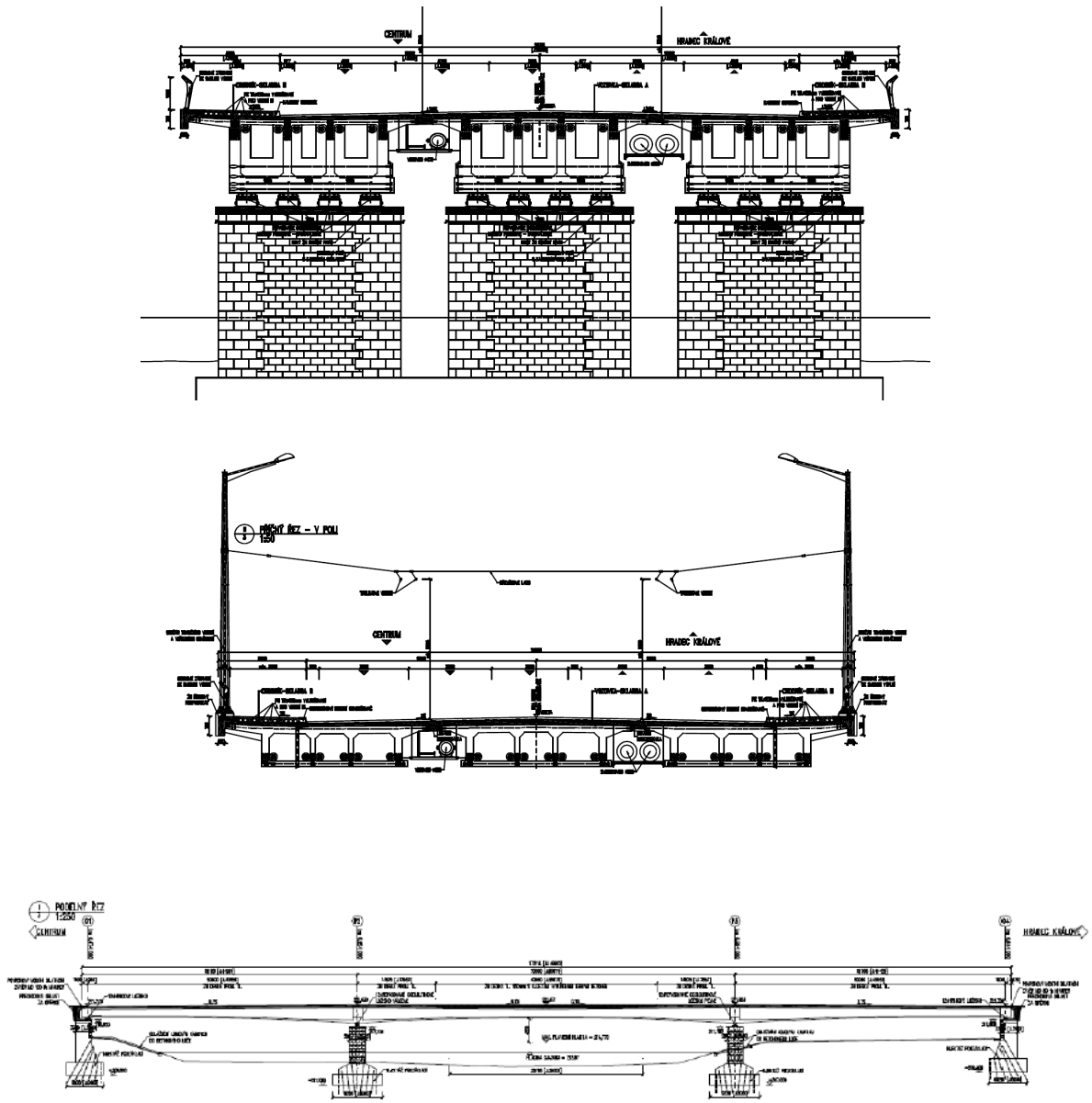


Figure 2: Dimensions of the bridge: drawing of cross section near the pillars, (top picture), mid-span cross section, (middle picture), and side (longitudinal) view of the bridge, (bottom picture) [11]

The structure was analyzed by program ATENA with implemented durability models [11]. The bridge is modeled by 4512 layered shell elements. The structure near supports and some other details are modeled by hexahedral and wedge solid elements. The pre-stressed tendons are realized by 3022 external cable truss elements, while the conventional reinforcement is introduced by embedded reinforcement within shell elements. A special material model for concrete and tendons has been employed with more details in [12].

The analysis of the bridge consists of three parts. The first analysis replicates in-situ load tests and measurements. The bridge was loaded by its self-weight and by tens of loaded trucks simulating a traffic load. It was used to calibrate the model of the bridge.

In the second part, numerical model was used to investigate service load state (SLS) and ultimate load state (ULS) of the bridge. Applied steps of the analysis and the associated loads were as follows:

1. Self-weight of the load-bearing structure and pre-stressing, (steps 1...10)
2. Weight of the top layers of the bridge, i.e. road etc., (steps 11...15)
3. Extra 35% of the load in the item 2, (steps 16...20)
4. 150% of the traffic load of the bridges according to ULS ČSN EN 1991-2, (steps 21...27)
5. Additional extra load according to the item 4. incremented up to failure of the bridge, (steps 28...78)

Deformation of the bridge at the 10th step and crack development at one end of the bridge is depicted in Figure 3 and Figure 4.

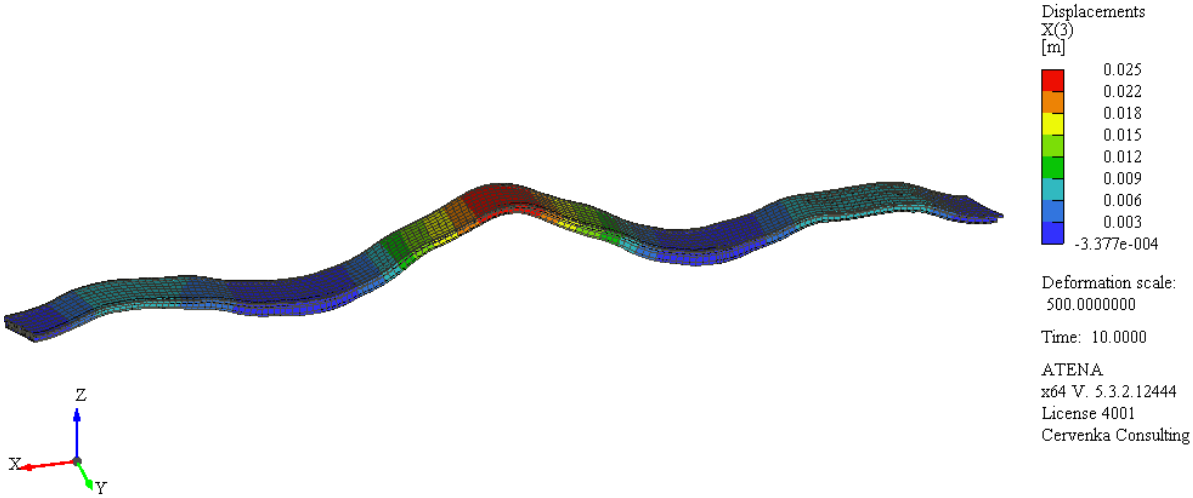


Figure 3: Deformation of the bridge at step 10, i.e. after self-weight and pre-stressing, (enlarged 500x)

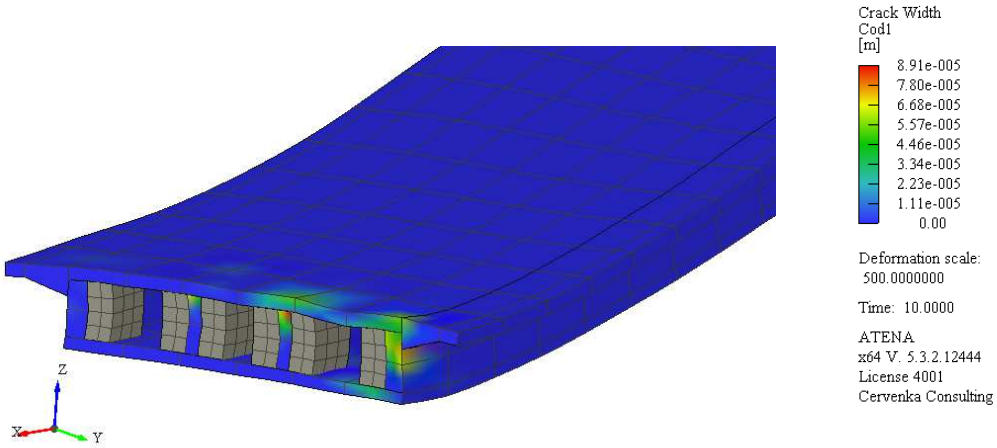


Figure 4: Crack development at the abutment of the bridge

Deformation of the bridge near its failure is shown further in Figure 5. It occurred at about 130% of the design load of the bridge, (i.e. steps 1...27 during loads 1-4). Load displacement diagram of the ULS analysis for centre of the central arch of the bridge is shown in Figure **6**, including prestress transfer.

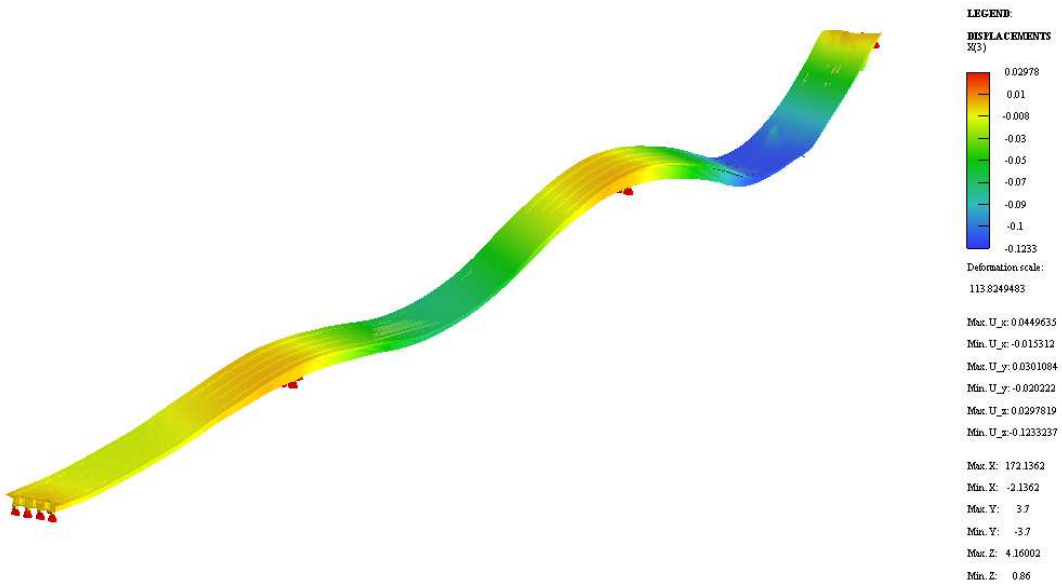


Figure 5: Deformation near failure of the bridge

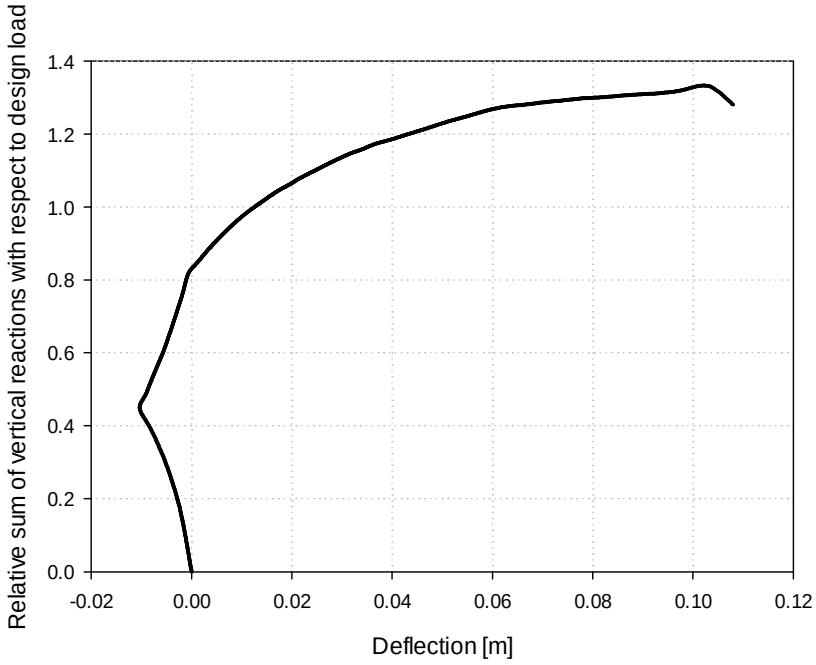


Figure 6: Load-displacement diagram of the ULS analysis - center of the central arch.

The third part of the analysis is devoted for durability study. According to direct concrete testing in 2003, concrete of the box-girder was classified as C35/45. An estimated composition yields CEM 42.5 350 kg/m^3 and water content 175 kg/m^3.

Concrete box-girder was loaded on its surface by carbonation and chlorides as described below. This analysis restarted from the step 27, (i.e. from the level of design load) and carbonation and chlorides load was applied for another 150 years simultaneously. The chlorides and carbonation loads used the following parameters:

- Carbonation: C_p = 350 kg/m^3, SCM = 0, W = 175 kg/m^3, CO_2 = 0.00036, RH = 0.60. Progressive period a_1 = 7.44e-5 m, a_2 = 7.30e-6 m, a_3 = -1.74e-5 m/MPa, $f_{t,ch}$ =

3.5 MPa, d_{ini} = 0.001 m, pitting corrosion R_{corr} =1, corrosion rate after spalling 30 μm/year.

- Chlorides: D_{ref} = 1.19e-7 m^2/day (mean value would be D_{ref} = 7.72e-13·86400 = 6.67e-08 m^2/day), t_{Dref} = 3650 days, m_{coeff} = 0.37, t_{mcoeff} = 10950 days, C_s = 0.103, Cl_{crit} = 0.0185. Progressive period a_1 = 7.44e-5 m, a_2 = 7.30e-6 m, a_3 = -1.74e-5 m/MPa, $f_{t,ch}$ = 3.5 MPa, w_d = 0.001 m, pitting corrosion R_{corr} = 3, corrosion rate after spalling 30 μm/year.

Chloride ingress assumed concentration of sea water on the surface; this resembles situation when salt brine water had leaked through insulation and the Cl concentration rose up substantially. According to Duracrete [3], spray zone from de-icing salts identified a slightly lower concentration C_s = 0.0776. Also, 90% confidence was considered for diffusivity D_{ref}, which is about twice higher than the mean value for this concrete strength class.

Figure **7** brings plot of maximum vertical displacement of the bridge vs. age of the structure. Note that the significant increase of the deflection at later times is due to tendons corrosion only as creep is neglected in computation and the force load is kept constant.

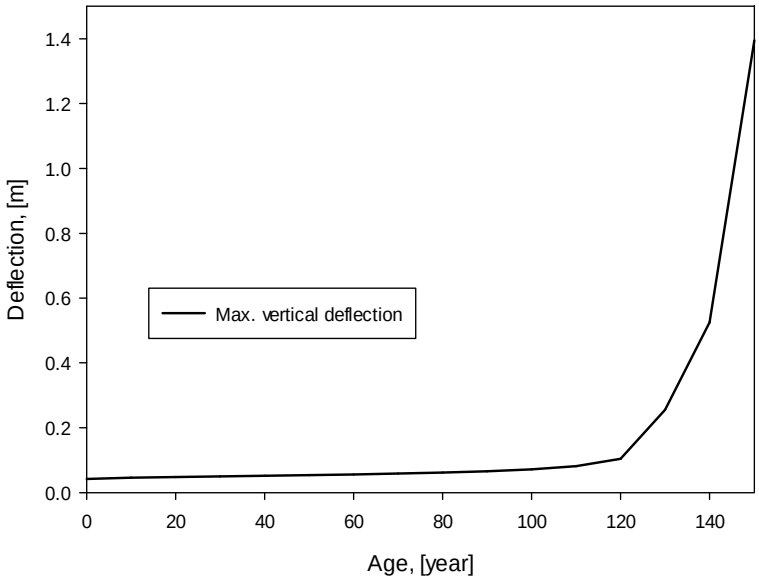


Figure 7 Maximum deflection of the bridge due to tendons corrosion

Figure 8 shows that carbonation depth is 58 mm for 150 years on uncracked concrete. Induction period for chlorides is much shorter, approximately 15 years with uncracked concrete cover 55 mm.

As far as steel corrosion is concerned, Figure 9 depicts calculated reduction coefficient for a pre-stressing tendon with concrete cover 55 mm. Reduction coefficient 1 means no corrosion, while value 0 signalizes total loss. For the first 12 years, the reinforcement does not corrode at all. Corrosion due to chlorides starts at 12 years. Corrosion due to carbonation begins after 110 years. At the age of 100 years about 60% of the reinforcement has corroded and at 150 years there is only about 35% of the original reinforcement. As expected, the effect of chlorides is much more devastating than that of carbonation. This is also documented by Fig-

ure 10, which shows total reinforcement reduction factor for the pre-stressing tendons at age of 150 years.

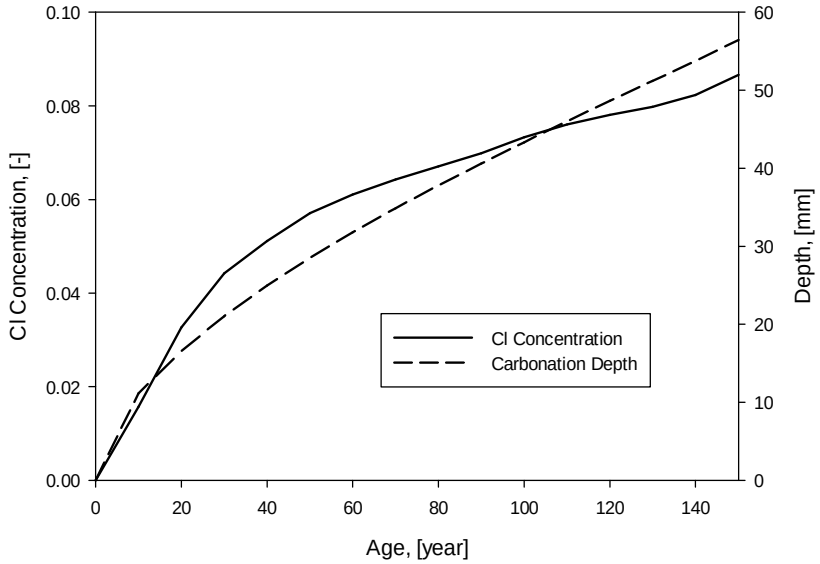


Figure 8: Characteristic carbonation depth for uncracked concrete. Chloride concentration for uncracked concrete at 55 mm from exposed surface

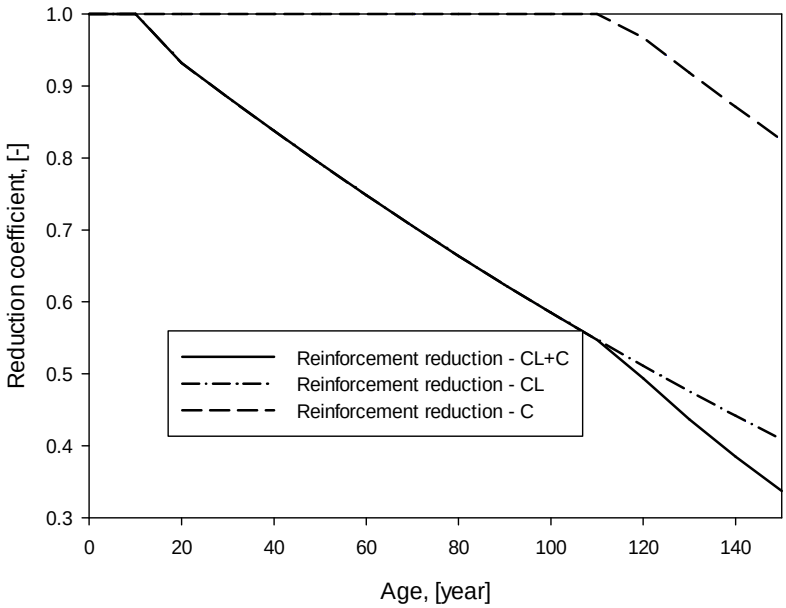


Figure 9 Computed reduction coefficient for a tendon, concrete cover 55 mm with the influence of cracks

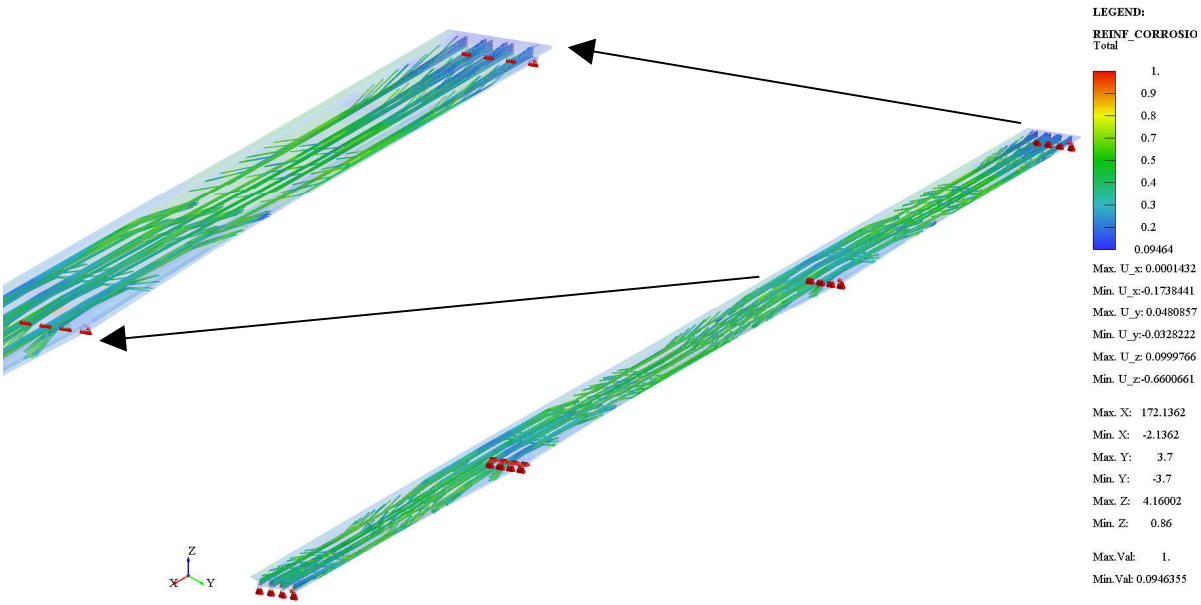


Figure 10: Total reduction coefficients for tendons at 150 years

4 CONCLUSIONS

The implemented durability model for reinforcement corrosion due to carbonation and chloride ingress focused on the propagation period where corrosion of reinforcement and spalling take place.

- The presented models provide induction time and reduction of tendons during the propagation period. The collapse of the bridge can be predicted when combined with ULS analysis.
- Results for a prestressed box-girder bridge indicate that tendon corrosion would start at 12 years if high external chloride concentration is assumed. At 100 years, about 60% of tendons' cross section remains and at 150 years this drops to only 30%.
- Durability analysis could be an integral part of bridge assessment, using slightly extended model. It allows prediction of negative effects of external environmental conditions on the structure and exploring deterioration.

Acknowledgement: We gratefully acknowledge the financial support from the Technology Agency of the Czech Republic TAČR under the project TA04031458 and SGS12/116/OHK1/2T/11 granted by Czech Technical University in Prague.

REFERENCES

[1] V. Šmilauer, L. Jendele, J. Červenka, Prediction of Carbonation and Chloride Ingress in Cracked Concrete Structures, *International conference on civil structural and environmental engineering computing*, 2013.

[2] F. Duprat, Reliability of RC Beams Under Chloride-ingress. *Construction and Building Materials*, **21**, 1605-1616, 2007.

[3] DuraCrete, Final Technical Report, 2000.

[4] J. Rodriguez, L. M. Ortega, J. Casal & J. M. Diez, Corrosion of Reinforcement and Service Life of Concrete Structures. In Proc. of Int. Conf. on Durability of Building Materials and Components, 7, Stockholm, 1:117-126, 1996.

[5] Page CL, Nature and properties of concrete in relation to environment corrosion, *Corrosion of Steel in Concrete*, Aachen, 1992.

[6] Liu Y, Weyers R.E., Modeling the Dynamic Corrosion Process in Chloride Contaminated Concrete Structures, *Cement and Concrete Research*, **28**(3), 365-367, 1998.

[7] Liu Y., Modelling the Time-to-corrosion Cracking of the Cover Concrete in Chloride Contaminated Reinforced Concrete Structures, *Ph.D. dissertation, Virginia Polytechnic Institute*, 1996.

[8] Spec-net, Corrosivity zones for steel construction [online] available from: http://www.spec-net.com.au/press/1014/gaa_081014/Corrosivity-Zones-for-Steel-Construction-Galvanizers-Association, 2015.

[9] J. A. Gonzales, C. Andrade, C. Alonso & S. Feliu, Comparison of Rates of General Corrosion and Maximum Pitting Penetration on Concrete Embedded Steel Reinforcement, *Cement and Concrete Research*, **25**(2), 257-264, 1995.

[10] M. S. Darmawan & M. G. Stewart, Effect of Pitting Corrosion on Capacity of Prestressing Wires, *Magazine of Concrete Research*, **59**(2), 131-139, 2007.

[11] V. Červenka, J. Cervenka, L. Jendele, *Atena Program Documentation*, Part 1-7, Červenka Consulting, Prague, Czech Republic, 2000-2013.

[12] J. Červenka and T. Sajdlová, *Most Pavla Wonky v Pardubicích*, Nelineární model nosné konstrukce - reprodukce zatěžovací zkoušky a výpočet únosnosti. (in Czech). Červenka Consulting, Prague, Czech Republic, 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THE STOCHASTIC FINITE ELEMENT METHOD FOR NUCLEAR APPLICATIONS

José David Arregui Mena[1], Lee Margetts[1], Llion Evans[1, 2], D. V. Griffiths[3], Anton Shterenlikht[4], Luis Cebamanos[5] and Paul M Mummery[2]

[1] University of Manchester
Oxford Road, Manchester, M13 9PL
e-mail: jose.arregui-mena@manchester.ac.uk, lee.margetts@manchester.ac.uk, paul.m.mummery@manchester.ac.uk

[2] Culham Centre for Fusion Energy
Culham Science Centre, Abingdon, OX14 3DB
e-mail: Llion.Evans@ukaea.uk

[3] Colorado School of Mines
1500 Illinois Street, Golden CO 80401, USA
e-mail: d.v.griffiths@mines.edu

[4] University of Bristol, Queen's Building
University Walk, Clifton BS8 1TR
e-mail: mexas@bristol.ac.uk

[5] University of Edinburgh
James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD
e-mail: l.cebamanos@epcc.ed.ac.uk

Keywords: Random Finite Element Method, Nuclear Energy, Nuclear Fusion.

Abstract. *Nuclear materials are subjected to demanding environments, encountering high temperature gradients and fast neutron fluxes that gradually damage its structure and therefore change the material properties. Some components of a nuclear reactor determine its lifetime, such as the graphite core and steel pressure vessel for fission reactors. In case of fusion reactors the tungsten divertor is expected to be replaced several times during its lifespan. All these materials contain defects and spatial material variability that may contribute to the failure of the component. The Stochastic Finite Element Method or a Random Finite Element Method was chosen in this research to model the spatial material variability in nuclear graphite and other key components of nuclear reactors. This research describes how a direct Monte Carlo Simulation approach was adapted to simulate the calibration of a random field and the modelling of these defects for nuclear graphite. It is also suggested that this methodology can be applied to fusion reactor modelling.*

1 INTRODUCTION

The Stochastic Finite Element Method (SFEM) or Random Finite Element Method (RFEM) is an extension of the FEM that considers the uncertainty of a system that arises through variations in initial conditions, forces, materials or geometry. These uncertainties can be taken into account by representing them through several mathematical tools i.e. random fields [1] or random media techniques [2]. The most common tool to represent a random variable with the SFEM is random fields, which can be described as a series of indexed random variables that describe a stochastic system [1]. A random field can potentially represent the stochastic nature of a system by including a) stochastic forces, b) spatial variability of material properties, or a c) randomly generated geometry (Figure 1). Several branches of the SFEM can be used to include, represent and solve a stochastic model; for a more in detail description of the SFEM the reader is referred to reference [3].

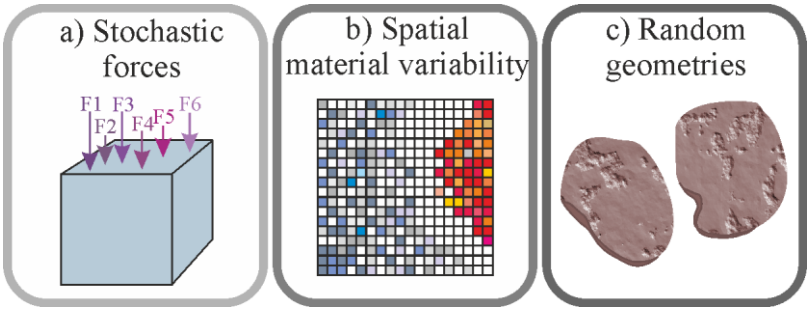


Figure 1. Stochastic variables that can be represented with random fields a) Stochastic forces, b) Spatial material variability, c) Random geometries

Random processes or uncertainties are present up to some degree in most systems. In nuclear reactors several stochastic variables arise from the defects of materials and the environment created by nuclear chain reactions. Figure 2 shows the types of chain reactions occurring in fission and fusion reactors.

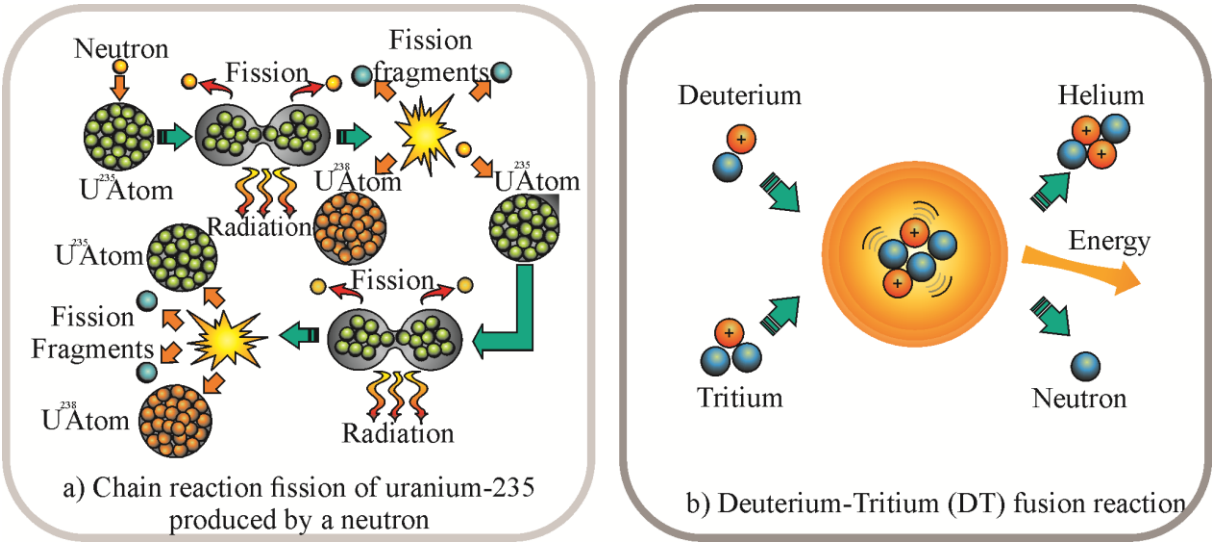


Figure 2. Types of chain reactions in fission and fusion reactors a) Fission reaction of uranium-235 (adapted from [4]), b) Fusion reaction of deuterium and tritium (adapted from [5])

The products of fusion and fission chain reactions are fast neutrons, the high energy of these particles produce a progressive damage to the components at the vicinity of these reactions. Furthermore, nuclear reactions can produce a transmutation of an element and generate heat. In this study we propose a methodology to study the effect of these environmental variables combined with the spatial material variability. The first case study proposes a methodology to predict the influence of spatial variability in nuclear graphite for British nuclear reactors. The second case study implements a random field generator of porosity to reproduce similar defects to the ones found in the tungsten components of a fusion reactor.

2 SFEM FOR FISSION NUCLEAR REACTORS MATERIALS

A Random Finite Element Method with a direct Monte Carlo Simulation scheme is proposed to study the UK's graphite moderated reactors called Advanced Gas-cooled Reactors (AGRs). The core of these reactors is formed by thousands of graphite bricks that serve as a moderator, a structural support and also provide a repository for fuel rods and other instruments. A summary of the most important parts of an AGR and its core components is shown in Figure 3.

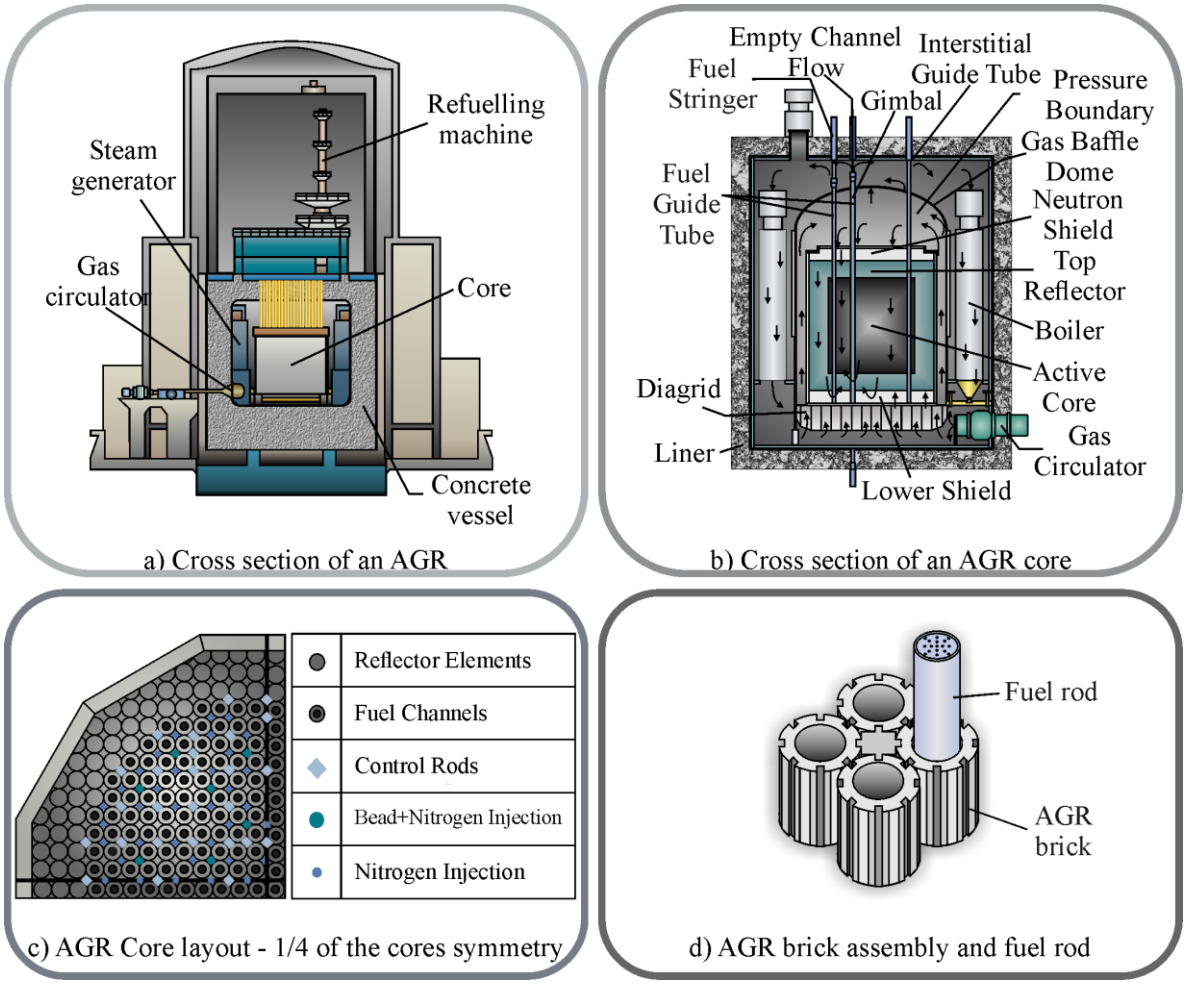


Figure 3. Parts of an AGR a) Cross section of an AGR (adapted from [6]), b) Cross section of an AGR core (adapted from [7]), c) AGR core layout (adapted from [8]), d) AGR brick assembly and fuel rod

The graphite bricks are subjected to several ageing processes that alter the material properties, produce dimensional changes induced by irradiation and temperature gradients and reduce the weight of the components. As a consequence the geometry of the graphite bricks could change. This, in turn, would cause a stress concentration, which may lead to cracking. The total strains ($\Delta\varepsilon^{Total}$) that a graphite brick suffers come from a combination of 4 factors that are: elastic strains ($\Delta\varepsilon^{e}$), creep strains ($\Delta\varepsilon^{c}$), dimensional strains ($\Delta\varepsilon^{dc}$) and thermal strains ($\Delta\varepsilon^{th}$). All these strains are represented in Equation 1.

$$\Delta\varepsilon^{Total} = \Delta\varepsilon^{e} + \Delta\varepsilon^{c} + \Delta\varepsilon^{dc} + \Delta\varepsilon^{th} \tag{1}$$

Irradiation creep strains are another product of the neutron irradiation, it is considered to relieve high stresses caused by dimensional changes that are produced by irradiation and thermal strains. The deformation of the core and graphite components can potentially impede the normal cooling of the reactor and the loading and unloading of fuel rods. Furthermore, graphite components cannot be replaced in AGRs, this makes them a critical component of AGR lifetime. The complex interactions that damage the graphite core lead to uncertainties in graphite core lifetime predictions and inspections cannot fully determine the actual condition of the components. In order to provide a different approach, the SFEM was adapted to incorporate spatial variability of material properties to the modeling of the thermomechanical response of graphite components.

The hypothesis of this research is that spatial variability in the coefficient of thermal expansion and Young's modulus can have a significant effect on the mechanical performance of nuclear graphite during its in-service lifetime. In order to test this hypothesis, the open source software ParaFEM [9] is being extended to interface with third party software libraries that can generate spatially variable 3D random fields for any material property.

2.1 Methodology

A deterministic and a stochastic approach were used to compare the influence of spatial material variability of material properties. For this comparison a profile of temperature change was applied to the deterministic and stochastic approach. Boundary conditions, temperature profiles are discussed in the next subsections.

In case of the stochastic approach several realizations of random fields were created and then incorporated into a thermoelastic finite element code using a Monte Carlo scheme. At the last step the influence of spatial variability of the material properties on the response variables e.g. strains and stress distributions in nuclear graphite were estimated. To recreate the spatial material variability the local average subdivision method was used. This random field generator requires the mean values, standard deviation and correlation length as inputs. An in detail explanation of the local average subdivision method can be found in reference [10]. The realisations produced with this random field generator were used as an input in the finite element software ParaFEM. For the deterministic approach only one simulation was created with an average of the material properties listed in Subsection 2.2. Both methodologies are summarised in Figure 4.

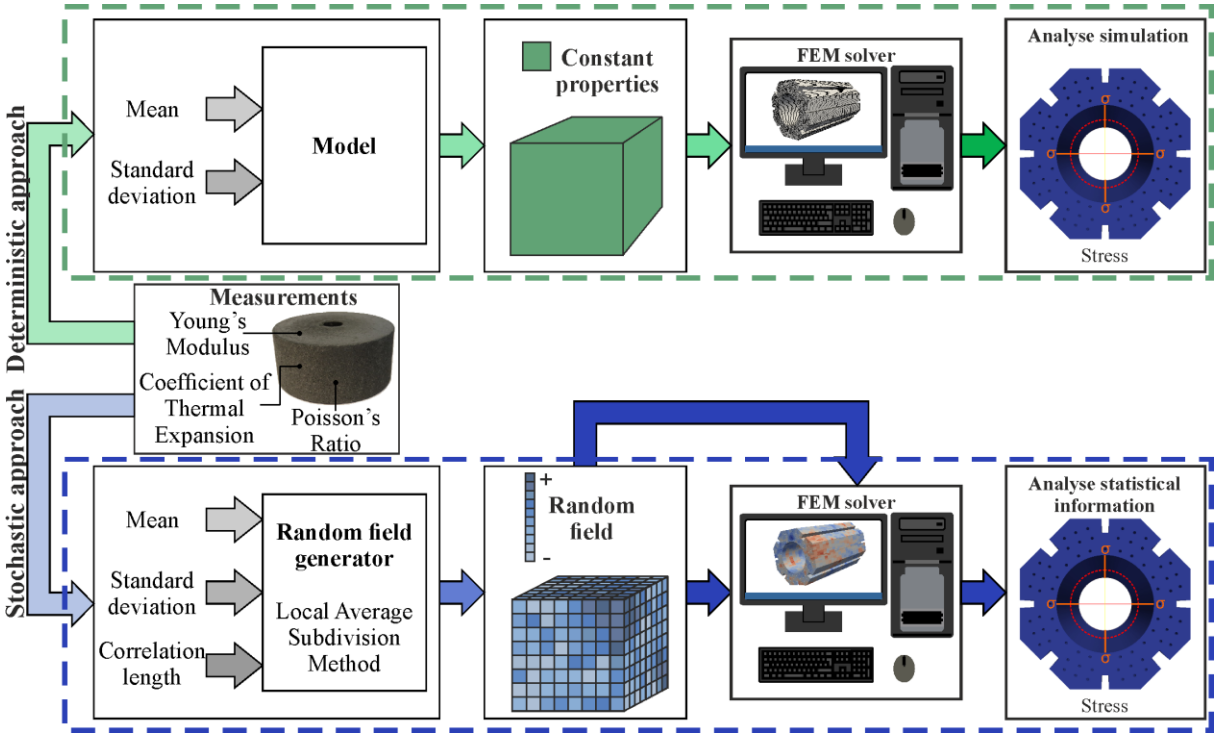


Figure 4. Methodologies for the deterministic and stochastic approach

2.2 Material properties

The material properties chosen for these realisations were obtained from [11] (Table 1); the material properties correspond to the average values for isotropic graphite. The standard deviation for the coefficient of thermal expansion was estimated by the authors. The correlation length was also defined by the authors due to the lack of calculations for this parameter. The correlation length is an important parameter required to describe the fluctuation of material properties in a random field. This value is a given length where the values of the random fields are correlated, beyond this distance the correlation between two data points tend to zero.

Material properties for virgin isotropic graphite	Mean values	Standard deviation
Mean coefficient of thermal expansion	4.35×10^{-6} (mm/mm°C) [11]	4.35×10^{-7} (mm/mm°C)
Poisson's ratio	0.2 [11]	
Dynamic Young's Modulus	10 GPa [11]	
Correlation length in all directions	100 mm	

Table 1: Material properties for all the analysis.

A relationship between Young's modulus and Coefficient of Thermal Expansion (CTE) was used. This relationship is given by equation (2)

$$E(\text{random}) = \alpha(\text{random})/\text{constant} \quad (2)$$

where E is Young's Modulus, α is the CTE and a constant. Therefore each Young's modulus value for an element would be calculated from the CTE random field value using equation (2).

2.3 Boundary conditions and field variables

The boundary conditions used for this research were a modified version of the 3-2-1 rule. The objectives of these boundary conditions are to allow the free thermal expansion, to minimize the stress concentrations around the fixed nodes and to remove any rigid body motions. Figure 5 shows the chosen boundary conditions selected for all the simulations.

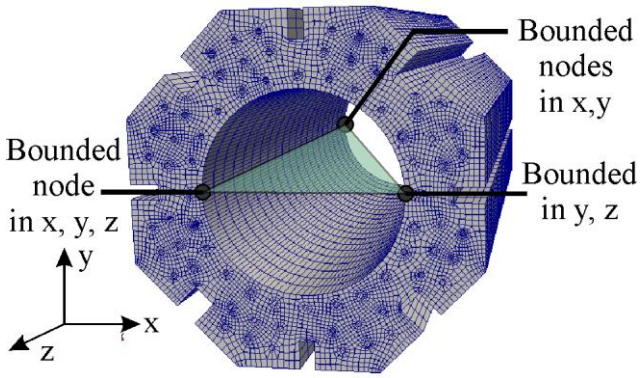


Figure 5. Boundary conditions used for this study

A similar temperature change profile as the ones found in AGR brick was applied to produce thermal expansion. Figure 6 shows this temperature change profile applied to the simulations.

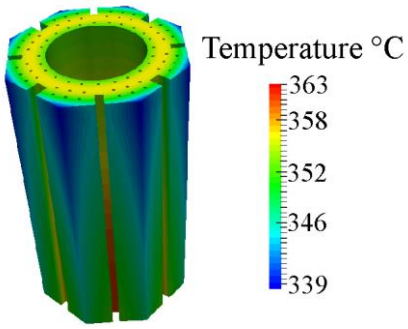


Figure 6. Temperature change profile for all the simulations

2.4 Results and discussion

The results of the deterministic and the stochastic approaches are compared in Figure 7. As can be seen on Figure 7 the deterministic approach produces stresses orders of magnitude less than by stochastic modelling. In the model where the stochastic spatial material variability was considered the stresses can reach a magnitude of 4.5 MPa. It is important to note that these stress concentrations would not lead to the cracking of graphite bricks. However, the combination of material spatial variability with thermal expansion, irradiation induced dimensional change and creep strains can increase these stress concentrations that might lead to the cracking of a brick.

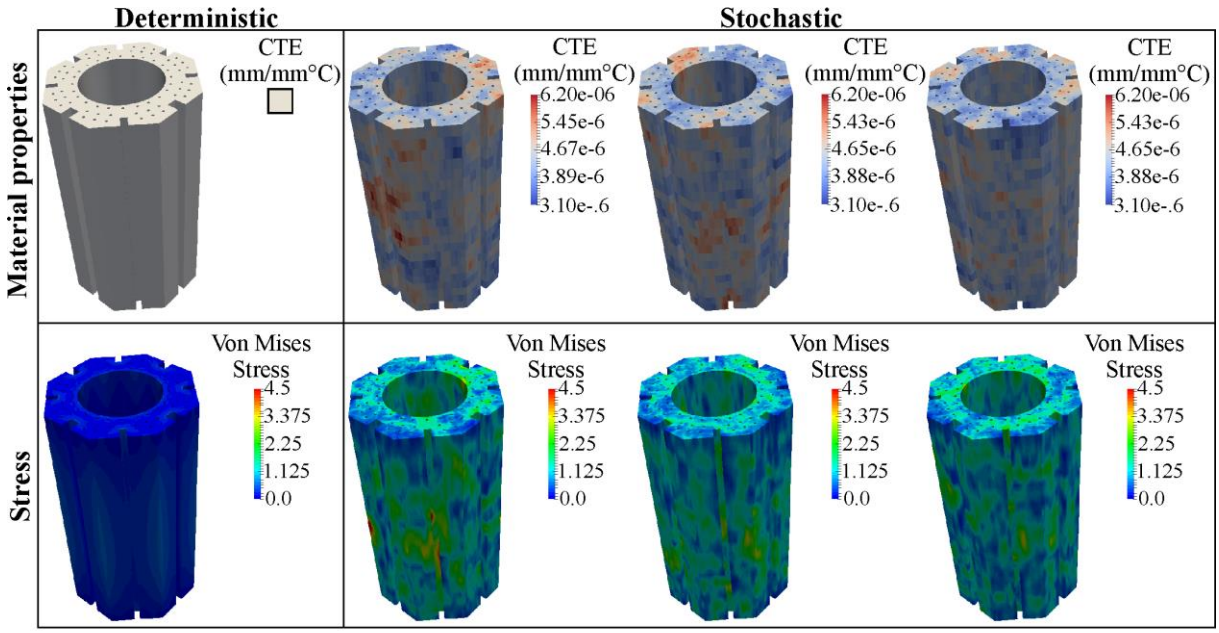


Figure 7. Von Misses stress comparison between the deterministic and stochastic approach

2.5 Conclusions

Spatial material variability is not usually considered for the modelling of nuclear graphite components. Variability of mechanical properties may be one of the factors that lead to the early cracking of nuclear components. The results show how material spatial variability increases the magnitude of Von Mises stress. Furthermore, stress concentrations are found in regions where they are not expected.

The current simulations only use only the estimations to recreate the input parameters for the correlation length. A new study focuses on calibrating the random fields from density and Young's modulus measurements made in a billet (Figure 8). These measurements will serve to obtain the average, standard deviation and correlation length for future simulations.

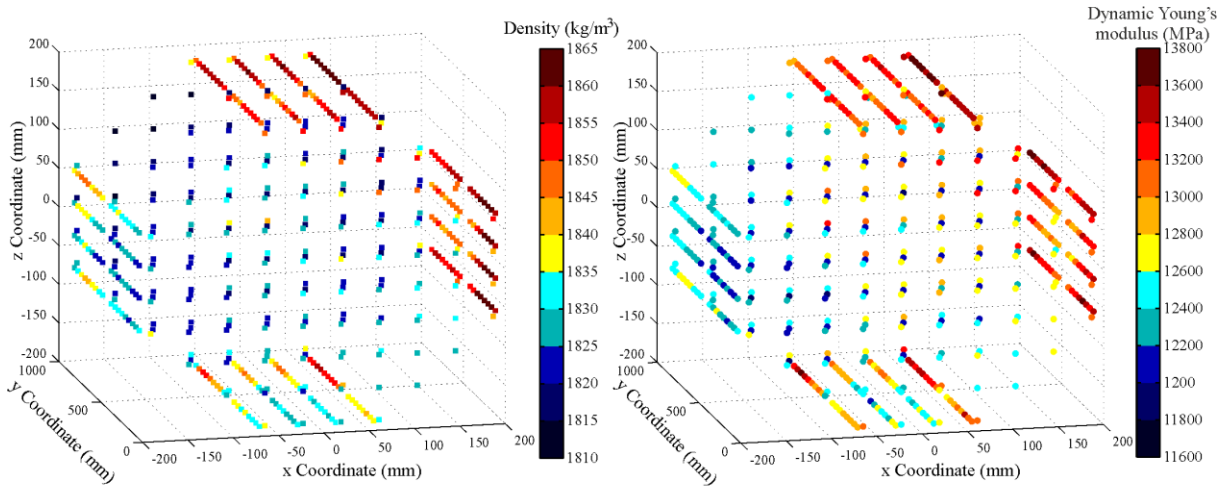


Figure 8. Density and dynamic Young's modulus measured in a graphite billet

Further improvements to the simulation of graphite modelling are to include a multiscale model by combining cellular automata modelling and the Finite Element Method (FEM) [12]. The aforementioned cellular automata methodology requires information of the crystal orientation of the material; this data has been obtained for nuclear graphite [13]. A random field generator can be modified to use these to implement the cellular automata methodology.

3 SFEM FOR FUSION NUCLEAR REACTORS MATERIALS

Materials in a fusion nuclear reactor need to endure extreme conditions. High thermal loads, high neutron fluxes and plasma erosion produce changes and damage in the materials that surround the reactions [14]. Two different designs lead the research to create future fusion reactors: magnetic confinement and inertial confinement. The magnetic confinement approach is currently considered to be the most technologically developed concept of the two. Numerous relatively small tokomaks already exist, but it is known that due to scaling laws that energy output increases at a greater rate than the energy input as plasma volume increases. For this reason, a much larger tokomak is being built in Cadarache in France called ITER [15]. Figure 9 shows the design and interior cross section of a tokamak. The objective of this reactor is to demonstrate the scientific feasibility of sustaining fusion reactions at electricity generating scales and producing an output ten times that of the energy required to initiate the process [16].

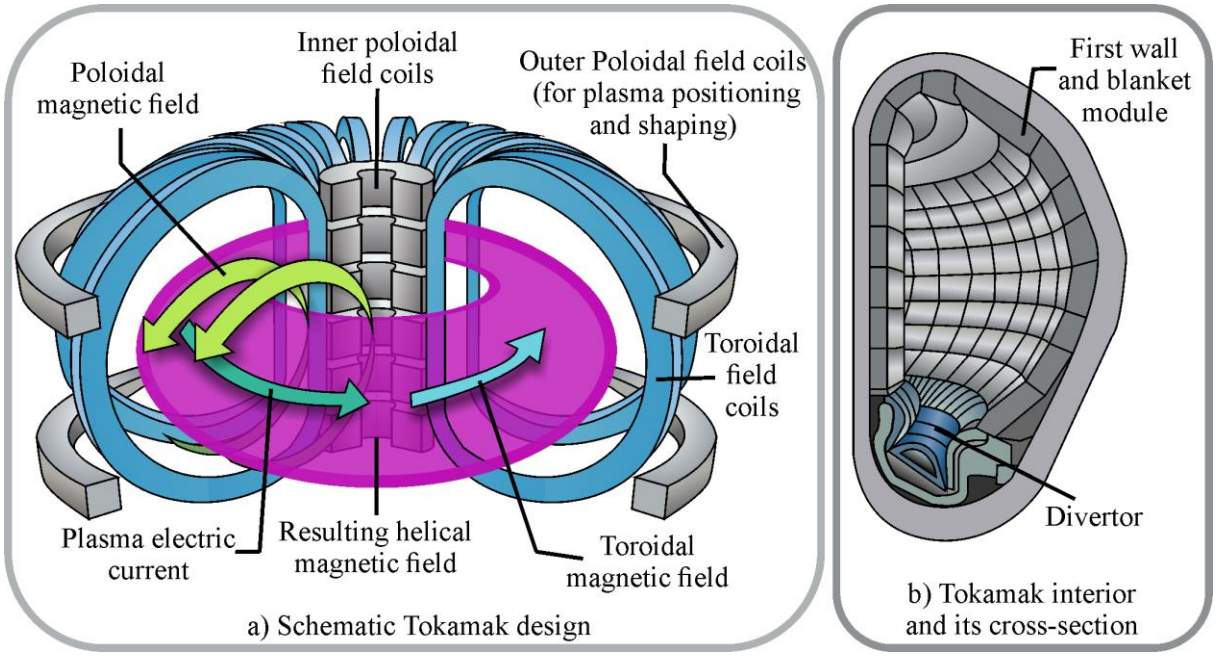


Figure 9. Tokamak system, a) Schematic of a Tokamak design (adapted from [17]), b) Tokamak interior and its cross-section (adapted from [18])

Vacuum plasma-sprayed tungsten (VPS-W) is a candidate for coating the walls of future fusion components [19]. The VPS-W is a complex material that contains a random porous structure; this porosity strongly influences the thermomechanical behavior of this material [20]. A previous study by Zivelonghi [21] combines synchrotron tomography and Image-Based FEM to model the thermomechanical response of VPS-W. Image-Based FEM uses stacks of 2D slices of a material to recreate 3D models of the intricate structures. Image-Based FEM have also been implemented to study other fusion materials such as carbon fiber composites [22]. However, Image-Based FEM is limited to the size, resolution and capacity

of the tomography equipment, thus large samples with highly detailed porous structure cannot be model through this technique. The information of small geometrical models can be used to calibrate a random field that mimics the porous structure found in VPS-W. These random fields can be incorporated to a SFEM model to calculate the response of larger components covered with VPS-W.

The next subsections describe the random field generator capable of recreating porous structures and some preliminary results.

3.1 Porosity random field generator

A random field generator proposed by Paiboon et al. [23] is capable of producing random fields that represent porosity in materials. This random field generator allows the user to control the volume fraction of porosity and pore size by controlling the correlation length and percentage of porosity. Examples of some realisations are shown in Figure 10.

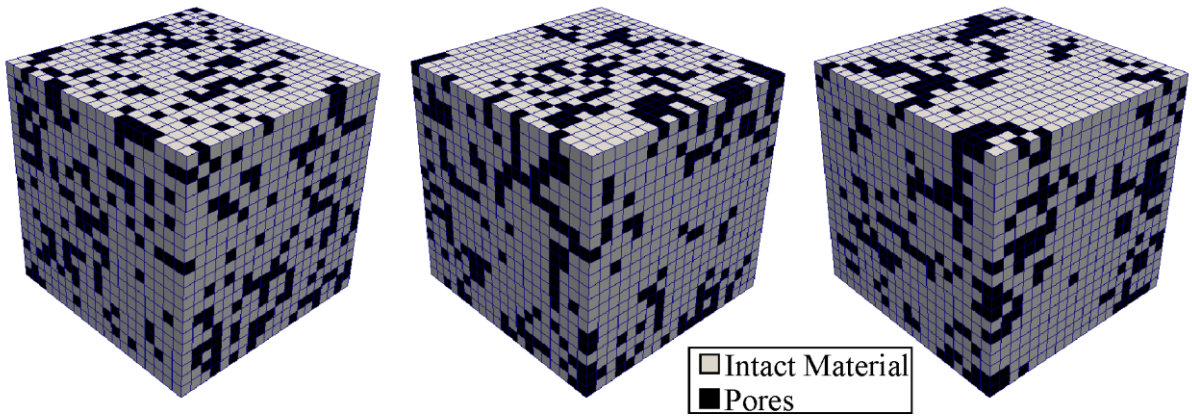


Figure 10. Realisations of porous materials

This random field generator is a modified version of the Local Average Subdivision method [10]. In this random field the intact material is assigned with a given material property value and the porous material property value are 100 times smaller than the value for intact material.

Several types of software are available to create 3D models from computer tomography data. The geometrical information of the pore can be manipulated and used to obtain the average size of the pore structure. Once the mean size of the pores has been found, it can be used as an input file for the aforementioned random field generator. The final step would be to assign the random field material property values to model the thermomechanical properties of larger components.

To test the influence of porosity in tungsten some preliminary results were obtained by calculating the effective material properties of one realisation.

3.2 Material properties

The material properties used for the intact material in the simulations are given in Table 2.

Material properties for virgin isotropic graphite	Values
Poisson's ratio	0.28
Young's Modulus	411 GPa
Correlation length in all directions	0.1 mm
Target porosity for porous simulation	0.8%

Table 2: Material properties for intact tungsten material.

3.3 Boundary conditions and field variables

The finite element mesh is formed by a cubic block of materials of 1000 mm side with 100×100×100 8 node cubic elements. The chosen boundary conditions are known as "tied freedom", with these boundary conditions the cube is forced to be deformed in an ideal cube, Figure 27 shows the geometry and boundary conditions. From this deformation the calculation of the elastic properties can be easily calculated from the elastic theory. The boundary conditions are: the base of the cube can move in x and y direction, the back left can move in the y-z plane, the right face in only allowed to movie in the z-x plane. The freedoms of the top face are tied in the z direction, the front left face freedoms are restrained in the in the y direction, finally the x freedoms are tied in the right front face. A vertical force is applied at the top of the cube with a value of Q = L x L.

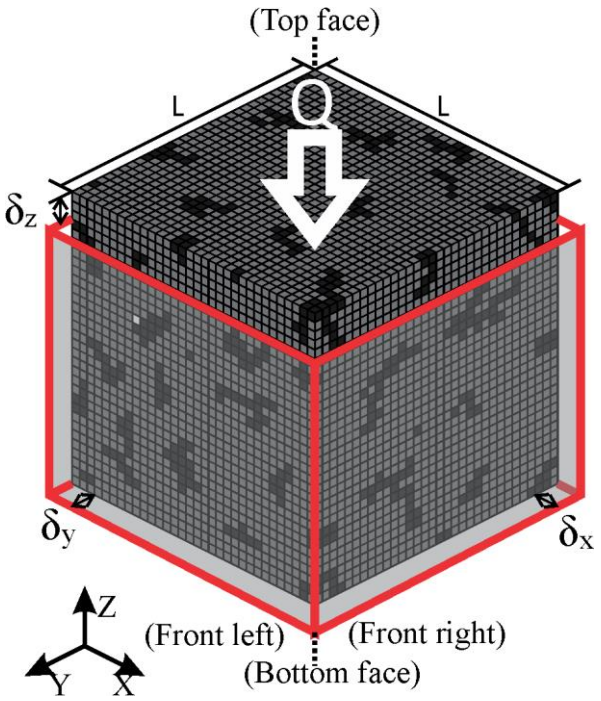


Figure 11. Compression test of a porous material to determine the influence of porosity in graphite

In order to obtain the effective properties of graphite the values for the deformations in each direction δ_x, δ_y and δ_z are retrieved from each simulation and substituted in the following equations:

$$E = \frac{Q}{L\delta_z}$$
$$\upsilon_x = \frac{\delta_x}{\delta_z} \quad (3)$$
$$\upsilon_y = \frac{\delta_y}{\delta_z}$$

where E is the effective elastic Young's modulus, Q is the loading force at the top of the cube and ν_x and ν_y are the effective Poisson's ratios in x and y direction.

3.4 Results and discussion

The results of the effective material properties of the deterministic and the stochastic simulations are summarised in Figure 12.

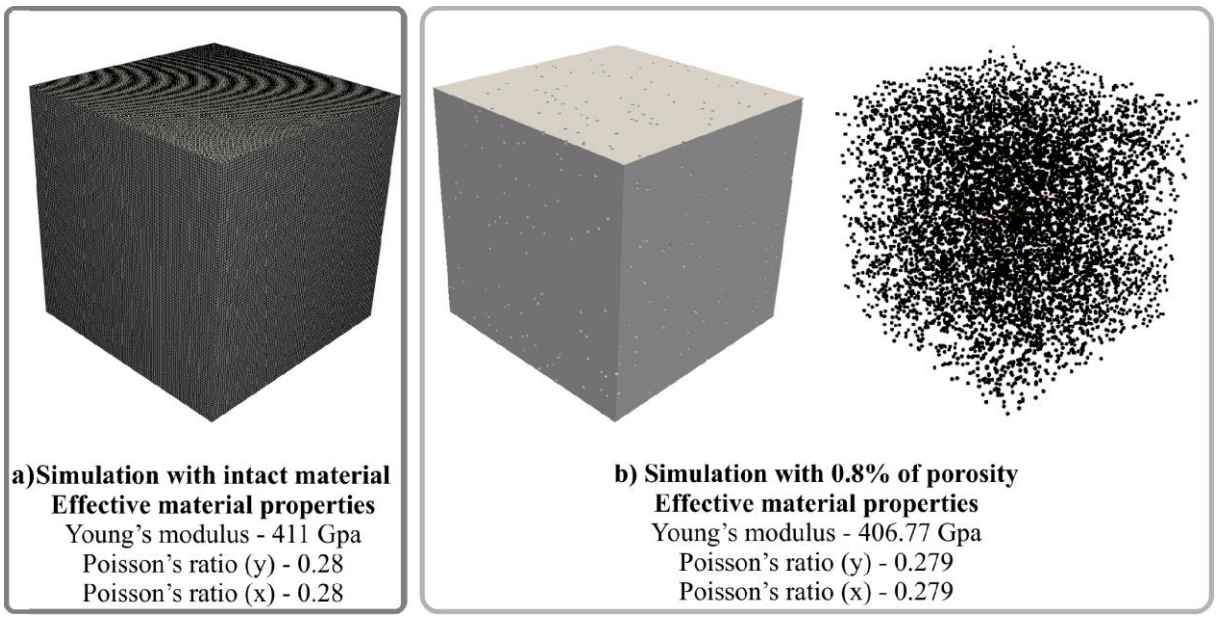


Figure 12. Results a) A deterministic simulation with intact material, b) Intact material and porous structure of the simulation

These tests show how porosity can affect the effective mechanical properties of VPS-W. The effective Young's modulus and Poisson's ratio remain fairly unaffected by this porous volume fraction. However, the extreme thermal loads and neutron dose would cause a more significant change on the mechanical behaviour of VPS-W. Other volume fractions will be used in future research to test different scenarios.

3.5 Conclusions

A random field to simulate the porous structure of VPS-W is presented in this study. A methodology to understand the effective Young's modulus and Poisson's ratio is proposed.

The software ParaFEM will be modified to analyse the thermomechanical response of different volume fractions of porosity in VPS-W using this random field generator.

4 GENERAL CONCLUSIONS

- This paper has found that the SFEM has the potential of simulating spatial material variability for nuclear materials. The simulated material properties for graphite, for example, Young's modulus and strength can be calibrated to experimentally measured values. This finding provides confidence for future work concerning the prediction of graphite behaviour over its lifetime in particular, the effect of microcracking of the bulk material. A criterion for crack initiation has to be developed and further work is required to test the accuracy of this method
- A random field for porous materials has been proposed to analyse the thermomechanical response of large components covered with VPS-W. The SFEM would enable to measure the thermomechanical response of whole parts of a fusion reactor. Moreover, preliminary results to obtain the effective material properties of VPS-W are provided.

REFERENCES

[1] E. Vanmarcke, *Random Fields: Analysis and Synthesis*. World Scientific, 2010

[2] S. Torquato, *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. Springer New York, 2005

[3] J. D. Arregui-Mena, L. Margetts, P. M. Mummery, *The Practical Stochastic Finite Element Method*, *Arch. Comput. Methods Eng.* (2014)

[4] R. A. Knief, *Nuclear Engineering: theory and technology of commercial nuclear power*. Second Edition, 2008

[5] "ITER – the way to new energy" [Online]. Available: http://www.iter.org/sci/whatisfusion. [Accessed 29 August 2013]

[6] D. G. Cacuci, *Handbook of nuclear engineering*. New York, Springer, 2010

[7] A. G. Steer, *AGR core design, operation and safety functions*. Management of ageing in graphite reactor core, 2007

[8] E. Nonbøl, *Description of the Advanced Gas Cooled Type of Reactor (AGR)*. Reactors in Nordic Surroundings. Roskilde, Denmark: Risø National Laboratory

[9] I. M. Smith, D. V. Griffiths, L. Margetts, *Programming the Finite Element Method, Fifth Edition*. Wiley, United Kingdom, 2014

[10] G. A. Fenton, D.V. Griffiths, *Risk Assessment in Geotechnical Engineering.* John Wiley & Sons, 2008

[11] D. K. L. Tsang, B. J. Marsden, The development of a stress analysis code for nuclear graphite components in gas-cooled reactors. *Journal of Nuclear Material*, Vol 350, 2006.

[12] A. Shterenlikht, L. Margetts, Three-dimensional cellular automata modelling of cleavage propagation across crystal boundaries in polycrystalline microstructures. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol 471

[13] M. Haverty, *The microstructure, texture and thermal expansion of nuclear graphite*. Thesis, Manchester, UK, The University of Manchester, 2015

[14] Ll.M. Evans, *Thermal Finite Element Analysis of Ceramic/Metal joining for fusion using X-ray tomography data*. Thesis, UK, The University of Manchester, 2013

[15] International Atomic Agency, *Fusion Physics*. Vienna, Austria, 2012

[16] "The tokamak" [Online], Available: http://www.ccfe.ac.uk/tokamak.aspx, [Accessed 01 March 2016]

[17] [Online]. Available: http://www.jet.efda.org/wp-content/gallery/graphics/jg05-537-1c.jpg. [Accessed 29 August 2013].

[18] "Specific manufacturing : heat removal system with cooling water system," [Online]. Available: http://www.atmostat-alcen.com/en/activities/specific-manufacturing. [Accessed 29 August 2013].

[19] Greuner, H. Bolt, B. Böswirth, S. Lindig, W. Kühnlein, T. Huber, K. Sato, S. Suzuki, Vacuum plasma-sprayed tungsten on EUROFER and 316L: Results of characterisation and thermal loading tests. *Fusion Engineering and Design*, 2005

[20] Exploring the pores in a heavy metal: the 3D microstructure of coatings for nuclear fusion reactors is revealed by high-resolution synchrotron microtomography [Online], http://www.esrf.eu/news/spotlight/spotlight113, [Accessed 01 March 2016]

[21] A. Zivelonghi, Thermomechanical behavior of two heterogeneous tungsten materials via 2D and 3D Image-Based FEM. Thesis, Technischen Universität München, München,Germany, 2010

[22] Ll. M. Evans, L. Margetts, V. Casalegno, L. M. Lever, J. Bushell, T, Lowe, A. Wallwork, P. Young, A. Lindemann, M. Schmidt, P.M. Mummery, Transient thermal finite element analysis of CFC-Cu ITER monoblock using X-ray tomography data, *Fusion Engineering and Design*. Vol 100, 2015

[23] J. Paiboon, D.V. Griffiths, Jinsong Huang, G. A. Fenton, Numerical analysis of effective elastic properties of geomaterials containing voids using 3D random fields and finite elements. *International Journal of Solids and Structures*, Vol 50, 2013

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DETERMINATION OF MESOSCALE RANDOM FIELDS FOR THE APPARENT PROPERTIES OF SPATIALLY RANDOM COMPOSITES

Dimitrios Savvas[1], **George Stefanou**[2]

[1]School of Civil Engineering, National Technical University of Athens
15780 Athens, Greece
e-mail: dimitriosavvas@yahoo.gr

[2] Department of Civil Engineering, Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
e-mail: gstefanou@civil.auth.gr

Keywords: Random Composites, Apparent Properties, Mesoscale random fields, RVE size, XFEM, Monte Carlo Simulation.

Abstract. *The macroscopic mechanical and physical properties of heterogeneous materials can be efficiently determined using either analytical or numerical homogenization techniques where the identification of a representative volume element (RVE) is required over which a fine-scale boundary value problem is solved. In this work, an efficient computational scheme is proposed for the determination of mesoscale random fields for the apparent properties and of the RVE size of particle-reinforced composites based on computer-simulated images of their microstructure. A variable number of microstructure models are directly constructed by segmentation of the composite material image into windows of certain size. The proposed numerical procedure takes into account the particle volume fraction variation through digital image processing of the microstructure models. The proposed approach couples the extended finite element method (XFEM) with Monte Carlo simulation in order to analyze the microstructure models and obtain statistical information (probability distribution, correlation structure) for the apparent properties of the composite in each window size. The XFEM analysis of the microstructure models also leads to upper and lower bounds for their constitutive behavior by solving Dirichlet and Neumann boundary value problems. The RVE is attained within a prescribed tolerance by examining the convergence of these two bounds with respect to the mesoscale size and useful conclusions are derived about the effect of matrix/inclusion stiffness ratio as well as of inclusion volume fraction on the apparent properties and on the RVE size.*

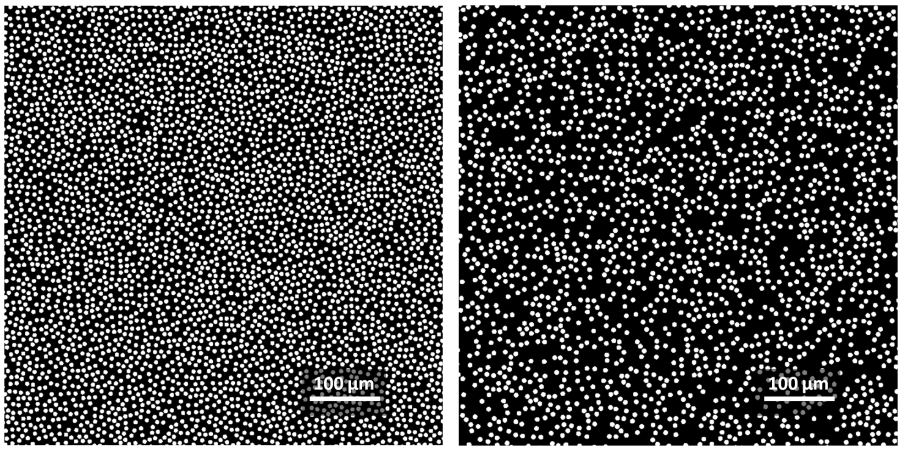
1 INTRODUCTION

The key issue in homogenization methods is the linking of micromechanical characteristics with the random variation of material properties at the macroscale [1, 2, 3, 4]. This link can be established using the Hill-Mandel macro-homogeneity condition [5] among other asymptotic or heterogeneous multiscale methods [6, 7]. In all these methods, the identification of a representative volume element (RVE) of the heterogeneous material is required over which a fine-scale boundary value problem is solved. Then effective material properties can be calculated using the link between the fine and coarse scale equations.

While the concept of unit cell is valid whenever periodicity of microstructure is present, in the case of spatial randomness (e.g. random position, shape, size of inclusions), the identification of RVE must be based on computational convergence schemes with respect to specific apparent properties. The typical procedure consists in setting multiple realizations of the microstructure followed by finite element simulation and statistical analysis of the results [8, 9, 10, 11, 12, 13, 14]. In the context of stochastic finite element analysis of composite structures, the RVE size determines the minimum element size that should be used for discretization at the macroscale [15, 16, 17].

Since the volume fraction is one of the primary microstructural features in homogenization problems, the constant volume fraction assumption can lead to inexact effective properties. In this work, a novel computational scheme is proposed for the determination of RVE size for particle-reinforced composites based on computer-simulated images of their microstructure. A variable number of microstructure models are directly constructed from the composite material image using the moving window technique. In this way, the scatter of inclusions for each sample model is not achieved in an arbitrary manner but is an exact geometrical representation of the corresponding window segment of the composite. Moreover, the volume fraction of inclusions is computed through digital image processing of the microstructure models. The proposed numerical procedure takes into account the local volume fraction variation and thus the obtained RVE can lead to more accurate homogenized properties.

This paper couples the extended finite element method (XFEM) with Monte Carlo simulation (MCS) in order to analyze the microstructure models and obtain statistical information for the constitutive properties of the samples in each window size. The number of MCS depends on the moving window technique and is increasing for small window sizes. The implementation of XFEM is particularly suitable for this type of problems since there is no need to generate conforming finite element (FE) meshes at each MCS [18, 19, 20]. The weak discontinuity in the solution due to material interfaces is captured using nodal enrichment functions within the framework of the partition of unity method to augment the FE approximations over a structured mesh [21]. The XFEM analysis of the microstructure models is then used to compute upper and lower bounds for their constitutive behavior by solving multiple boundary value problems under kinematic and static uniform boundary conditions [22]. The RVE is attained within a prescribed tolerance by examining the convergence of these two bounds with respect to the mesoscale size. Mesoscale random fields describing the spatial variation of the components of the apparent elasticity tensor are also determined.



(a) (b)

Figure 1: Computer-simulated images of composite materials containing circular inclusions with a) vf=40% (image 1) and b) vf=20% (image 2).

2 COMPUTATION OF LOCAL VOLUME FRACTION VARIATION

The variability of local volume fraction (vf) in computer-simulated composites is studied in this section exploiting image analysis techniques available in MATLAB software. Specifically, images of two composites containing an initial amount (vf) of 40% and 20% of inclusions, respectively, are processed (see Fig. 1). Both images have dimensions $L_{image} \times L_{image}$ ($L_{image} = 640\mu m$) and an approximate analysis of 5800×5800 pixels. The represented composites contain 4096 and 2048 non-overlapping randomly scattered inclusions, respectively. Each inclusion has circular shape with constant area 40 μm^2 which is equivalent to a constant diameter $d \simeq 7.14\ \mu m$.

The computation of the local volume fraction variation is based on the moving window technique. In this method, an initial window of area $L \times L$ is set at a starting point O of the image and then, by moving this window over the image by a vector $\vec{\xi_p} = \xi_{x_p}\vec{e_x} + \xi_{y_p}\vec{e_y}$ a set of windows of the same size can be extracted (see Fig. 2). In this paper, the moving distance step $\Delta\xi$ is assumed to be the same along both directions. This means that $\xi_{x_p} = i\Delta\xi$ and $\xi_{y_p} = j\Delta\xi$, with i, j the number of steps along the x and y directions, respectively. The total number of windows n_w belonging to each set depends on the image size L_{image}, the investigated window size L and the selected moving distance step $\Delta\xi$ as follows:

$$n_w = \left(\frac{L_{image} - L}{\Delta\xi} + 1 \right)^2 \tag{1}$$

By choosing $\Delta\xi = L$ with L an integer divisor of L_{image} the whole image is segmented into $n_w = (L_{image}/L)^2$ non-overlapping windows (mesoscale models). Specifically 10 sets containing n_w=4, 16, 25, 64, 100, 256, 400, 1024, 1600, 4096 windows have been generated by choosing $\Delta\xi = L$= 320, 160, 128, 80, 64, 40, 32, 20, 16, 10 μm, respectively.

The vf of inclusions is computed in each window by using image analysis tools. The histograms of the vf corresponding to each set of windows are shown in Fig. 3 for the two images of Fig. 1. As the dispersion of the inclusions within the composite materials is not uniform, it is obvious that there are regions rich or poor in inclusions. This is clearly illustrated in the histograms where a very large variability of vf is observed for the sets containing windows of small size. Thus in the context of homogenization methods, choosing a RVE size for composite ma-

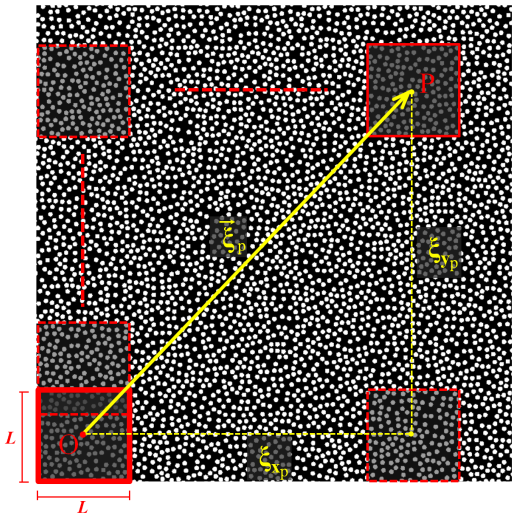


Figure 2: Illustration of the moving window technique.

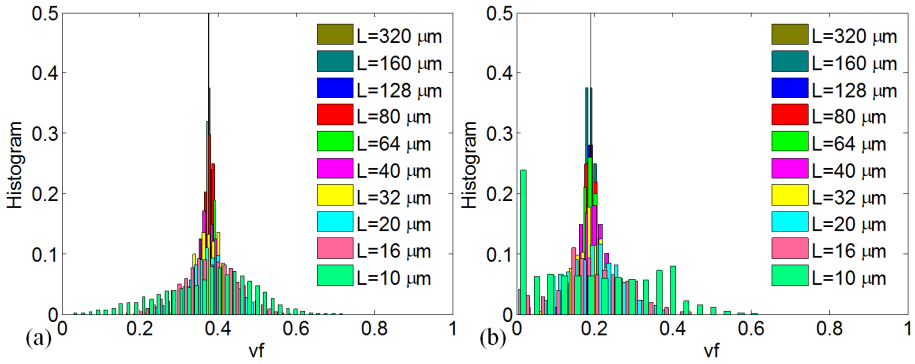


Figure 3: Histograms of vf for various window sizes L for a) image 1, b) image 2.

terials with spatial randomness without considering local vf variability, can lead to unrealistic estimations of their mechanical behavior [23].

3 COMPUTATIONAL PROCEDURE FOR RVE DETERMINATION

According to [24] the homogenization process is based on the fundamental assumption of statistical homogeneity of the heterogeneous medium. This means that all statistical properties of the state variables are the same at any material point and thus a RVE can be identified. While in case of composites with periodic or nearly periodic geometry the RVE is explicitly defined, in case of spatial randomness, the existence of RVE has to be sought using computational methods. In this paper, identification of RVE is performed using a computational procedure based on Hill's definition [5] which postulates separation of scales in the form:

$$d \ll L \ll L_{macro} \tag{2}$$

In the above inequality, the microscale parameter d denotes a characteristic size of the fillers, e.g. their diameter in case of circular inclusions, the mesoscale parameter L denotes the size of the RVE and the macroscale parameter L_{macro} denotes the characteristic length over which the

macroscopic loading varies in space, or in the case of complete scale separation, the size of the macroscopic homogeneous medium.

In the proposed computational procedure, the RVE will be defined with respect to the non-dimensional parameter $\delta = L/d$ with $\delta \in [1, \infty]$. Note that in the absence of spatial periodicity in composite materials the RVE is exactly obtained only in the limit $\delta \to \infty$. However, the RVE can be attained at a finite mesoscale size δ when a particular apparent property derived from statistical volume elements (SVEs) is almost approaching to a constant effective property. At this mesoscale size, the estimation of the particular effective property is not changing with increasing the number of realizations.

In this paper, the RVE is identified in terms of the convergence of the components of the spatially averaged apparent elasticity tensor, which are calculated on mesoscale models of increasing size by applying both kinematic and static uniform boundary conditions. In each investigated mesoscale size δ, variability of vf is taken into account in the computational procedure by processing microstructure models directly extracted from images of the particular composite material (section 2). The RVE obtained from this computational approach is expected to lead to more realistic homogenization results. In the numerical examples of section 4, it will be shown that the convergence rate of the mesoscale models to the RVE is mainly depending on the matrix/inclusion stiffness ratio and volume fraction of the composite material.

3.1 Problem formulation

Let us denote a mesoscale realization of a composite material as $B_\delta(\theta)$ and its boundary surface as $\partial B_\delta(\theta)$, $\theta \in \Theta$ where Θ is the sample space (see Fig. 4). The governing equilibrium equation for the elastostatic problem of the medium is

$$\text{div}\boldsymbol{\sigma}(\theta, \boldsymbol{Y}) + \boldsymbol{b} = 0 \ \text{ in } B_\delta(\theta) \tag{3}$$

where $\boldsymbol{b}$ are the body forces acting on the medium and $\boldsymbol{Y}$ denotes the coordinate system on the mesoscale model $B_\delta(\theta)$ with the stress and strain fields connected by Hooke's elasticity tensor $C(\theta, \boldsymbol{Y})$

$$\boldsymbol{\sigma}(\theta, \boldsymbol{Y}) = C(\theta, \boldsymbol{Y}) : \boldsymbol{\varepsilon}(\theta, \boldsymbol{Y}) \tag{4}$$

These fields can be expressed as a superposition of means ($\bar{\boldsymbol{\sigma}}$ and $\bar{\boldsymbol{\varepsilon}}$) and of zero-mean fluctuations ($\boldsymbol{\sigma'}$ and $\boldsymbol{\varepsilon'}$) as follows:

$$\boldsymbol{\sigma}(\theta, \boldsymbol{Y}) = \bar{\boldsymbol{\sigma}} + \boldsymbol{\sigma'}(\theta, \boldsymbol{Y}) \quad , \quad \boldsymbol{\varepsilon}(\theta, \boldsymbol{Y}) = \bar{\boldsymbol{\varepsilon}} + \boldsymbol{\varepsilon'}(\theta, \boldsymbol{Y}) \tag{5}$$

The means of stress and strain tensors at some point $\boldsymbol{X}$ of the macro-continuum are computed as volume averages over $B_\delta(\theta)$ in the form [5]:

$$\bar{\boldsymbol{\sigma}}(\boldsymbol{X}) = \frac{1}{V_\delta} \int\limits_{B_\delta(\theta)} \boldsymbol{\sigma}(\theta, \boldsymbol{Y}) \, \mathrm{d}V_\delta, \quad \bar{\boldsymbol{\varepsilon}}(\boldsymbol{X}) = \frac{1}{V_\delta} \int\limits_{B_\delta(\theta)} \boldsymbol{\varepsilon}(\theta, \boldsymbol{Y}) \, \mathrm{d}V_\delta \tag{6}$$

Also the volume average of the strain energy can be calculated as:

$$\bar{U} = \frac{1}{2V_\delta} \int\limits_{B_\delta(\theta)} \boldsymbol{\sigma}(\theta, \boldsymbol{Y}) : \boldsymbol{\varepsilon}(\theta, \boldsymbol{Y}) \, \mathrm{d}V_\delta = \frac{1}{2}\overline{\boldsymbol{\sigma} : \boldsymbol{\varepsilon}} = \frac{1}{2}\bar{\boldsymbol{\sigma}} : \bar{\boldsymbol{\varepsilon}} + \frac{1}{2}\overline{\boldsymbol{\sigma'} : \boldsymbol{\varepsilon'}} \tag{7}$$

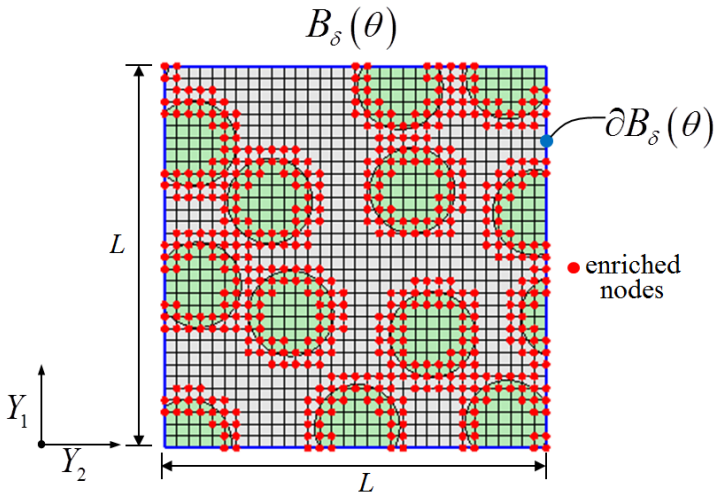


Figure 4: Discretization of a mesoscale model of a spatially random composite material.

Note that for an unbounded space domain $(\delta \to \infty)$ the fluctuation terms in (7) become negligible $\left(\overline{\boldsymbol{\sigma}' : \boldsymbol{\varepsilon}'} = 0\right)$ and thus the following equation holds:

$$\overline{\boldsymbol{\sigma} : \boldsymbol{\varepsilon}} = \bar{\boldsymbol{\sigma}} : \bar{\boldsymbol{\varepsilon}} \tag{8}$$

which is known as Hill's condition. However, at a finite mesoscale, Hill's condition is valid provided that the following constraint is satisfied [25]:

$$\int_{\partial B_\delta} (t - \bar{\boldsymbol{\sigma}} \cdot n) \cdot (u - \bar{\boldsymbol{\varepsilon}} \cdot \boldsymbol{Y}) \, \mathrm{d}S = 0 \tag{9}$$

The constraint equation (9) is a priori satisfied by the following types of boundary conditions:

1. Uniform strains (kinematic, essential or Dirichlet):

$$u(\boldsymbol{Y}) = \bar{\boldsymbol{\varepsilon}} \cdot \boldsymbol{Y} \ , \ \forall \, \boldsymbol{Y} \in \partial B_\delta \tag{10}$$

2. Uniform stresses (static, natural or Neumann):

$$t(\boldsymbol{Y}) = \bar{\boldsymbol{\sigma}} \cdot n \ , \ \forall \, \boldsymbol{Y} \in \partial B_\delta \tag{11}$$

3. Uniform orthogonal-mixed:

$$(t(\boldsymbol{Y}) - \bar{\boldsymbol{\sigma}} \cdot n) \cdot (u(\boldsymbol{Y}) - \bar{\boldsymbol{\varepsilon}} \cdot \boldsymbol{Y}) \ , \ \forall \, \boldsymbol{Y} \in \partial B_\delta \tag{12}$$

4. Periodic:

$$\begin{gathered} u\left(\boldsymbol{Y}^+\right) - u\left(\boldsymbol{Y}^-\right) = \bar{\boldsymbol{\varepsilon}} \cdot \left(\boldsymbol{Y}^+ - \boldsymbol{Y}^-\right) \text{ and } t\left(\boldsymbol{Y}^+\right) + t\left(\boldsymbol{Y}^-\right) = 0 \ , \\ \forall \, \boldsymbol{Y}^+ \in \partial B_\delta^+ \text{ and } \forall \, \boldsymbol{Y}^- \in \partial B_\delta^- \text{ with } \partial B_\delta = \partial B_\delta^+ \cup \partial B_\delta^- \end{gathered} \tag{13}$$

3.2 Computation of apparent properties on mesoscale

Miehe and Koch [22] proposed a computational procedure to define apparent properties (homogenized stresses and overall tangent moduli) of microstructures undergoing small strains. They have shown that apparent properties can be defined in terms of discrete forces and stiffness properties on the boundary of discretized microstructures. Using these deformation-driven algorithms, the apparent stiffness or compliance tensor of a mesoscale model of size δ can be calculated by solving a uniform strain or a uniform stress boundary value problem, respectively. The extended finite element method (XFEM) is used to model the microstructural problems. In the context of XFEM, weak discontinuities (material interfaces) can be captured by a discontinuous approximation of the displacement function $u^h(\boldsymbol{Y})$ as follows [26]:

$$u^h(\boldsymbol{Y}) = \sum_{i \in I} N_i(\boldsymbol{Y})\, u_i + \sum_{j \in J} N_j(\boldsymbol{Y}) \left(\sum_{k=1}^{n_0} \psi_k(\boldsymbol{Y})\, \alpha_{jk} \right) \tag{14}$$

where N are the shape functions of the elements used, I is the set of all nodes in the mesh, J is the set of nodes (denoted as red circles in Fig. 4) that are enriched with the enrichment functions ψ_k, u_i are the classical dofs, α_{jk} are the enriched dofs corresponding to node j whose support is cut by the kth inclusion. To improve the accuracy and convergence of XFEM solution the following enrichment function is used [18]:

$$\psi_k(\boldsymbol{Y}) = \sum_{i \in I} N_i(\boldsymbol{Y}) \left| \phi_i^k \right| - \left| \sum_{i \in I} N_i(\boldsymbol{Y})\, \phi_i^k \right| \tag{15}$$

with ϕ_i^k the nodal values of the level set function corresponding to the kth circular inclusion. The adopted computational procedure is outlined below for the two cases of uniform boundary conditions. More details can be found in [27].

3.2.1 Uniform strains

A prescribed uniform strain tensor $\bar{\boldsymbol{\varepsilon}} = [\bar{\varepsilon}_{11} \;\; \bar{\varepsilon}_{22} \;\; 2\bar{\varepsilon}_{12}]^T$ is applied on the boundary ∂B_δ of a discretized mesoscale model, as that of Fig. 4 through displacement boundary conditions (Dirichlet) in the form:

$$u_b = D_b^T \bar{\boldsymbol{\varepsilon}} \tag{16}$$

where D_b is a geometric matrix which depends on the coordinates of the boundary node b and is defined as:

$$D_b = \frac{1}{2} \begin{bmatrix} 2Y_1 & 0 \\ 0 & 2Y_2 \\ Y_2 & Y_1 \end{bmatrix} \quad \text{with} \quad (Y_1, Y_2) \in \boldsymbol{Y} \tag{17}$$

The static problem is denoted by:

$$\begin{bmatrix} K_{ii} & K_{ib} \\ K_{bi} & K_{bb} \end{bmatrix} \begin{bmatrix} U_i \\ U_b \end{bmatrix} = \begin{bmatrix} F_i \\ F_b \end{bmatrix} \tag{18}$$

where the subscripts i and b denote internal and boundary nodes, respectively. Then the apparent stiffness tensor $C_\delta^D(\theta)$ of the mesoscale model of size δ under Dirichlet boundary conditions

can be calculated in terms of the condensed stiffness matrix $\tilde{K}_{bb} = K_{bb} - K_{bi}K_{ii}^{-1}K_{ib}$ in the form:

$$C_{\delta}^{D}(\theta) = \frac{1}{V_{\delta}}\mathbb{D}\tilde{K}_{bb}\mathbb{D}^{T} \tag{19}$$

with $\mathbb{D} = [D_1\ D_2\ ...\ D_M]$, M the total number of boundary nodes.

3.2.2 Uniform stresses

A prescribed uniform stress tensor $\bar{\boldsymbol{\sigma}} = [\bar{\sigma}_{11}\ \ \bar{\sigma}_{22}\ \ 2\bar{\sigma}_{12}]^{T}$ is applied on the boundary surface ∂B_{δ} of a discretized mesoscale model, as that of Fig. 4 through traction boundary conditions (Neumann) as follows:

$$F_b = S_b^{T}\bar{\boldsymbol{\sigma}} \tag{20}$$

where S_b is a matrix depending on the components of the discrete area vector $\mathbf{a}_b$ which is given in terms of the nodal coordinates of the neighboring boundary nodes $b-1$, b and $b+1$ in the form:

$$\mathbf{a}_b = \frac{1}{2}\left[\boldsymbol{Y_{b+1}} - \boldsymbol{Y_{b-1}}\right] \times \boldsymbol{e_n} \tag{21}$$

These nodes are oriented so that the cross product with the Cartesian, out-of-plane, base vector $\boldsymbol{e_n}$ yields $\mathbf{a}_b$ as an outward normal vector at the boundary node b. Thus the matrix S_b is defined as:

$$S_b = \begin{bmatrix} \mathrm{a}_{b_1} & 0 \\ 0 & \mathrm{a}_{b_2} \\ \mathrm{a}_{b_2} & \mathrm{a}_{b_1} \end{bmatrix} \tag{22}$$

Then the apparent compliance tensor $S_{\delta}^{N}(\theta)$ of the mesoscale model of size δ under Neumann boundary conditions can be calculated in terms of the condensed stiffness matrix $\tilde{K}_{bb}$ in the form:

$$S_{\delta}^{N}(\theta) = \frac{1}{V_{\delta}}\mathbb{S}\tilde{K}_{bb}^{-1}\mathbb{S}^{T} \tag{23}$$

where $\mathbb{S} = [S_1\ S_2\ ...\ S_M]$ with M the total number of boundary nodes. For this type of boundary conditions the apparent stiffness tensor is obtained by inverting the compliance tensor of Eq. (23):

$$C_{\delta}^{N}(\theta) = \left[S_{\delta}^{N}(\theta)\right]^{-1} \tag{24}$$

Dirichlet and Neumann boundary conditions provide upper and lower bounds of the strain energy which converge to each other as the mesoscale size δ is increasing. Thus the following relation holds [28]:

$$\left\{\begin{array}{ll} \frac{1}{2}\left[\bar{\boldsymbol{\varepsilon}} : C_{\delta}^{N}(\theta) : \bar{\boldsymbol{\varepsilon}}\right] < \frac{1}{2}\left[\bar{\boldsymbol{\varepsilon}} : C_{\delta}^{D}(\theta) : \bar{\boldsymbol{\varepsilon}}\right] & \text{for } \delta \text{ finite} \\ \frac{1}{2}\left[\bar{\boldsymbol{\varepsilon}} : C_{\infty}^{N}(\theta) : \bar{\boldsymbol{\varepsilon}}\right] = \frac{1}{2}\left[\bar{\boldsymbol{\varepsilon}} : C_{\infty}^{D}(\theta) : \bar{\boldsymbol{\varepsilon}}\right] & \text{for } \delta \to \infty \end{array}\right\} \tag{25}$$

In case of uniaxial strains $\bar{\boldsymbol{\varepsilon}} = \left([1 \;\; 0 \;\; 0]^T \text{ or } [0 \;\; 1 \;\; 0]^T \right)$ or simple shear $\bar{\boldsymbol{\varepsilon}} = [0 \;\; 0 \;\; 1]^T$ the following notation can be used:

$$C_\delta^N(\theta) \leq C_\delta^D(\theta) \tag{26}$$

4 RESULTS AND DISCUSSION

In this section, the spatial average (mean) of the apparent moduli is calculated at each size δ from n_w mesoscale models extracted by using the window technique on the composite images of section 2. Thus the following formula for the mean is used:

$$\overline{C_\delta(\theta)} = \frac{1}{n_w} \sum_{p=1}^{n_w} C_\delta\left(\vec{\xi}_p, \theta\right) \tag{27}$$

with $\vec{\xi}_p$ the position vector of the p-th mesoscale window model on the image (see Fig. 2). Following the notation of Eq. (26) the bounds $\overline{C_\delta^N(\theta)}$ and $\overline{C_\delta^D(\theta)}$ are calculated for the axial and shear components of the apparent elasticity tensor. Note that, while the constituent materials of the composite are considered isotropic, the computational method of section 3.2 results in an anisotropic apparent elasticity tensor due to spatial randomness. Composites containing either stiff or compliant inclusions are investigated which are obtained by keeping the Young's modulus of the matrix constant at $E_m = 1$ GPa and varying the Young's modulus of the inclusions E_{in}. The Poisson ratio is assumed to be the same for the two phases, $\nu_m = \nu_{in} = 0.3$. All components of the apparent elasticity tensor computed in this section are in GPa.

4.1 Effect of volume fraction and stiffness ratio on RVE size

The convergence of $\overline{C_\delta^N(\theta)}$ and $\overline{C_\delta^D(\theta)}$ with respect to the non-dimensional parameter δ is illustrated in Fig. 5 for the composite materials of images 1 and 2 (see Fig. 1). The results refer to the average axial stiffness $C_{ii}/2$ ($i = 1, 2$) and shear stiffness $C_{33} \equiv G$ components of the mean elasticity tensor. As mentioned before in section 2, image 1 corresponds to a composite with initial vf=40%, whereas image 2 corresponds to a composite with initial vf=20%. Note that both composites contain circular inclusions of constant size ($d \simeq 7.14\ \mu m$). The effect of volume fraction on the mechanical properties of the composite is also illustrated in Fig. 5. It can be observed that the convergence rate of the two numerical bounds with respect to δ is slower as vf increases. This means that, in order to perform homogenization in composites with high volume fraction, a much larger RVE has to be used.

The results of a parametric study with respect to matrix/inclusion stiffness ratio are shown in Figs. 6 and 7 for the composite of image 1. The cases of stiff (E_{in}/E_m=10, 100, 1000) and compliant (E_{in}/E_m=0.1, 0.01, 0.001) inclusions have been tested. As the Young's modulus of the inclusions increases, the computed apparent properties of the composite increase as well. The effect of stiffness ratio on the convergence rate of the apparent properties is also illustrated in these figures. This effect is better illustrated in Fig. 8 where the discrepancy (tolerance) e_δ between the numerical bounds is plotted against the non-dimensional parameter δ for both composites examined (images 1 and 2). The discrepancy is calculated as:

$$e_\delta = \left| \frac{\overline{C_{ij}^D} - \overline{C_{ij}^N}}{\overline{C_{ij}^D}} \right|_\delta \tag{28}$$

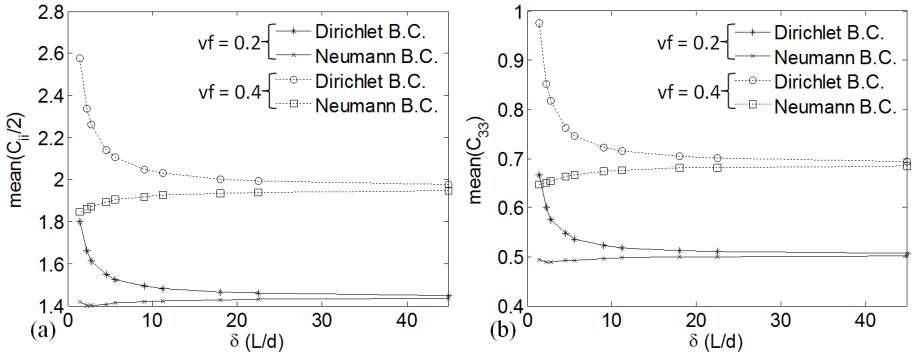


Figure 5: Comparison of numerical bounds of apparent properties for the composites of images 1 and 2 containing stiff inclusions ($E_{in}/E_m = 10$).

where $\overline{C_{ij}^D}$ and $\overline{C_{ij}^N}$ are the components of the spatially averaged apparent elasticity tensor obtained by applying Dirichlet and Neumann boundary conditions, respectively.

4.2 Determination of mesoscale random fields

In this section, random fields for the volume fraction and the apparent elasticity tensor of the composite of image 1 (see section 2) are obtained in mesoscale sizes δ from the simulation of a large number of SVE models extracted using the moving window technique. In order to obtain accurate statistical properties, the moving distance step of the method has been chosen as $\Delta\xi = 0.25L$, with L the moving window size. Specifically, random fields for δ =5.605, 11.210 and 22.420 have been derived by simulating n_w =3721, 841 and 169 SVEs, respectively (see Eq. (1) in section 2 for n_w). All the presented results refer to kinematic uniform boundary conditions.

Figs. 9 and 10 depict the computed random fields along with the respective empirical distributions and 2-D spatial correlations of the average axial and shear components of the apparent elasticity tensor, for the three mesoscale sizes δ examined. Note that the spatial correlations ρ_A^B have been calculated for every lag $(\xi_x,\ \xi_y)$ according to the following formula:

$$\rho_A^B\left(\xi_x,\ \xi_y\right) = \frac{1}{n_w - 1}\sum_{i=0}^{\sqrt{n_w}}\sum_{j=0}^{\sqrt{n_w}}\left(\frac{A\left(x_i,\ y_j\right) - \bar{A}}{\sigma_A}\right)\left(\frac{B\left(x_i + \xi_x,\ y_j + \xi_y\right) - \bar{B}}{\sigma_B}\right) \tag{29}$$

$$-\sqrt{n_w}\Delta\xi \leq \xi_x \leq \sqrt{n_w}\Delta\xi\ ,\ -\sqrt{n_w}\Delta\xi \leq \xi_y \leq \sqrt{n_w}\Delta\xi$$

with ρ_A^B denoting auto-correlations when quantity $A \equiv B$, otherwise cross-correlations are defined. $\bar{A}$, $\bar{B}$ are the spatial average values while σ_A, σ_B are the standard deviations of quantities A, B, respectively. The mesoscale random fields of the components of the apparent elasticity tensor in Figs. 9 and 10 have been derived for stiffness ratio $E_{in}/E_m = 10$. A general observation for all mesoscale random fields is that their empirical PDFs become narrower as the mesoscale size δ increases. In other words, the random field tends to a random variable and thus the SVE tends to the RVE as δ increases. The auto-correlations for lag $(\xi_x = 0,\ \xi_y = 0)$ are 1 and tend to zero for lag values $|\xi_x| > L$ and $|\xi_y| > L$.

The 1-D cross correlations $\rho_{\nu f}^{0.5(C_{11}+C_{22})}$ and $\rho_{\nu f}^{C_{33}}$ are depicted in Figs. 11 and 12, respectively, for lag values along x $(\xi_x,\ \xi_y = 0)$ and y $(\xi_x = 0,\ \xi_y)$ for stiffness ratio E_{in}/E_m =10 and 1000.

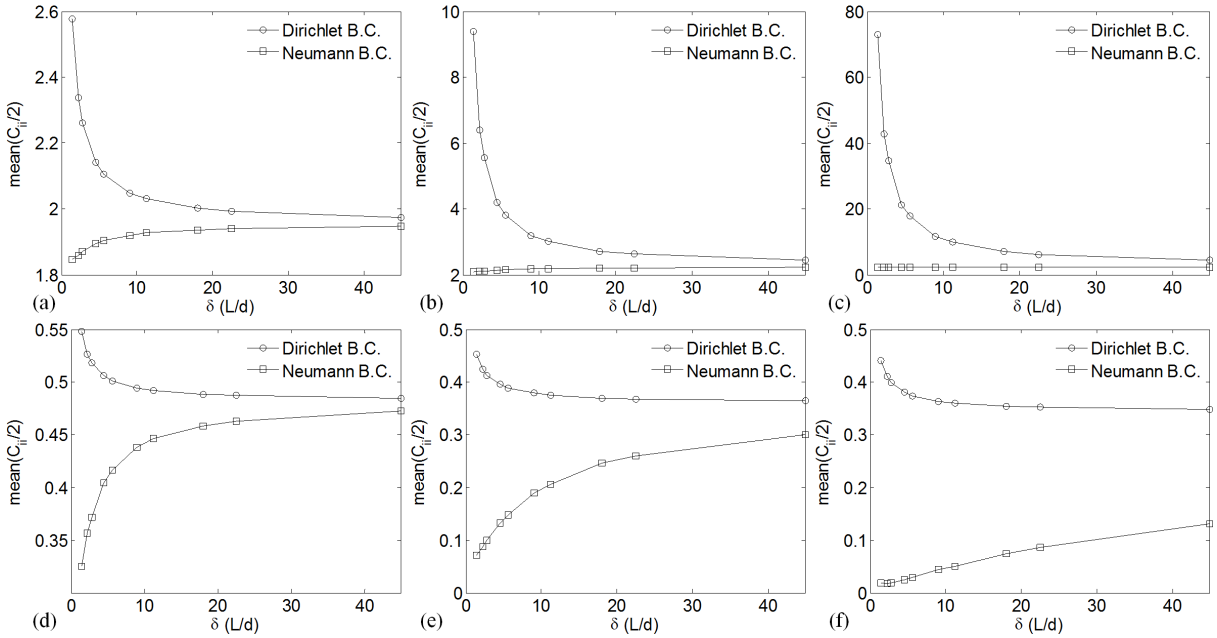


Figure 6: Convergence of numerical bounds on the average axial stiffness ($C_{ii}/2$) with respect to mesoscale size δ for stiffness ratio a) $E_{in}/E_m = 10$, b) $E_{in}/E_m = 100$, c) $E_{in}/E_m = 1000$, d) $E_{in}/E_m = 0.1$, e) $E_{in}/E_m = 0.01$, f) $E_{in}/E_m = 0.001$.

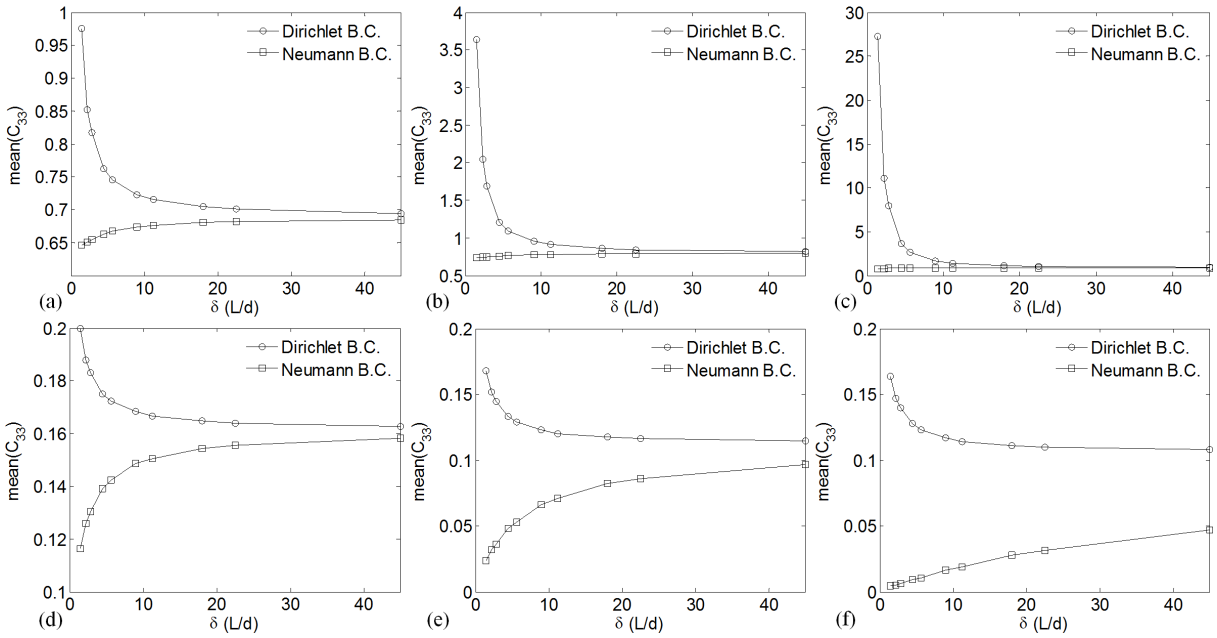


Figure 7: Convergence of numerical bounds on the shear stiffness (C_{33}) with respect to mesoscale size δ for stiffness ratio a) $E_{in}/E_m = 10$, b) $E_{in}/E_m = 100$, c) $E_{in}/E_m = 1000$, d) $E_{in}/E_m = 0.1$, e) $E_{in}/E_m = 0.01$, f) $E_{in}/E_m = 0.001$.

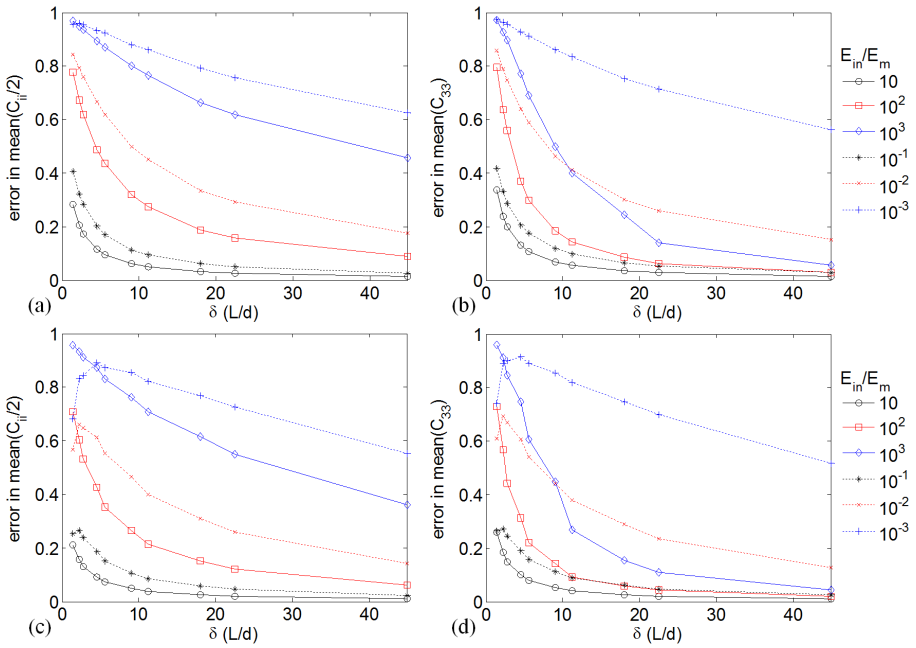


Figure 8: Discrepancy between upper and lower bounds with respect to mesoscale size δ for the composites of image 1 (a-b) and image 2 (c-d) and various cases of stiffness ratio.

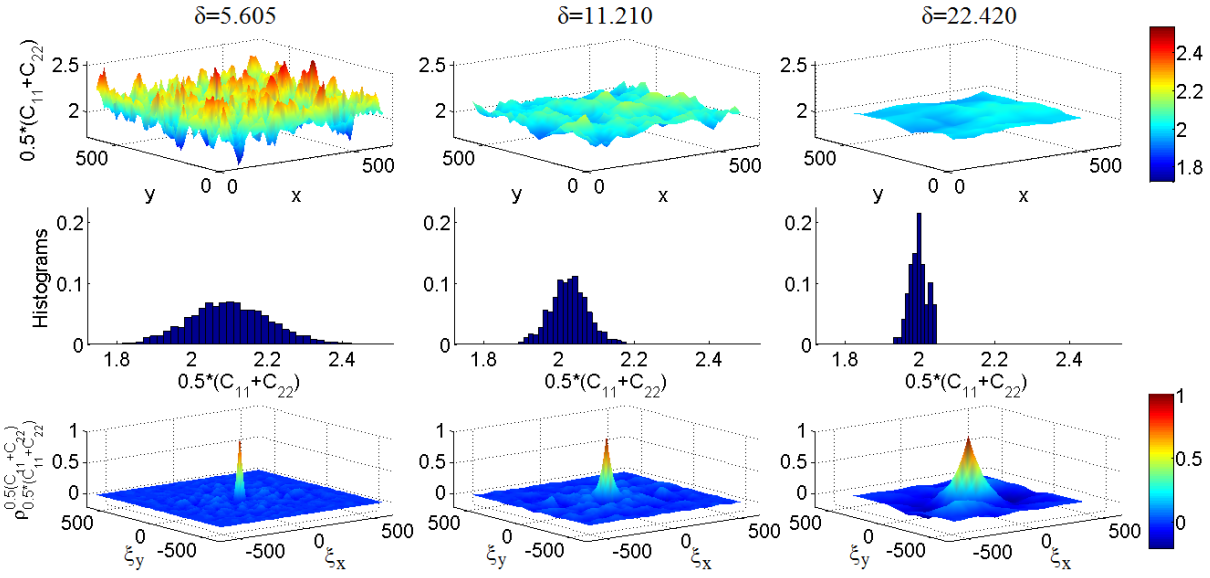


Figure 9: Mesoscale random fields for average axial stiffness ($E_{in}/E_m = 10$).

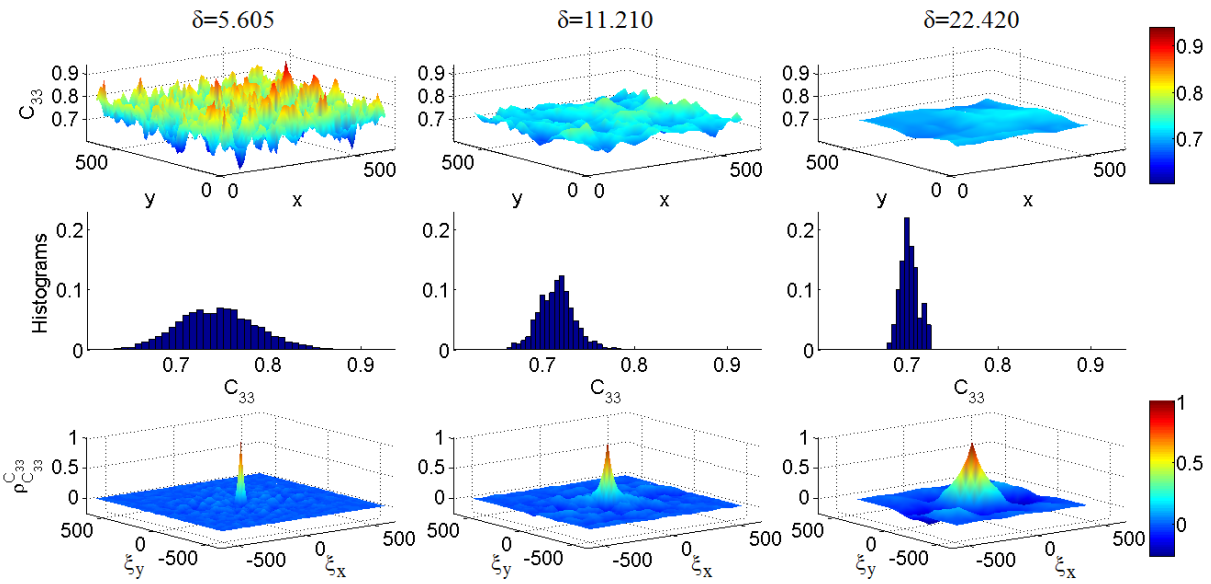


Figure 10: Mesoscale random fields for shear stiffness ($E_{in}/E_m = 10$).

All cross-correlations are between -1 and +1 and tend to zero for $|\xi_x| > L$ or $|\xi_y| > L$. This can be attributed to the fact that the probability of SVE models sharing common inclusions decreases as the length of the vector $\vec{\xi}_p$ (see Fig. 2) increases. A decrease on the values of cross-correlations with the increase of stiffness ratio is also observed, as in [29], especially for $|\xi_x| \leq L$ and $|\xi_y| \leq L$ for all sizes δ. This could be an explanation of the fact that the RVE size can not be defined within a reasonable tolerance for the case of high stiffness ratio (see results of section 4.1). More information about the statistical characteristics of mesoscale random fields is provided in [27].

5 CONCLUSIONS

In this paper, a novel computational procedure based on XFEM and MCS has been proposed for the determination of mesoscale random fields for the apparent properties and of the RVE size of spatially random composites. The proposed approach takes into account the local volume fraction variation by processing computer-simulated images of composites with randomly dispersed inclusions. A variable number of microstructure models were derived directly from the images using the moving window technique. The XFEM analysis of the microstructure models was used to compute upper and lower bounds on the apparent material properties by solving multiple boundary value problems under kinematic and static uniform boundary conditions. The RVE was attained within a prescribed tolerance by examining the convergence of these two bounds with respect to the mesoscale size. The effect of matrix/inclusion stiffness ratio as well as of the volume fraction of inclusions on the RVE size was highlighted. The proposed computational procedure can provide with a more realistic RVE because it incorporates the local volume fraction variation present in real composites and can be extended to any filler-reinforced composite as it is based on image analysis of computer-simulated microstructures. Finally, the obtained results can be used in the framework of homogenization methods and stochastic finite element analysis of composite structures.

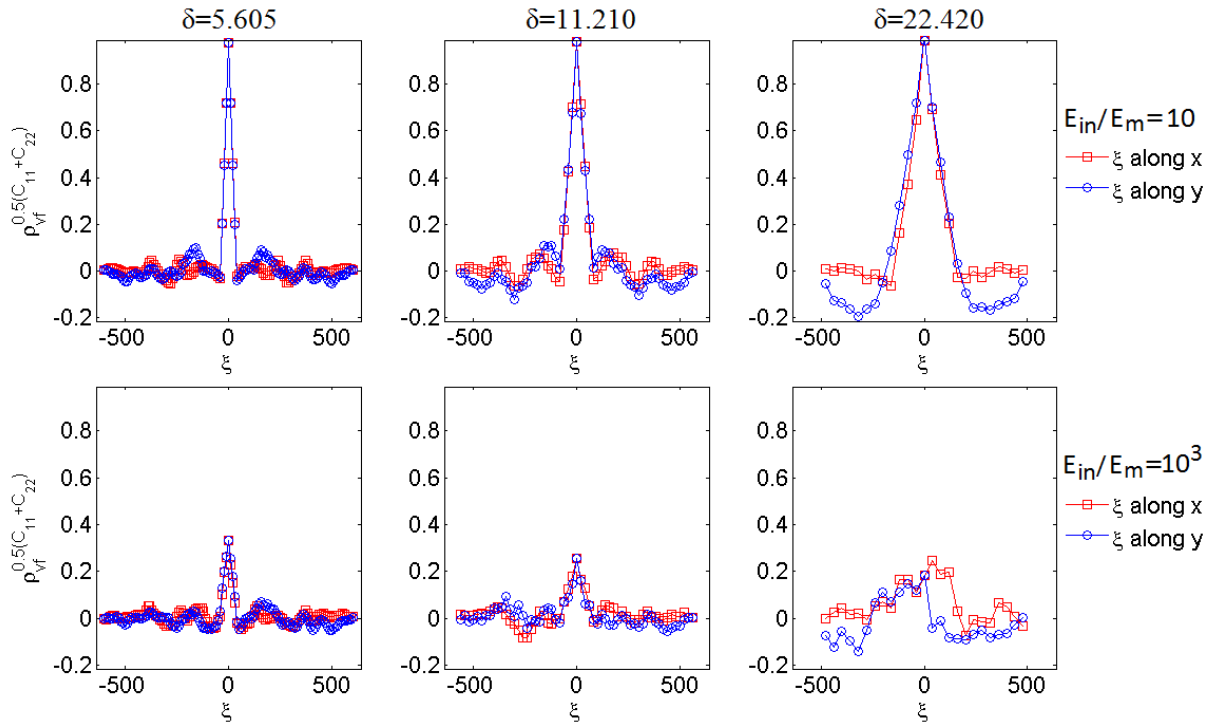


Figure 11: Cross correlations at $(\xi_x,\ \xi_y = 0)$ and $(\xi_x = 0,\ \xi_y)$ for $E_{in}/E_m = 10$ and $E_{in}/E_m = 1000$.

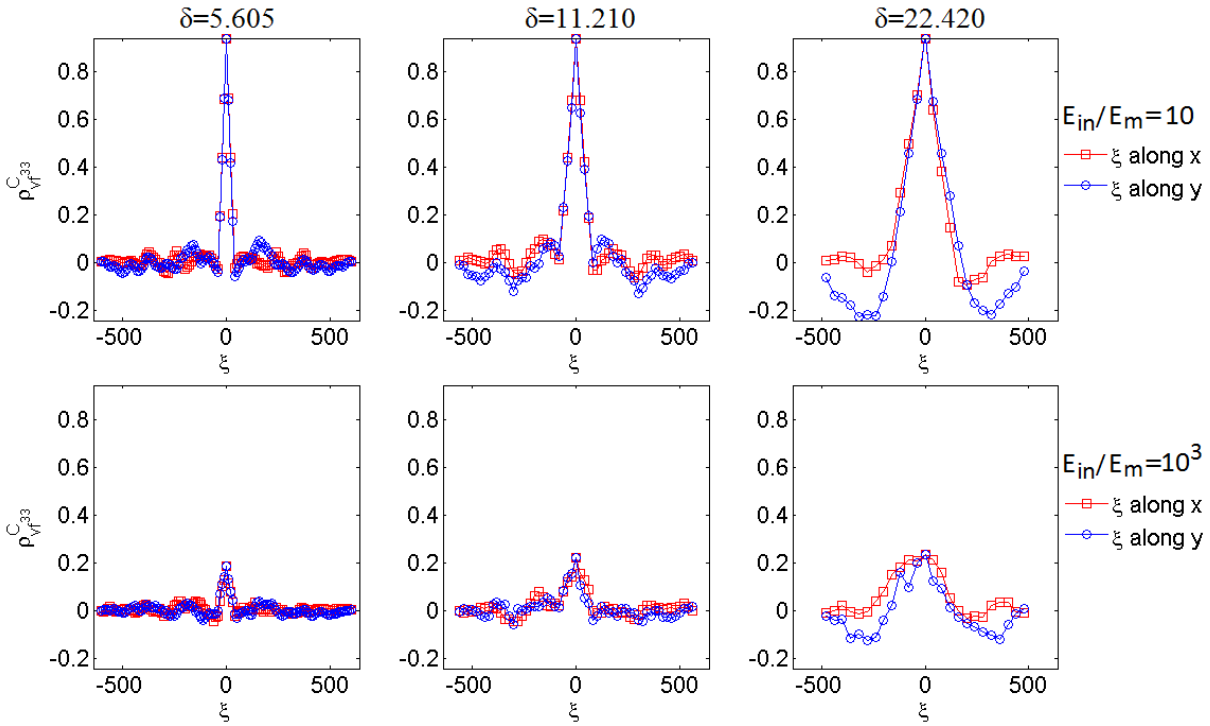


Figure 12: Cross correlations at $(\xi_x,\ \xi_y = 0)$ and $(\xi_x = 0,\ \xi_y)$ for $E_{in}/E_m = 10$ and $E_{in}/E_m = 1000$.

ACKNOWLEDGEMENTS

The financial support provided by the European Research Council Advanced Grant "MASTER - Mastering the computational challenges in numerical modeling and optimum design of CNT reinforced composites" (ERC-2011-ADG 20110209) is gratefully acknowledged by the authors.

REFERENCES

[1] S. Baxter and L. Graham, "Characterization of random composites using moving-window technique," *Journal of Engineering Mechanics*, vol. 126, no. 4, pp. 389–397, 2000.

[2] X. F. Xu and L. Graham-Brady, "A stochastic computational method for evaluation of global and local behavior of random elastic media," *Computer Methods in Applied Mechanics and Engineering*, vol. 194, no. 42, pp. 4362–4385, 2005.

[3] D. Charmpis, G. Schuëller, and M. Pellissetti, "The need for linking micromechanics of materials with stochastic finite elements: A challenge for materials science," *Computational Materials Science*, vol. 41, no. 1, pp. 27–37, 2007.

[4] A. Clément, C. Soize, and J. Yvonnet, "Uncertainty quantification in computational stochastic multiscale analysis of nonlinear elastic materials," *Computer Methods in Applied Mechanics and Engineering*, vol. 254, pp. 61–82, 2013.

[5] R. Hill, "Elastic properties of reinforced solids: some theoretical principles," *Journal of the Mechanics and Physics of Solids*, vol. 11, no. 5, pp. 357–372, 1963.

[6] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic analysis for periodic structures*, vol. 374. American Mathematical Soc., 2011.

[7] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, "Heterogeneous multiscale methods: a review," *Commun. Comput. Phys*, vol. 2, no. 3, pp. 367–450, 2007.

[8] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin, "Determination of the size of the representative volume element for random composites: statistical and numerical approach," *International Journal of solids and structures*, vol. 40, no. 13, pp. 3647–3679, 2003.

[9] M. Ostoja-Starzewski, "Material spatial randomness: From statistical to representative volume element," *Probabilistic engineering mechanics*, vol. 21, no. 2, pp. 112–132, 2006.

[10] X. F. Xu and X. Chen, "Stochastic homogenization of random elastic multi-phase composites and size quantification of representative volume element," *Mechanics of Materials*, vol. 41, no. 2, pp. 174–186, 2009.

[11] M. Silani, H. Talebi, S. Ziaei-Rad, P. Kerfriden, S. P. Bordas, and T. Rabczuk, "Stochastic modelling of clay/epoxy nanocomposites," *Composite Structures*, vol. 118, pp. 241–249, 2014.

[12] J. Ma, J. Zhang, L. Li, P. Wriggers, and S. Sahraee, "Random homogenization analysis for heterogeneous materials with full randomness and correlation in microstructure based on finite element method and monte-carlo method," *Computational Mechanics*, vol. 54, no. 6, pp. 1395–1414, 2014.

[13] J. Zeman and M. Sejnoha, "From random microstructures to representative volume elements," *Modelling and Simulation in Materials Science and Engineering*, vol. 15, no. 4, pp. 325–335, 2007.

[14] J. Wimmer, B. Stier, J.-W. Simon, and S. Reese, "Computational homogenisation from a 3d finite element model of asphalt concrete–linear elastic computations," *Finite Elements in Analysis and Design*, vol. 110, pp. 43–57, 2016.

[15] M. Ostoja-Starzewski and X. Wang, "Stochastic finite elements as a bridge between random material microstructure and global response," *Computer Methods in Applied Mechanics and Engineering*, vol. 168, no. 1, pp. 35–49, 1999.

[16] V. Lucas, J.-C. Golinval, S. Paquay, V.-D. Nguyen, L. Noels, and L. Wu, "A stochastic computational multiscale approach; application to mems resonators," *Computer Methods in Applied Mechanics and Engineering*, vol. 294, pp. 141–167, 2015.

[17] G. Stefanou, D. Savvas, and M. Papadrakakis, "Stochastic finite element analysis of composite structures based on material microstructure," *Composite Structures*, vol. 132, pp. 384–392, 2015.

[18] N. Moës, M. Cloirec, P. Cartraud, and J.-F. Remacle, "A computational approach to handle complex microstructure geometries," *Computer Methods in Applied Mechanics and Engineering*, vol. 192, no. 28, pp. 3163–3177, 2003.

[19] B. Hiriyur, H. Waisman, and G. Deodatis, "Uncertainty quantification in homogenization of heterogeneous microstructures modeled by XFEM," *International Journal for Numerical Methods in Engineering*, vol. 88, no. 3, pp. 257–278, 2011.

[20] D. Savvas, G. Stefanou, M. Papadrakakis, and G. Deodatis, "Homogenization of random heterogeneous media with inclusions of arbitrary shape modeled by XFEM," *Computational Mechanics*, vol. 54, no. 5, pp. 1221–1235, 2014.

[21] J. M. Melenk and I. Babuška, "The partition of unity finite element method: basic theory and applications," *Computer Methods in Applied Mechanics and Engineering*, vol. 139, no. 1, pp. 289–314, 1996.

[22] C. Miehe and A. Koch, "Computational micro-to-macro transitions of discretized microstructures undergoing small strains," *Archive of Applied Mechanics*, vol. 72, no. 4-5, pp. 300–317, 2002.

[23] S. H. R. Sanei and R. S. Fertig, "Uncorrelated volume element for stochastic modeling of microstructures based on local fiber volume fraction variation," *Composites Science and Technology*, vol. 117, pp. 191–198, 2015.

[24] Z. Hashin, "Analysis of composite materials: A survey," *Journal of Applied Mechanics*, vol. 50, no. 2, pp. 481–505, 1983.

[25] S. Hazanov and C. Huet, "Order relationships for boundary conditions effect in heterogeneous bodies smaller than the representative volume," *Journal of the Mechanics and Physics of Solids*, vol. 42, no. 12, pp. 1995–2011, 1994.

[26] T. Belytschko, R. Gracie, and G. Ventura, "A review of extended/generalized finite element methods for material modeling," *Modelling and Simulation in Materials Science and Engineering*, vol. 17, no. 4, pp. 1–24, 2009.

[27] D. Savvas, G. Stefanou, and M. Papadrakakis, "Determination of RVE size for random composites with local volume fraction variation," *Computer Methods in Applied Mechanics and Engineering*, in press, 2016.

[28] P. M. Suquet, "Elements of homogenization for inelastic solid mechanics," in *E. Sanchez-Palencia and A. Zaoui, Eds., Homogenization Techniques for Composite Media*, pp. 193–278, Springer, 1987.

[29] M. P. Sena, M. Ostoja-Starzewski, and L. Costa, "Stiffness tensor random fields through upscaling of planar random materials," *Probabilistic Engineering Mechanics*, vol. 34, pp. 131–156, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

NUMERICAL SIMULATION OF THE AUTOGENOUS SHRINKAGE OF HARDENING PORTLAND CEMENT PASTE

P. Gao[1,2a], G. Ye[2b], J.X. Wei[1c], Q.J. Yu[1d]

[1] School of Materials Science and Engineering, South China University of Technology
Guangzhou, 510640, People's Republic of China
[2] Microlab, Faculty of Civil Engineering and Geosciences, Delft University of Technology
Delft, 2628 CN, The Netherlands

[a]p.gao@tudelft.nl, [b]g.ye@tudelft.nl, [c]jxwei@scut.edu.cn, [d]concyuq@scut.edu.cn,

Keywords: Autogenous Shrinkage, Numerical Simulation, Capillary Pressure.

Abstract. *Autogenous shrinkage is the volume deformation of a closed, isothermal, cementitious material system not subjected to external forces. How to accurately predict the autogenous shrinkage in hardening cement-based materials is an important issue in concrete technology because autogenous shrinkage increases cracking risk and reduces the durability and service life of reinforced concrete structures. Many models simulated the autogenous shrinkage of cement-based materials based on some mechanisms such as capillary pressure and disjoining pressure. These models are normally empirical and cannot reveal the deformation behavior of cement-based materials under the internal load of the driving force of autogenous shrinkage in microscale. As a consequence, the reliability of these models are questionable.*

This paper proposes a numerical model to simulate the autogenous shrinkage of hardening cement paste. A cement hydration and microstructure model HYMOSTRUC3D is used to simulate the microstructure of cement paste. A pore morphology based method is applied to describe the water and empty capillary pores distribution in the microstructure. Capillary pressure in the microstructure of cement paste is calculated from relative humidity measured by experiment based on Kelvin equation. A discrete algorithm is proposed to divide the hydration time into several static times. At each static time, a lattice finite element fracture method is used to simulate the deformation of simulated microstructure of cement paste under capillary pressure imposing. The autogenous shrinkage of hardening cement paste is equal to the sum of deformation of cement paste at each static time. The autogenous shrinkage of Portland cement paste with water to cement ratio of 0.3 is predicted. The simulation results are in a good agreement with experiments.

1 INTRODUCTION

Autogenous shrinkage of cement-based materials (cement pastes, mortars and concretes, etc.) is defined as "*the bulk deformation of a closed, isothermal, cementitious material system not subjected to external forces*" [1]. Internal factors including the chemical shrinkage due to the cement hydration, the change of internal relative humidity caused by the consumption of water, and the pore structure of concrete are believed to result in autogenous shrinkage [2].

Autogenous shrinkage is harmful for concrete structures because it increases cracking risk and potentially reduces service life of concrete structures. Due to widespread usage of superplasticizers and high amount of supplementary cementitious materials (e.g. silica fume), modern concretes are characterized of low water to binder ratio (W/B) and fine pore structure, which significantly increases autogenous shrinkage [1-3].

Self-desiccation is considered as the main reason for autogenous shrinkage [1,2]. The products of cement hydration occupies less volume than the reactants during cement hydration. The volume difference between products and reactants is defined as chemical shrinkage [2]. Under sealed-conditions, chemical shrinkage causes "empty spaces" in the pore structure of cement-based materials, because the eternal water is not available to fill the volume difference due to cement hydration [3]. These "empty spaces" is then filled with vapour evaporated from water. After that, surface tension forms in the interface of gas phase (vapour) and liquid phase (water). Due to the effect of surface tension, meniscus forms [2,3]. Relative humidity (RH) in the vapour decreases at the same time due to meniscus formation. Thus this process is called self-desiccation. It is believed that self-desiccation rises internal tensile stresses in pore solution, which is considered as the driving force of autogenous shrinkage [2,3].

Intensive studies have investigated how does self-desiccation rise internal tensile stresses. Three fundamental mechanisms were proposed: surface tension mechanism, capillary pressure mechanism, and disjoining pressure mechanism [4-9]. Based on these mechanisms, some numerical models [4,7,8] were established to predict autogenous shrinkage. For example, Lura et al. [4] modelled the deformation of cement paste by a semi-macroscopic equation:

$$\varepsilon_{LIN} = \frac{S\sigma_{cap}}{3}\left(\frac{1}{K} - \frac{1}{K_S}\right) \tag{1}$$

where S is saturation fraction (-), σ_{cap} is the capillary pressure in the pore fluid (MPa), K is bulk modulus of the whole porous body (MPa) and K_S is bulk modulus of the solid material (MPa).

Hua et al. also proposed the models for predicting autogenous shrinkage at macroscopic scale [7] and the scale of hydrating grains [8]. Capillary pressure was considered as the driving force of autogenous shrinkage in both scales. At macroscopic scale, an empirical equation was adopted to calculate deformation [7], while at the scale of hydrating grains an finite element approach was applied to simulate the shrinkage of a hydrating grain [8].

Although these models can predict the macroscopic linear or volume deformation of cement-based materials, however, they cannot illustrate neither the impose process of internal tensile stresses on the solid skeleton of cement paste nor the corresponding deformation of solid skeleton in microscale. Solid skeleton is defined as the solid phase of cement paste that provides the mechanical strength of cement paste.

This paper aims to propose a numerical model to simulate the autogenous shrinkage of cement paste in microscale. In this model, both the impose process of internal tensile stress and the deformation of solid skeleton in microscale are simulated. Internal capillary pressure

is considered as the driving force of autogenous shrinkage. The numerical cement hydration and microstructure model , HYMOSTRUC3D [10-12] is used to simulate the microstructure of cement paste. The pore morphology-based method [13] is employed to distribute the water and empty capillary pore in the pore space. A discrete algorithm is proposed to divided the hydration time into several static times. At each static time, the lattice finite element fracture method [14] is adapted to predict the deformation of solid skeleton of cement paste under the load of internal capillary pressure. The autogenous shrinkage of hardening cement paste is equal to the sum of the deformation of cement paste at each static time.

2 MODELLING APPROACH

Internal capillary pressure is considered as the driving force of autogenous shrinkage in this simulation. The outline of the simulation process is shown in Figure 1. The simulation process can be divided into three stages. In stage 1, HYMOSTRUC3D [10-12] is used to obtain the degree of hydration and the microstructure of cement paste. In stage 2, the water and empty capillary pores distribution in the simulated pore structure of cement paste is modelled. In stage 3, the lattice finite element fracture method is applied to simulate the deformation of skeleton phase of cement paste under the impose of capillary pressure calculated from measured RH.

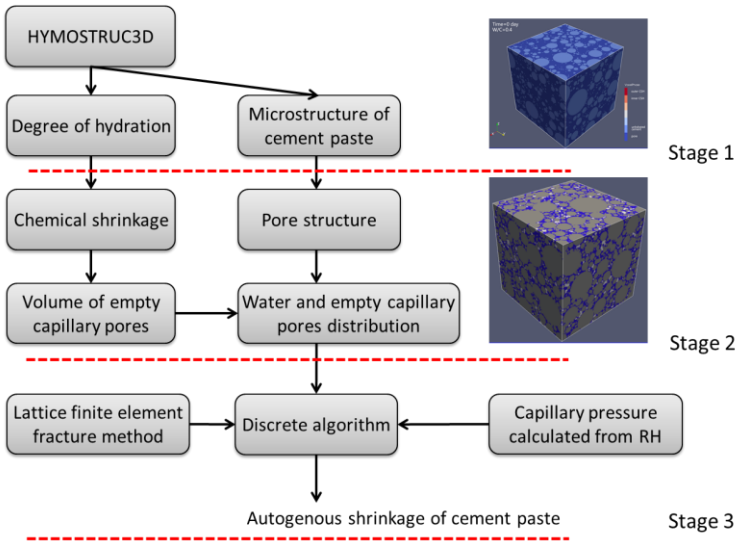


Figure 1 Simulation process of autogenous shrinkage of cement paste

2.1 Hydration and microstructure of Portland cement

HYMOSTRUC3D is a numerical model that can simulate the hydration and microstructure of Portland cement paste [10-12]. The numerical microstructure outputted by HYMOSTRUC3D contains several phases of Portland cement paste, including capillary pores, unhydrated cement, inner C-S-H, outer C-S-H, and calcium hydroxide (CH). For example, Figure 2 shows the simulated microstructure of Portland cement paste with a water to cement ratio (W/C) of 0.3 at the age of 7 days. The main input parameters of HYMOSTRUC3D include mineral composition of cement, particle size distribution of cement, W/C, and model parameters like penetration rate (K_0) and transition thickness (δ_{tr}) of cement. More information about the input parameters and model parameters of HYMOSTRUC3D can be referred to [10-12].

In this simulation, the mineral composition of Portland cement is: C_3S (62.0%), C_2S (10.5%), C_3A (7.3%), C_4AF (10.2%) [12]. The particle size distribution of Portland cement is shown in Figure 3 [12]. The W/C of cement paste is 0.3. As indicated in Table 1, the model parameters K_0 and δ_{tr} are calculated from the mineral composition of Portland cement based on the method described by Tuan [15].

Table 1 Two model parameters K_0 and δ_{tr} of different cement minerals [15]

Phase	K_0 (μm/hour)	δ_{tr} (μm)
C_3S	$0.0635+0.0195\times(1-\%C_3S)$	$2.1199+1.4707\times(1-\%C_3S)$
C_2S	$0.0033+0.0020\times(1-\%C_2S)$	$2.0730+1.1528\times(1-\%C_2S)$
C_3A	$1.2118-1.1714\times(1-\%C_3A)$	$2.3280+1.2758\times(1-\%C_3A)$
C_4AF	0.02	1.19

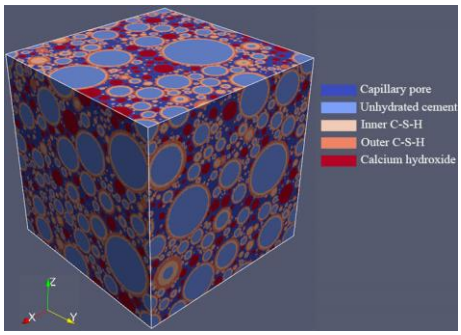


Figure 2 Simulated microstructure of Portland cement paste by extended HYMOSTRUC3D (100×100×100 μm^3, W/C=0.3, 1 day)

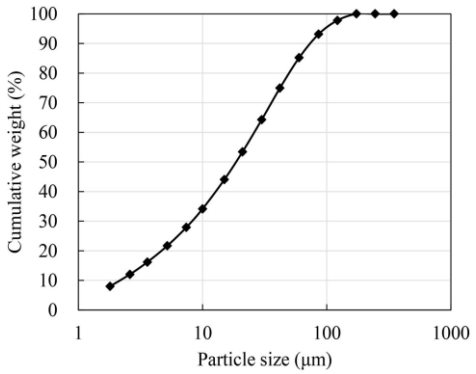


Figure 3 Particle size distribution of Portland cement

2.2 Water and empty capillary pores distribution

Because the products of cement hydration occupies less volume than the reactants of cement hydration (cement and water) some space that is not filled with water in the pore structure will be formed. The water free pace is defined as empty capillary pores. In this study we assume the pore space of cement paste contains two phases: water and empty capillary pores. The capillary pressure only exists in the liquid phase after the distribution of water and empty capillary pores is obtained. Therefore, the simulation of water and empty capillary pores is an necessary step for the simulation of the impose process of capillary pressure on the solid skeleton.

Water and empty capillary pores distribution can be simulated by pore-morphology based method established by Hilpert and Miller [13]. This method is commonly used to determine

the capillary pressure-saturation curves for different porous media [13,16]. In this study the input parameters of pore-morphology based method are the volume of empty capillary pores and the pore structure of cement paste. The volume of empty capillary pores equals to:

$$V_{\text{Empty capillary pores}} = V_{\text{Chemical shrinkage}} / V_{\text{cement paste}} \quad (2)$$

where $V_{\text{Empty capillary pores}}$ is the volume of empty capillary pore (ml/ml). $V_{\text{Chemical shrinkage}}$ is the volume of chemical shrinkage (ml/g). It can be simulated by HYMOSTRUC3D. $V_{\text{cement paste}}$ is the volume of cement paste (ml/g).

The pore structure of cement paste is obtained by HYMOTRUC3D.

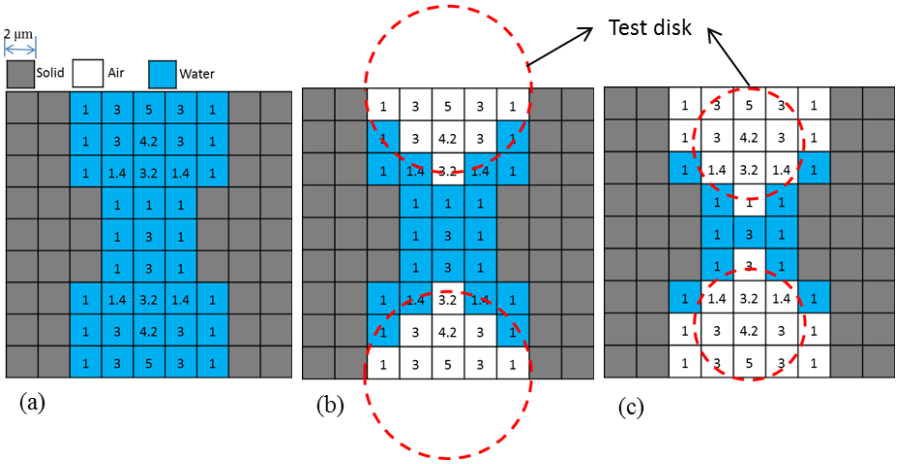


Figure 4 Schematic diagram of modelling self-desiccation process in cement paste based on pore-morphology based method (In 2D). Grey is solid, blue is water, and white is empty capillary pore. Note: Sphere is used to replace disk in the algorithm for 3D. The size of voxel in the 3D simulation is 1 μm.

(1) Figure out the distance map of pore space.

(2) All the voxels that belong to disk (R=5 μm) are labelled as empty capillary pore.

(3) In the next step, all the voxels that belong to disk (R=3.2 μm) are labelled as empty capillary pore. After this step, the radii of disk changes to 3 μm. The "drying" process continues until the volume of empty voxels is equal to that of the empty capillary pores.

The schematic diagram for simulating the distribution water and empty capillary pores is shown in Figure 4. The algorithm contains the following steps:

(1) The simulated microstructure is digitalized into small voxels with identical size. These voxels represent solid and pore space (Figure 4).

(2) Then the distance map of pore space is figured out. In the distance map, each voxel in pore space has a distance value. This distance value is defined as the shortest distance between the centre of this pore voxel to the solid phase [17]. Figure 4a shows the schematic diagram of a distance map. Besides, all the pore voxels are labelled as water in this step (Figure 4a).

(3) A test disk (in 2D) or sphere (in 3D) is defined to determine which pore voxels should be empty. The radii of test disk or sphere is initially equal to the largest distance value in the distance map. For each pore voxel with largest distance value, a test disk or sphere is set up, and the centre of test disk or sphere is fixed at the centre of this pore voxel. After that, all the pore voxels in the range of test disks or spheres are labelled as empty capillary pores. For the example in Figure 4b, because only two pore voxels with largest pore distance exist, two test disks with radii of 5 μm are given. The centres of the test disks are fixed at the centres of the pore voxels with the distance value of 5 μm, respectively. Then, all the pore voxels in the range of test disks are labelled as empty capillary pores.

(4) In the next step, the radii of test disk or sphere changes to be second largest distance value. Similarly, for each pore voxels with this distance value, a test disk or sphere is set up, and the centre of test disk or sphere is fixed at the centre of this pore voxel. Then, all the pore voxels in the range of test disks or spheres are labelled as empty capillary pores. For the example shown in Figure 4c, there are two test disks with the radii of 4.2 μm, because the number of the voxels with the radii of 4.2 μm is two. The centres of the test disks are fixed at the centres of the pore voxels with the distance value of 4.2 μm, respectively. All voxels in the range of test disks are labelled as empty capillary pores.

The iteration process described in (3) and (4) continues until the volume of total empty voxels are equal to that of empty capillary pores.

2.3 Autogenous shrinkage of cement paste

In stage 3 of Figure 1, a discrete algorithm is proposed to divide the hydration time into several static times. At each static time, a lattice finite element fracture analysis method is applied to simulate the impose process of capillary pressure on the solid skeleton of cement paste. The deformation of cement paste in microscale is then obtained. The autogenous shrinkage of hardening cement paste is equal to the sum of the deformation of cement paste at each static time. The model process are described as follow.

(1) Lattice finite element fracture analysis method

Lattice model has been used to simulate the fracture process of cement-based materials for decades [14]. Based on lattice model, Qian et al. [14] proposed a 3D lattice finite element fracture analysis to simulate the fracture process of cement paste. A GLAK (Generalized Lattice Analysis Kernel) package was developed to simulate the deformation of lattice mesh and the fracture process [14]. The main input files of GLAK package include lattice mesh and impose load. In this simulation the GLAK package is applied to simulate the impose of internal capillary pressure on the solid skeleton of cement paste and the corresponding deformation. The lattice mesh for GLAK package is generated from the microstructure of cement paste simulated by HYMOSTRUC3D. The internal capillary pressure, which is calculated from RH based on Kelvin equation, is considered as the impose load.

(2) Lattice mesh construction

The lattice mesh for GLAK package contains two fundamental parameters: lattice nodes and lattice beams [14]. Based on the method described in [14], this paper generates lattice mesh from the microstructure of cement paste simulated by HYMOSTRUC3D. The simulated microstructure of cement paste (Figure 5a and b) are classed as three phases (Figure 5c): solid phase (cement, inner product and outer product and CH indicated as grey), liquid phase (water indicated as blue), and vapor phase (empty capillary pore indicated as white). The lattice nodes are then randomly generated in each solid voxel. After that, beams are generated by connecting neighbor nodes (Figure 5d). The resolution of lattice mesh is 2 μm.

Table 2: Mechanical properties of different phases in microstructure (modified from [14])

Node type	Young's modulus (GPa)	Shear modulus (GPa)
Cement	135	52
Inner product	30	12
Outer product	22	8.9
CH	30	12

The mechanical properties of beams are obtained according to the method described by Qian et al. [14]. Since each beam in the lattice mesh connects two nodes, the mechanical properties of each beam is assumed to be the average of that of two nodes.

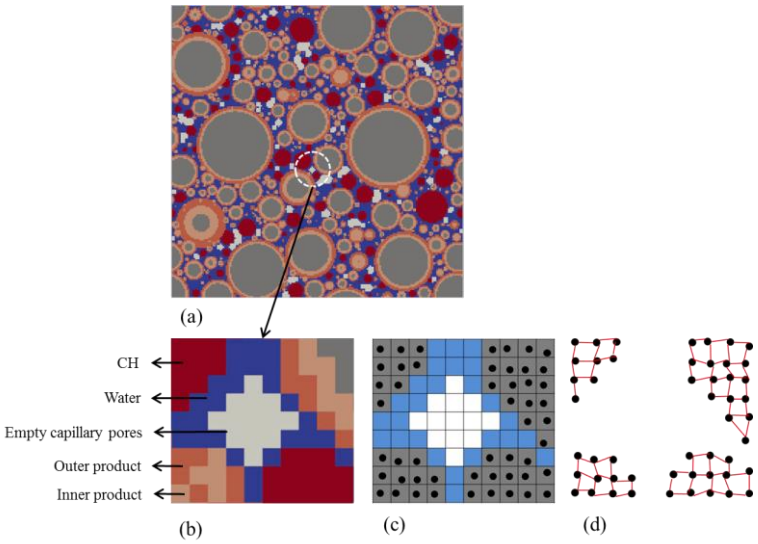


Figure 5 Schematic diagram of the construction of lattice mesh

(3) Impose of capillary pressure

During self-desiccation process, meniscus forms on the interfaces of water and vapour along with the formation of empty capillary pores. Due to meniscus formation, capillary pressure arises in the liquid and becomes an negative force that imposes on the solid skeleton.

In this simulation, the capillary pressure is assumed to be the same in any place of liquid. The capillary pressure can be calculated from RH [4]:

$$\Delta p = \frac{2\gamma}{r} = \frac{\ln RH_m \times RT}{V_m} \tag{3}$$

where Δp is capillary pressure, γ is surface tension, r is radii of pore, RH_m is modified RH based on the method described in [4]. RH can be experimentally tested [4]. R is the universal gas constant, T is temperature, V_m is the molar volume of the water.

The modified RH can be calculated by [4]:

$$RH_m = \frac{RH}{RH_s} \tag{4}$$

Because the salt concentration in the pore solution of cement paste normally changes with cement hydration, RH_S should be a function of hydration time. In this simulation, RH_S is simplified as a constant (98.2%) based on the report [4].

The capillary force imposed on the node is calculated by:

$$F_{capillary} = \Delta p \times S \tag{5}$$

where S is the area of the node. It depends on the resolution of lattice mesh.

The direction of capillary force on each node is figured out according to the rule: only the nodes that contact with liquid phase are imposed capillary force. For instance, the direction of capillary force on each node in Figure 5d is indicated in Figure 6.

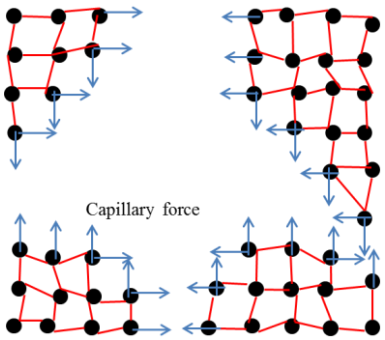


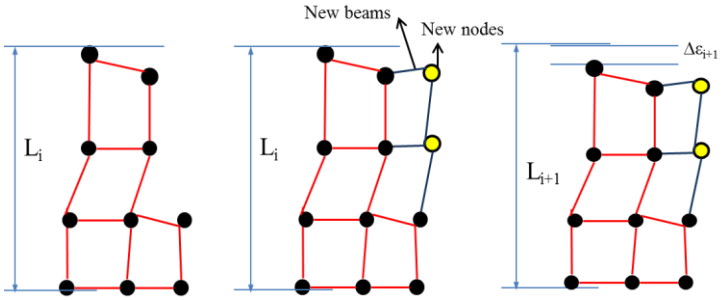
Figure 6 Impose of capillary pressure on lattice

(4) Discrete algorithm for calculating autogenous shrinkage of hardening cement paste

The above steps can calculate the deformation of cement paste at a static time. However, because both the microstructure of hardening cement paste and the capillary pressure in the pore solution change with the hydration time, the autogenous shrinkage of hardening cement paste is not static. We propose a discrete algorithm to calculate dynamic process of the autogenous shrinkage of hardening cement paste.

The schematic diagram of discrete algorithm for calculating autogenous shrinkage of hardening cement paste is shown in Figure 7. We assume at time i, the imposed force is F_i and the length of specimen is L_i after imposing. In the next time, new hydration products are generated in the microstructure, and the volume of solid phase increases due to cement hydration. Thus, the lattice mesh of microstructure at this time contains more nodes and beams than that at the last time. The increased nodes and beams are defined as new nodes and new beams (Figure 7b). The imposed force becomes to F_{i+1}. The extra force is defined as ΔF_{i+1}, being equal to F_{i+1}-F_i. After loading this extra force, an extra strain ($\Delta\varepsilon_{i+1}$) is formed. Then, the length of specimen (L_{i+1}) at time i+1 after loading extra force is equal to:

$$L_{i+1} = (1-\Delta\varepsilon_{i+1}) \times L_i \tag{6}$$



(a) After loading F_i at time i. (b) Before loading ΔF_{i+1} at time i+1. (c) After loading ΔF_{i+1} at time i+1
Figure 7 Schematic diagram of discrete algorithm for calculating autogenous shrinkage of hardening cement paste

The autogenous shrinkage of cement paste at time n is calculated by:

$$L_n = (1-\Delta\varepsilon_n) \times (1-\Delta\varepsilon_{n-1}) \cdots (1-\Delta\varepsilon_1) L_0 \tag{7}$$

where L_0 is the length of specimen at time zero.

The volume shrinkage of specimen is calculated by:

$$V_{Shrinkage} = 1-(1-L_n/L_0)^3 \tag{8}$$

In the discrete algorithm, the hydration time (from time zero to final testing time) is divided into several static times. The interval between two static times is not identical. Since cement commonly hydrates more quickly at early age than at later age, the interval is short at early age, and becomes long with increasing age. In this simulation, the period from time zero (5 hours after mixing) up to 7 days is divided into 37 static times (Figure 8).

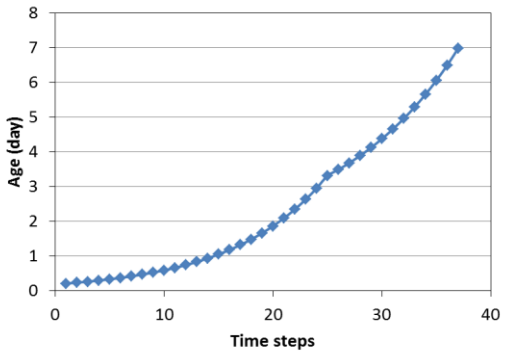


Figure 8 Time steps in the discrete algorithm

3 SIMULATION RESULTS AND DISCUSSION

3.1 Hydration and microstructure

The simulated degree of hydration of Portland cement paste with W/C of 0.3 is shown in Figure 9. The value is in a good agreement with experiment obtained by non-evaporable water method [12].

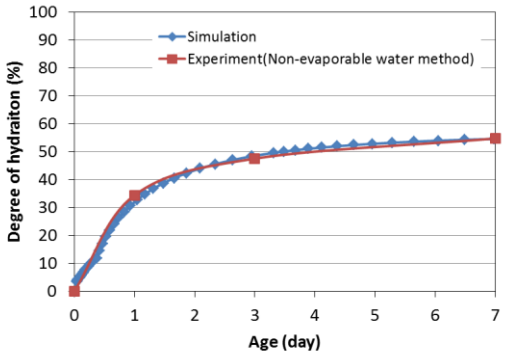


Figure 9 Simulated and experimental measured degree of hydration of cement paste

The simulated microstructures of Portland cement paste are show in Figure 10. The solid skeletons of cement paste are obtained.

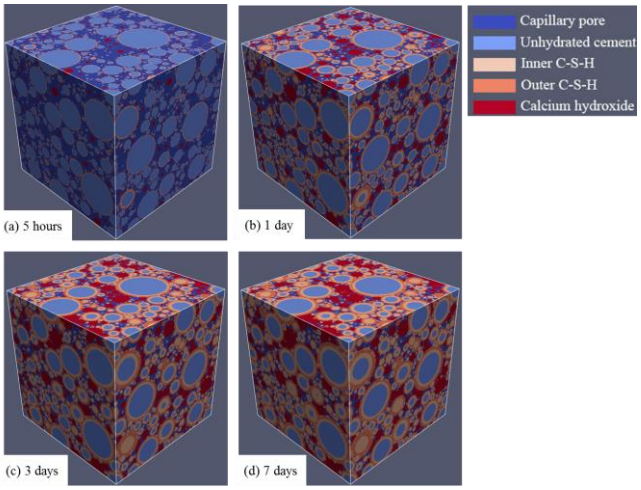


Figure 10 Simulated microstructures of cement pastes up to 7 days (100×100×100 μm^3)

3.2 Water and empty capillary pores distribution

The simulated water and empty capillary pores distribution is shown in Figure 11. It is clear that the empty capillary pores are firstly formed in "big" pore space.

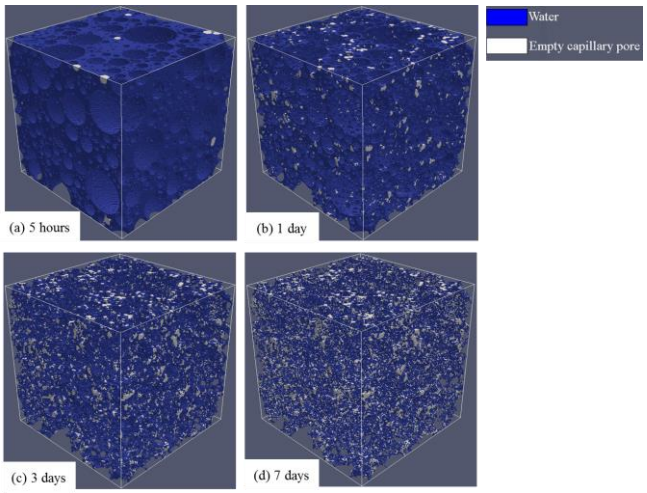
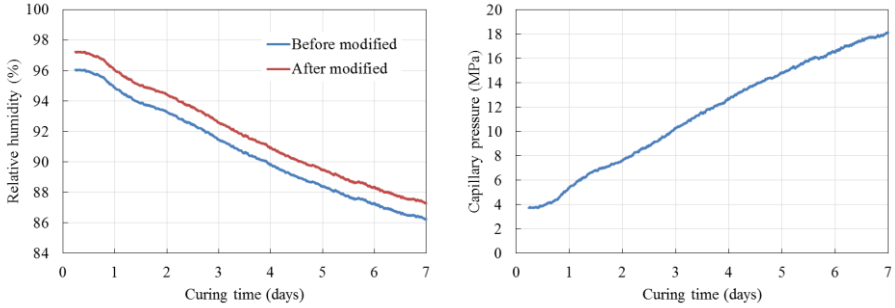


Figure 11 Simulated water and empty capillary pores distribution of cement pastes up to 7 days (100×100×100 μm^3)

3.3 Autogenous shrinkage of hardening cement paste

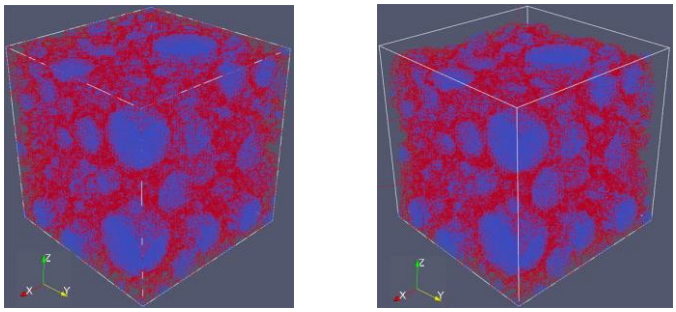
The RH used in the simulation is obtained by the experimental study from Lu and van Breugel [18]. The results are shown in Figure 12a. After modification by using Eq. 4, RH is increased (red line in Figure 12a). The capillary pressure is calculated based on Kelvin equation (Eq.3) and shown in Figure 12b.



(a) RH (b) Capillary pressure

Figure 12 Relative humidity (measured and modified by Eq. 4) and capillary pressure used in the simulation

The deformation of solid skeleton due to capillary pore is indicated in Figure 13. It illustrates that the solid skeleton shrinks due to the effect of capillary pressure (Figure 13b). The simulation results indicate that the surface of solid skeleton after shrinking is not smooth. This is because the beams have different mechanical properties. The mechanical properties of blue beams are higher than that of red beams, because blue beams consist of inner product and unhydrated cement, while red beams consist of outer product. In consequence, the deformation is not homogenous.



(a) Before imposing capillary pressure (b) After imposing capillary pressure

Figure 13 Simulated deformation of solid skeleton due to capillary pressure effect. (100×100×100 μm^3, 1 day) Note: the deformation is magnified for the purpose of visualization. Red beams are imposed load, and blue beam are not imposed load.

The simulated volume deformation of cement paste is compared with experimental data measured by Lu and van Breugel [18]. Lu and van Breugel used corrugated tube method to determine the autogenous shrinkage of Portland cement with a W/C of 0.3. Final setting time was adopted as time zero. The results of autogenous shrinkage of cement paste are shown in Figure 14. It is observed that the simulation result fits the experiment curve well.

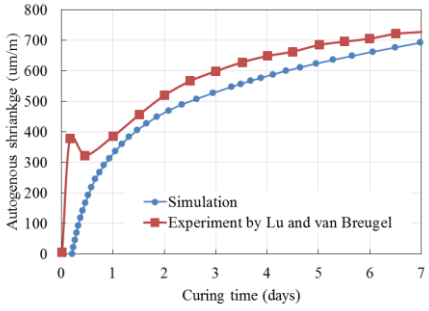


Figure 14 Autogenous shrinkage of cement paste

4 CONCLUSIONS

This paper proposes a numerical model to predict the autogenous shrinkage of hardening Portland cement paste. Capillary pressure is considered as the driving force for autogenous shrinkage. A discrete algorithm is proposed to calculate the autogenous shrinkage of hardening cement paste. In the discrete algorithm, the hydration time from 5 hours to 7 days is divided into 37 static times. For each static time, the lattice finite element fracture analysis method is employed to simulate the impose of increased capillary pressure on solid skeleton of cement paste. The dynamic autogenous shrinkage is equal to the sum of the deformation of cement paste at each step. The simulated autogenous shrinkage of hardening Portland cement paste (W/C=0.3) is in a good agreement with that measured from experiment.

5 ACKNOWLEDGEMENTS

This work was funded by the China Scholarship Council (CSC), the Research Centre of TU Delft in Urban System and Environment (No. C36103) and the National Basic Research Program of China (973 Program: 2011CB013800).

REFERENCES

[1] O.M. Jensen, P.F. Hansen, Autogenous deformation and RH-change in perspective. *Cement and Concrete Research*, **31**, 1859-1865, 2001.

[2] D.P. Bentz, O.M. Jensen, Mitigation strategies for autogenous shrinkage cracking. Cement and Concrete Composites, **26**, 677-685, 2004.

[3] D.P. Bentz, A review of early-age properties of cement-based materials. *Cement and Concrete Research*, **31**, 1859-1865, 2008.

[4] P. Lura, O.M. Jensen, K. van Breugel, Autogenous shrinkage in high-performance cement paste: An evaluation. *Cement and Concrete Research*, **33**, 913-922, 2003.

[5] K. van Breugel, Numerical modelling of volume changes at early ages-Potential, pitfalls and challenges. *Materials and Structures*, **34**, 293-301, 2001.

[6] F.H. Wittmann, Surface tension skrinkage and strength of hardened cement paste. *Matériaux et Construction*. **1**, 547-552, 1968.

[7] C. Hua, P. Acker, A. Ehrlacher, Analyses and models of the autogenous shrinkage of hardening cement paste: I. Modelling at macroscopic scale. *Cement and Concrete research*, **25**, 1457-1468, 1995.

[8] C. Hua, A. Ehrlacher, P. Acker, Analyses and models of the autogenous shrinkage of hardening cement paste II. Modelling at scale of hydrating grains. *Cement and concrete research*, **27**, 245-258, 1997.

[9] F. Beltzung, F.H. Wittmann, Role of disjoining pressure in cement based materials. *Cement and Concrete Research*, **35**, 2364-2370, 2005.

[10] K. van Breugel, Simulation of hydration and formation of structure in hardening cement-based materials, PhD Thesis, Delft, Delft University of Technology, The Netherlands, 1991.

[11] E.A.B. Koenders, Simulation of volume changes in hardening cement-based materials, PhD Thesis, Delft University of Technology, Delft, The Netherlands, 1997.

[12] G. Ye, Experimental Study and Numerical Simulation of the Development of the microstructure and permeability of cementitious materials, PhD Thesis, Delft, Delft University of Technology, The Netherlands, 2003.

[13] M. Hilpert, and C.T. Miller, Pore-morphology-based simulation of drainage in totally wetting porous media. *Advances in Water Resources*, **24**, 243-255, 2001.

[14] Z.W. Qian, E. Schlangen, G. Ye, K. van Breugel, Prediction of mechanical properties of cement paste at microscale. *Materiales de Construccion*, **60**, 7-18, 2010.

[15] Tuan N.V., Rice husk ash as a mineral admixture for ultra high performance concrete. PhD Thesis, Delft, Delft University of Technology, The Netherlands, 2011.

[16] V.P. Schulz, E.A. Wargo, E.C. Kumbur, Pore-Morphology-Based Simulation of Drainage in Porous Media Featuring a Locally Variable Contact Angle. *Transport in Porous Media*, 107, 13-25, 2015.

[17] E. Vergés, D. Ayala, S. Grau, D. Tost, 3D reconstruction and quantification of porous structures. *Computers and Graphics*, **32**, 438-444, 2008.

[18] T.S. Lu, K. van Bregugel, Analyzing and modelling time dependent-component of autogenous shrinkage of hardening cement paste and mortar. *14th International Congress on the Chemistry of Cement*, Beijing, China, October 13-16, 2015

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ANALYSIS OF LAMINATED COMPOSITE PLATES USING ISOGEOMETRIC COLLOCATION METHOD

G. S. Pavan [1], and K. S. Nanjunda Rao[2]

[1]Research Scholar
Department of Civil Engineering, Indian Institute of Science, Bangalore, India
e-mail: pavangs@civil.iisc.ernet.in

[2] Prinicipal Research Scientist
Department of Civil Engineering, Indian Institute of Science, Bangalore, India
e-mail: ksn@civil.iisc.ernet.in

Keywords: Isogeometric collocation Method, NURBS, Laminated composite plate, Sandwich plate, Reissner Mindlin plate theory

Abstract. *Isogeometric analysis (IGA) is a relatively new class of computational methods primarily aimed at integrating finite element analysis with Computer Aided Design (CAD). IGA adopts Non - Uniform Rational B Spline (NURBS) functions as the basis functions to represent geometry and approximate the unknown field variable. The standard Galerkin based isogeometric approach has been extensively explored by researchers during the past decade and is being continued. Apart from the standard Galerkin based isogeometric approach, several other numerical schemes exist within the framework of isogeometric methods. Isogeometric collocation method is one such relatively new numerical procedure. Isogeometric collocation methods discretize the governing PDEs in strong form. In this study, the potential of isogeometric collocation methods for the static analysis of laminated composite plates using Reissner - Mindlin plate theory is explored. Results of benchmark problems on bending of rectangular laminated composite plates subjected to sinusoidal and uniform loading are presented. The results obtained from the proposed isogeometric collocation method for the analysis of laminated composite plates have been assessed by comparing them with the solutions available in literature.*

1 INTRODUCTION

Isogeometric analysis (IGA) is a relatively new class of numerical methods proposed by Hughes et al. [1] in 2005 mainly aimed at integrating Computer Aided Design (CAD) and finite element analysis. IGA simplifies the complex procedure of mesh generation and mesh refinement thereby lowering the huge computational cost incurred during the analysis of large systems. IGA achieves this smooth flow of information by adopting the same NURBS functions employed to represent geometry by CAD models to approximate the unknown field variable in an isoparametric fashion. NURBS functions possess high continuity properties which lends IGA a significant edge over traditional finite element methods during discretization of higher order partial differential equations.

Over the last decade, IGA has been successfully applied on wide ranging topics such as solid and structural mechanics [2–10], fluid mechanics [11–17], phase field modeling [18], fluid structure interaction [19–21] etc. One of the important issue related to the efficient implementation of IGA was the development of an efficient integration rule for the Galerkin based IGA. The typical Gauss quadrature rules adopted for element wise integration in Galerkin based finite element methods is sub-optimal for Galerkin-IGA since it does not take into account the higher inter element continuity of NURBS functions.

One of the recent developments in the field of IGA is the emergence of Isogeometric collocation methods (IGA collocation) proposed by Auricchio et al. [22]. Isogeometric collocation method discretizes the governing differential equations in strong form while retaining the inherent IGA feature of adopting NURBS functions in an isoparametric sense. A comparison study between IGA collocation, Galerkin-IGA and finite element method by Schillinger et al. [23] revealed accuracy of the method achieved at a low cost of computation. The advantages offered by IGA collocation method are more evident when higher order approximations are employed.

Studies on IGA collocation have been performed on solid mechanics [24], phase field modeling [25], contact [26]. IGA collocation method has been successfully applied to Euler - Bernoulli beams and Kirchhoff plates [27]. It is worthy to mention that the high continuity properties possessed by NURBS functions enabled the application of IGA collocation method on Euler-Bernoulli beam and Kirchhoff plate theories which are governed by fourth order differential equations. Studies on performance of IGA collocation on shear deformable beams, rods and plates [28–30] has also been conducted. Kiendl et al. [30] showed that mixed formulation within the framework of IGA collocation is successful in alleviating shear locking observed during lower order approximations for plate bending.

In this paper, we propose IGA collocation method for the analysis of laminated composite plates described by Reissner - Mindlin plate theory. Standard formulation with lateral displacement and two rotations are the three unknown field variables within the framework of IGA collocation. Performance of IGA collocation is assessed by solv-

ing benchmark problems and comparing the results obtained with the ones existing in literature.

2 REISSNER MINDLIN PLATE THEORY FOR LAMINATED COMPOSITE PLATES

A laminated composite plate of total thickness *t* composed of *N* orthotropic layers is considered. The undeformed midplane of the plate is taken as the reference plane Ω_0. The kinematic assumptions underlying the Reissner - Mindlin theory are made, namely, cross sections remain straight during deformation but may not remain normal to the reference plane. The displacement field according to Reissner Mindlin plate theory is

$$\begin{aligned} u(x,y,z) &= u_0(x,y) + z\phi_x(x,y) \\ v(x,y,z) &= v_0(x,y) + z\phi_y(x,y) \\ w(x,y,z) &= w_0(x,y) \end{aligned} \tag{1}$$

where u_0,v_0, w_0 are the displacement components for the midplane in x, y and z directions respectively. The rotations in y-z plane and x-z plane are denoted as ϕ_x and ϕ_y respectively. The strains associated with the displacement field mentioned in equation (1) are

$$\begin{aligned} \epsilon_{xx} &= \frac{\partial u_0}{\partial x} + z\frac{\partial \phi_x}{x} & \gamma_{xy} &= \left(\frac{\partial u_0}{\partial y} + \frac{\partial v_0}{\partial x}\right) + z\left(\frac{\partial \phi_x}{\partial y} + \frac{\partial \phi_y}{\partial x}\right) \\ \epsilon_{yy} &= \frac{\partial v_0}{\partial y} + z\frac{\partial \phi_y}{\partial y} & \gamma_{yz} &= \frac{\partial w_0}{\partial y} + \phi_y \\ \gamma_{xz} &= \frac{\partial w_0}{\partial x} + \phi_x & \epsilon_{zz} &= 0 \end{aligned}$$

The stresses at the kth layer in local coordinates are given by the following constitutive relationship.

$$\begin{Bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{Bmatrix}^k = \begin{bmatrix} Q_{11} & Q_{12} & 0 & 0 & 0 \\ Q_{21} & Q_{22} & 0 & 0 & 0 \\ 0 & 0 & Q_{66} & 0 & 0 \\ 0 & 0 & 0 & Q_{55} & 0 \\ 0 & 0 & 0 & 0 & Q_{44} \end{bmatrix}^k \begin{Bmatrix} \epsilon_{xx} \\ \epsilon_{yy} \\ \gamma_{xy} \\ \gamma_{xz} \\ \gamma_{yz} \end{Bmatrix}^k \tag{2}$$

where

$$\begin{aligned} Q_{11} &= \frac{E_1}{1-\nu_{12}\nu_{21}} & Q_{12} &= \frac{E_2\nu_{12}}{1-\nu_{12}\nu_{21}} & Q_{22} &= \frac{E_2}{1-\nu_{12}\nu_{21}} \\ Q_{66} &= G_{12} & Q_{55} &= G_{13} & Q_{44} &= G_{23} \end{aligned} \tag{3}$$

where E_1 and E_2 are the elastic modulus in directions 1 and 2 respectively. G_{12}, G_{13} and G_{23} are the shear modulus in x-y, x-z and y-z planes respectively. ν_{12} and ν_{21} are the two Poisson ratios for orthotropic materials. The material constants in equation (2) v in local coordinates can be expressed in the global coordinates by the following set of relations

$$
\begin{aligned}
Q_{11}^{'} &= Q_{11}cos^4\theta + 2\left(Q_{12} + 2Q_{66}\right) sin^2\theta cos^2\theta + Q_{22}sin^4\theta \\
Q_{12}^{'} &= \left(Q_{11} + Q_{22} - 4Q_{66}\right) sin^2\theta cos^2\theta + Q_{12}\left(sin^4\theta + cos^4\theta\right) \\
Q_{22}^{'} &= Q_{11}sin^4\theta + 2\left(Q_{12} + 2Q_{66}\right) sin^2\theta cos^2\theta + Q_{22}cos^4\theta \\
Q_{16}^{'} &= \left(Q_{11} - Q_{12} - 2Q_{66}\right) sin\theta cos^3\theta + \left(Q_{12} - Q_{22} + 2Q_{66}\right) sin^3\theta cos\theta \\
Q_{26} &= \left(Q_{11} - Q_{12} - 2Q_{66}\right) sin^3\theta cos\theta + \left(Q_{12} - Q_{22} + 2Q_{66}\right) sin\theta cos^3\theta \\
Q_{66}^{'} &= \left(Q_{11} + Q_{22} - 2Q_{12} - 2Q_{66}\right) sin^2\theta cos^2\theta + Q_{66}\left(sin^4\theta + cos^4\theta\right) \\
Q_{44}^{'} &= Q_{44}cos^2\theta + Q_{55}sin^2\theta \\
Q_{45}^{'} &= \left(Q_{55} - Q_{44}\right) cos\theta sin\theta \\
Q_{55}^{'} &= Q_{44}sin^2\theta + Q_{55}cos^2\theta
\end{aligned} \tag{4}
$$

Here, θ is the inclination between the local co-ordinate system at the kth layer and the global co-ordinate axes.

2.1 Equations of equilibrium

The equations of equilibrium expressed in terms of the bending moments M_{xx}, M_{yy}, the twisting moments M_{xy}, shear forces Q_x, Q_y and the in - plane forces N_{xx}, N_{yy} and N_{xy} are given below

$$N_{xx,x} + N_{xy,y} = 0 \tag{5a}$$
$$N_{xy,x} + N_{yy,y} = 0 \tag{5b}$$
$$Q_{x,x} + Q_{y,y} = -f \tag{5c}$$
$$M_{xx,x} + M_{xy_y} - Q_x = 0 \tag{5d}$$
$$M_{yy,y} + M_{yx,x} - Q_y = 0 \tag{5e}$$

The stress resultants in equations (5a) are related to displacement variables by the following relations

$$\begin{Bmatrix} N_{xx} \\ N_{yy} \\ N_{xy} \end{Bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{16} \\ A_{12} & A_{22} & A_{26} \\ A_{16} & A_{26} & A_{66} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial u_0}{\partial x} \\ \dfrac{\partial v_0}{\partial y} \\ \dfrac{\partial u_0}{\partial y} + \dfrac{\partial v_0}{\partial x} \end{Bmatrix} + \begin{bmatrix} B_{11} & B_{12} & B_{16} \\ B_{12} & B_{22} & B_{26} \\ B_{16} & B_{26} & B_{66} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial \phi_x}{\partial x} \\ \dfrac{\partial \phi_y}{\partial y} \\ \dfrac{\partial \phi_x}{\partial y} + \dfrac{\partial \phi_y}{\partial x} \end{Bmatrix} \tag{6a}$$

$$\begin{Bmatrix} M_{xx} \\ M_{yy} \\ M_{xy} \end{Bmatrix} = \begin{bmatrix} B_{11} & B_{12} & B_{16} \\ B_{12} & B_{22} & B_{26} \\ B_{16} & B_{26} & B_{66} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial u_0}{\partial x} \\ \dfrac{\partial v_0}{\partial y} \\ \dfrac{\partial u_0}{\partial y} + \dfrac{\partial v_0}{\partial x} \end{Bmatrix} + \begin{bmatrix} D_{11} & D_{12} & D_{16} \\ D_{12} & D_{22} & D_{26} \\ D_{16} & D_{26} & D_{66} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial \phi_x}{\partial x} \\ \dfrac{\partial \phi_y}{\partial y} \\ \dfrac{\partial \phi_x}{\partial y} + \dfrac{\partial \phi_y}{\partial x} \end{Bmatrix} \tag{6b}$$

$$\begin{Bmatrix} Q_y \\ Q_x \end{Bmatrix} = K \begin{bmatrix} A_{44} & A_{45} \\ A_{45} & A_{55} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial w_0}{\partial y} + \phi_y \\ \dfrac{\partial w_0}{\partial x} + \phi_x \end{Bmatrix} \tag{6c}$$

The A_{ij}, B_{ij} and the D_{ij} present in the above set of equations (6) are termed as extensional stiffness, bending stiffness and bending - extensional stiffness respectively and are defined as following

$$(A_{ij}, B_{ij}, D_{ij}) = \int_{-t/2}^{t/2} Q_{ij}^{'} \left(1, z, z^2\right) dz \tag{7}$$

Substituting the stress resultant - displacement variable relationships mentioned in equation (6) into (5a), we obtain the equilibrium equations in terms of displacement variables u_0,v_0, w_0, ϕ_x and ϕ_y as shown below

$$A_{11}\frac{\partial^2 u_0}{\partial x^2}+2A_{16}\frac{\partial^2 u_0}{\partial x\partial y}+A_{66}\frac{\partial^2 u_0}{\partial y^2}+A_{12}\frac{\partial^2 v_0}{\partial x\partial y}+A_{16}\frac{\partial^2 v_0}{\partial x^2}+A_{26}\frac{\partial^2 v_0}{\partial y^2}+A_{66}\frac{\partial^2 v_0}{\partial x\partial y}+$$
$$B_{11}\frac{\partial^2 \phi_x}{\partial x^2}+2B_{16}\frac{\partial^2 \phi_x}{\partial x\partial y}+B_{66}\frac{\partial^2 \phi_x}{\partial y^2}+B_{12}\frac{\partial^2 \phi_y^h}{\partial x\partial y}+B_{16}\frac{\partial^2 \phi_y^h}{\partial x^2}+B_{26}\frac{\partial^2 \phi_y^h}{\partial y^2}+B_{66}\frac{\partial^2 \phi_y^h}{\partial x\partial y}=0 \tag{8a}$$

$$A_{16}\frac{\partial^2 u_0}{\partial x^2}+A_{66}\frac{\partial^2 u_0}{\partial x\partial y}+A_{12}\frac{\partial^2 u_0}{\partial x\partial y}+A_{26}\frac{\partial^2 u_0}{\partial y^2}+A_{26}\frac{\partial^2 v_0}{\partial x\partial y}+A_{66}\frac{\partial^2 v_0}{\partial x^2}+A_{22}\frac{\partial^2 v_0}{\partial y^2}+$$
$$A_{26}\frac{\partial^2 v_0}{\partial x\partial y}+B_{16}\frac{\partial^2 \phi_x}{\partial x^2}+B_{66}\frac{\partial^2 \phi_x}{\partial x\partial y}+B_{12}\frac{\partial^2 \phi_x}{\partial y\partial x}+B_{26}\frac{\partial^2 \phi_x}{\partial y^2}+B_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y}+B_{66}\frac{\partial^2 \phi_y^h}{\partial x^2}+$$
$$B_{22}\frac{\partial^2 \phi_y^h}{\partial y^2}+B_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y}=0 \tag{8b}$$

$$KA_{55}\frac{\partial \phi_x}{\partial x}+KA_{45}\frac{\partial \phi_x}{\partial y}+KA_{45}\frac{\partial \phi_y^h}{\partial x}+KA_{44}\frac{\partial \phi_y^h}{\partial y}+KA_{55}\frac{\partial^2 w_0}{\partial x^2}+2KA_{45}\frac{\partial^2 w_0}{\partial x\partial y}$$
$$+KA_{44}\frac{\partial^2 w_0}{\partial y^2}+f=0 \tag{8c}$$

$$B_{11}\frac{\partial^2 u_0}{\partial x^2}+2B_{16}\frac{\partial^2 u_0}{\partial x\partial y}+B_{66}\frac{\partial^2 u_0}{\partial y^2}+B_{12}\frac{\partial^2 v_0}{\partial x\partial y}+B_{16}\frac{\partial^2 v_0}{\partial x^2}+B_{26}\frac{\partial^2 v_0}{\partial y^2}+B_{66}\frac{\partial^2 v_0}{\partial x\partial y}+$$
$$D_{11}\frac{\partial^2 \phi_x}{\partial x^2}+2D_{16}\frac{\partial^2 \phi_x}{\partial x\partial y}+D_{66}\frac{\partial^2 \phi_x}{\partial y^2}+D_{12}\frac{\partial^2 \phi_y^h}{\partial x\partial y}+D_{16}\frac{\partial^2 \phi_y^h}{\partial x^2}+D_{26}\frac{\partial^2 \phi_y^h}{\partial y^2}+D_{66}\frac{\partial^2 \phi_y^h}{\partial x\partial y}-$$
$$KA_{55}\left(\frac{\partial w_0}{\partial x}+\phi_x\right)-KA_{45}\left(\frac{\partial w_0}{\partial y}+\phi_y\right)=0 \tag{8d}$$

$$B_{16}\frac{\partial^2 u_0}{\partial x^2}+B_{66}\frac{\partial^2 u_0}{\partial x\partial y}+B_{12}\frac{\partial^2 u_0}{\partial x\partial y}+B_{26}\frac{\partial^2 u_0}{\partial y^2}+B_{26}\frac{\partial^2 v_0}{\partial x\partial y}+B_{66}\frac{\partial^2 v_0}{\partial x^2}+B_{22}\frac{\partial^2 v_0}{\partial y^2}+$$
$$B_{26}\frac{\partial^2 v_0}{\partial x\partial y}+D_{16}\frac{\partial^2 \phi_x}{\partial x^2}+D_{66}\frac{\partial^2 \phi_x}{\partial x\partial y}+D_{12}\frac{\partial^2 \phi_x}{\partial y\partial x}+D_{26}\frac{\partial^2 \phi_x}{\partial y^2}+D_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y}+D_{66}\frac{\partial^2 \phi_y^h}{\partial x^2}+$$
$$D_{22}\frac{\partial^2 \phi_y^h}{\partial y^2}+D_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y}-KA_{45}\left(\frac{\partial w_0}{\partial x}+\phi_x\right)-KA_{44}\left(\frac{\partial w_0}{\partial y}+\phi_y\right)=0 \tag{8e}$$

The above equilibrium equations eq. (8) are valid in the interior of the domain. The equilibrium equations at the Neumann boundary are given by:

$$\begin{bmatrix} n_x^2 & n_y^2 & 2n_xn_y \\ -n_xn_y & n_xn_y & n_x^2-n_y^2 \end{bmatrix}\begin{Bmatrix} N_{xx} \\ N_{yy} \\ N_{xy} \end{Bmatrix}=\begin{Bmatrix} \widehat{N}_{nn} \\ \widehat{N}_{ns} \end{Bmatrix} \tag{9a}$$

$$\begin{bmatrix} n_x^2 & n_y^2 & 2n_xn_y \\ -n_xn_y & n_xn_y & n_x^2-n_y^2 \end{bmatrix}\begin{Bmatrix} M_{xx} \\ M_{yy} \\ M_{xy} \end{Bmatrix}=\begin{Bmatrix} \widehat{M}_{nn} \\ \widehat{M}_{ns} \end{Bmatrix} \tag{9b}$$

$$\begin{bmatrix} n_x & n_y \end{bmatrix}\begin{Bmatrix} Q_x \\ Q_y \end{Bmatrix}=\begin{Bmatrix} \widehat{Q}_n \end{Bmatrix} \tag{9c}$$

where $\widehat{N}_{nn}$ & $\widehat{N}_{ns}$ are the in - plane axial forces specified at the Neumann boundary. Similary, $\widehat{M}_{nn}$ & $\widehat{M}_{ns}$ are the bending moments and, $\widehat{Q}_n$ are the shear forces specified at Neumann boundary.

Using the stress resultant - displacement relationship in eq. (6), we obtain eqs. (9) in terms of displacement variables as given below:

$$\widehat{N}_{nn} = \left(n_x^2 A_{11} + n_y^2 A_{12} + 2n_x n_y A_{16}\right)\frac{\partial u_0}{\partial x} + \left(n_x^2 A_{16} + n_y^2 A_{26} + 2n_x n_y A_{66}\right)\frac{\partial u_0}{\partial y} + \left(n_x^2 A_{16} + n_y^2 A_{26} + 2n_x n_y A_{66}\right)\frac{\partial v_0}{\partial x} + \left(n_x^2 A_{12} + n_y^2 A_{22} + 2n_x n_y A_{26}\right)\frac{\partial v_0}{\partial y} + \left(n_x^2 B_{11} + n_y^2 B_{12} + 2n_x n_y B_{16}\right)\frac{\partial \phi_x}{\partial x} + \left(n_x^2 B_{16} + n_y^2 B_{26} + 2n_x n_y B_{66}\right)\frac{\partial \phi_x}{\partial y} + \left(n_x^2 B_{16} + n_y^2 B_{26} + 2n_x n_y B_{66}\right)\frac{\partial \phi_y}{\partial x} + \left(n_x^2 B_{12} + n_y^2 B_{22} + 2n_x n_y B_{26}\right)\frac{\partial \phi_y}{\partial y} \tag{10a}$$

$$\widehat{N}_{ns} = \left(-n_x n_y A_{11} + n_x n_y A_{12} + \left(n_x^2 - n_y^2\right) A_{16}\right)\frac{\partial u_0}{\partial x} + \left(-n_x n_y A_{16} + n_x n_y A_{26} + \left(n_x^2 - n_y^2\right) A_{66}\right)\frac{\partial u_0}{\partial y} + \left(-n_x n_y A_{16} + n_x n_y A_{26} + \left(n_x^2 - n_y^2\right) A_{66}\right)\frac{\partial v_0}{\partial x} + \left(-n_x n_y A_{12} + n_x n_y A_{22} + \left(n_x^2 - n_y^2\right) A_{26}\right)\frac{\partial v_0}{\partial y} + \left(-n_x n_y b_{11} + n_x n_y B_{12} + \left(n_x^2 - n_y^2\right) B_{16}\right)\frac{\partial \phi_x}{\partial x} + \left(-n_x n_y B_{16} + n_x n_y B_{22} + \left(n_x^2 - n_y^2\right) B_{66}\right)\frac{\partial \phi_x}{\partial y} + \left(-n_x n_y B_{16} + n_x n_y B_{26} + \left(n_x^2 - n_y^2\right) B_{66}\right)\frac{\partial \phi_y}{\partial x} + \left(-n_x n_y B_{12} + n_x n_y B_{22} + \left(n_x^2 - n_y^2\right) B_{26}\right)\frac{\partial \phi_y}{\partial y} \tag{10b}$$

$$\widehat{Q}_n = K\left[A_{45}\frac{\partial w_0}{\partial y}n_x + A_{55}\frac{\partial w_0}{\partial x}n_x + A_{44}\frac{\partial w_0}{\partial y}n_y + A_{45}\frac{\partial w_0}{\partial x}n_y\right] + K\left[A_{55}\phi_x n_x + A_{45}\phi_x n_y\right] + K\left[A_{45}\phi_y n_x + A_{44}n_y\phi_y\right] \tag{10c}$$

$$\widehat{M}_{nn} = \left(n_x^2 B_{11} + n_y^2 B_{12} + 2n_x n_y B_{16}\right)\frac{\partial u_0}{\partial x} + \left(n_x^2 B_{16} + n_y^2 B_{26} + 2n_x n_y B_{66}\right)\frac{\partial u_0}{\partial y} + \left(n_x^2 B_{16} + n_y^2 B_{26} + 2n_x n_y B_{66}\right)\frac{\partial v_0}{\partial x} + \left(n_x^2 B_{12} + n_y^2 B_{22} + 2n_x n_y B_{26}\right)\frac{\partial v_0}{\partial y} + \left(n_x^2 D_{11} + n_y^2 D_{12} + 2n_x n_y d_{16}\right)\frac{\partial \phi_x}{\partial x} + \left(n_x^2 B_{16} + n_y^2 B_{26} + 2n_x n_y B_{66}\right)\frac{\partial \phi_x}{\partial y} + \left(n_x^2 B_{16} + n_y^2 B_{26} + 2n_x n_y B_{66}\right)\frac{\partial \phi_y}{\partial x} + \left(n_x^2 B_{12} + n_y^2 B_{22} + 2n_x n_y B_{26}\right)\frac{\partial \phi_y}{\partial y} \tag{10d}$$

$$\widehat{M}_{ns} = \left(-n_x n_y B_{11} + n_x n_y B_{12} + \left(n_x^2 - n_y^2\right) B_{16}\right)\frac{\partial u_0}{\partial x} + \left(-n_x n_y B_{16} + n_x n_y B_{26} + \left(n_x^2 - n_y^2\right) B_{66}\right)\frac{\partial u_0}{\partial y} + \left(-n_x n_y B_{16} + n_x n_y B_{26} + \left(n_x^2 - n_y^2\right) B_{66}\right)\frac{\partial v_0}{\partial x} + \left(-n_x n_y B_{12} + n_x n_y B_{22} + \left(n_x^2 - n_y^2\right) B_{26}\right)\frac{\partial v_0}{\partial y} + \left(-n_x n_y D_{11} + n_x n_y D_{12} + \left(n_x^2 - n_y^2\right) D_{16}\right)\frac{\partial \phi_x}{\partial x} + \left(-n_x n_y D_{16} + n_x n_y D_{22} + \left(n_x^2 - n_y^2\right) D_{66}\right)\frac{\partial \phi_x}{\partial y} + \left(-n_x n_y D_{16} + n_x n_y D_{26} + \left(n_x^2 - n_y^2\right) D_{66}\right)\frac{\partial \phi_y}{\partial x} + \left(-n_x n_y D_{12} + n_x n_y D_{22} + \left(n_x^2 - n_y^2\right) D_{26}\right)\frac{\partial \phi_y}{\partial y} \tag{10e}$$

3 Numerical formulation

In order to solve the equations presented in the previous section using Isogeometric collocation method, the unknown field variables are expressed as a linear combination of Non - Uniform Rational B - Splines (NURBS).

3.1 B - Splines and NURBS

B-spline functions are polynomials, say of degree p, are defined piecewise on the parametric space. The parametric space is given by a knot vector. A knot vector Ξ, is a set of non - decreasing coordinates ξ_i in the parametric space, $\Xi = \{\xi_1, \xi_2, \xi_3, ..., \xi_{n+p+1}\}$

, where n is the number of basis functions. The space between two consecutive knots in the parametric space is defined as a knot span. Knot values can be repeated at a particular location in the parametric space and are called as multiple knots. The B-splines are C^{p-1} continuous at a particular knot and C^{p-k} continuous at multiple knot locations with multiplicity k.

The ith B-spline basis function of degree p for a given knot vector Ξ is defined recursively as shown below

$$N_{i,0}(\xi) = \begin{cases} 1, & \xi_i \leq \xi < \xi_{i+1} \\ 0, & otherwise \end{cases} \tag{11a}$$

$$N_{i,p} = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(\xi) \tag{11b}$$

B-Splines in two dimensions are obtained by taking the tensor product of their univariate counterparts as shown below

$$B_{ij,pq}(\xi.\eta) = N_{i,p}(\xi) M_{j,q}(\eta) \qquad i = 1, ..., n \qquad j = 1, ..., m. \tag{12}$$

where ξ and η are the parametric coordinates in two directions.

A bivariate NURBS function R_{ij} is defined as the weighted bivariate B-Spline B_{ij} function,

$$R_{ij}(\xi, \eta) = \frac{B_{ij,pq}(\xi, \eta)\, \omega_{ij}}{\sum_{k=1}^{n} \sum_{l=1}^{m} B_{ij,pq}(\xi, \eta)\, \omega_{ij}} \tag{13}$$

ω_{ij} are weights adopted.

The isogeometric class of numerical methods adopt NURBS functions to discretize the unknown field variables and geometry. Likewise, isogeometric collocation methods adopt NURBS to approximate the unknown field variables, namely, u, v, w, ϕ_x & ϕ_y as shown below:

$$u_0^h(\xi,\eta) = \sum_{i=1}^{n}\sum_{i=1}^{m} R_{ij}^w(\xi,\eta)\,\widehat{u}^{ij} \tag{14a}$$

$$v_0^h(\xi,\eta) = \sum_{i=1}^{n}\sum_{i=1}^{m} R_{ij}^w(\xi,\eta)\,\widehat{v}^{ij} \tag{14b}$$

$$w_0^h(\xi,\eta) = \sum_{i=1}^{n}\sum_{i=1}^{m} R_{ij}^w(\xi,\eta)\,\widehat{w}^{ij} \tag{14c}$$

$$\phi_x^h(\xi,\eta) = \sum_{i=1}^{n}\sum_{i=1}^{m} R_{ij}^w(\xi,\eta)\,\hat{\phi_x}^{ij} \tag{14d}$$

$$\phi_y^h(\xi,\eta) = \sum_{i=1}^{n}\sum_{i=1}^{m} R_{ij}^w(\xi,\eta)\,\hat{\phi_y}^{ij} \tag{14e}$$

3.2 Collocation using NURBS functions

The discretized strong form of the equilibrium equations are collocated on the images of greville points in physical space. The greville points are defined in the parametric space. The greville points $\hat{\xi}_i$ defined for a spline space of degree p and knot vector $\{\xi_i, ..., \xi_{i+p+1}\}$ are given by

$$\hat{\xi}_i = \frac{\xi_{i+1} + \xi_{i+2} + ... + \xi_{i+p}}{p} \tag{15}$$

In two dimensional problems like plate problems, a greville point is defined by $\left(\hat{\xi}_i, \hat{\eta}_i\right)$, which are the greville coordinates in two parametric directions. From eq. (15), it can be observed that some of the greville points are located within the domain and others are located at the boundary. Since the equations of equilibrium are different for the domain interior and boundary, the equations to be collocated need to be chosen suitably depending on the location of the greville point.

Using eqs.(14), we discretize the equilibrium equations eqs. (8) defined in the interior of the domain. The residual of these equations $\left(R_u^\Omega, R_v^\Omega, R_w^\Omega, R_{\phi_x}^\Omega \& R_{\phi_y}^\Omega\right)$ are presented as follows:

$$R_u^\Omega = A_{11}\frac{\partial^2 u_0^h}{\partial x^2} + 2A_{16}\frac{\partial^2 u_0^h}{\partial x\partial y} + A_{66}\frac{\partial^2 u_0^h}{\partial y^2} + A_{12}\frac{\partial^2 v_0^h}{\partial x\partial y} + A_{16}\frac{\partial^2 v_0^h}{\partial x^2} + A_{26}\frac{\partial^2 v_0^h}{\partial y^2} + A_{66}\frac{\partial^2 v_0^h}{\partial x\partial y} + B_{11}\frac{\partial^2 \phi_x^h}{\partial x^2} + 2B_{16}\frac{\partial^2 \phi_x^h}{\partial x\partial y} + B_{66}\frac{\partial^2 \phi_x^h}{\partial y^2} + B_{12}\frac{\partial^2 \phi_y^h}{\partial x\partial y} + B_{16}\frac{\partial^2 \phi_y^h}{\partial x^2} + B_{26}\frac{\partial^2 \phi_y^h}{\partial y^2} + B_{66}\frac{\partial^2 \phi_y^h}{\partial x\partial y} \tag{16a}$$

$$R_v^\Omega = A_{16}\frac{\partial^2 u_0^h}{\partial x^2} + A_{66}\frac{\partial^2 u_0^h}{\partial x\partial y} + A_{12}\frac{\partial^2 u_0^h}{\partial x\partial y} + A_{26}\frac{\partial^2 u_0^h}{\partial y^2} + A_{26}\frac{\partial^2 v_0^h}{\partial x\partial y} + A_{66}\frac{\partial^2 v_0^h}{\partial x^2} + A_{22}\frac{\partial^2 v_0^h}{\partial y^2} + A_{26}\frac{\partial^2 v_0^h}{\partial x\partial y} + B_{16}\frac{\partial^2 \phi_x^h}{\partial x^2} + B_{66}\frac{\partial^2 \phi_x^h}{\partial x\partial y} + B_{12}\frac{\partial^2 \phi_x^h}{\partial y\partial x} + B_{26}\frac{\partial^2 \phi_y^h}{\partial y^2} + B_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y} + B_{66}\frac{\partial^2 \phi_y^h}{\partial x^2} + B_{22}\frac{\partial^2 \phi_y^h}{\partial y^2} + B_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y} \tag{16b}$$

$$R_w^\Omega = KA_{55}\frac{\partial \phi_x^h}{\partial x} + KA_{45}\frac{\partial \phi_x^h}{\partial y} + KA_{45}\frac{\partial \phi_y^h}{\partial x} + KA_{44}\frac{\partial \phi_y^h}{\partial y} + KA_{55}\frac{\partial^2 w_0^h}{\partial x^2} + 2KA_{45}\frac{\partial^2 w_0^h}{\partial x\partial y} + KA_{44}\frac{\partial^2 w_0^h}{\partial y^2} + f \tag{16c}$$

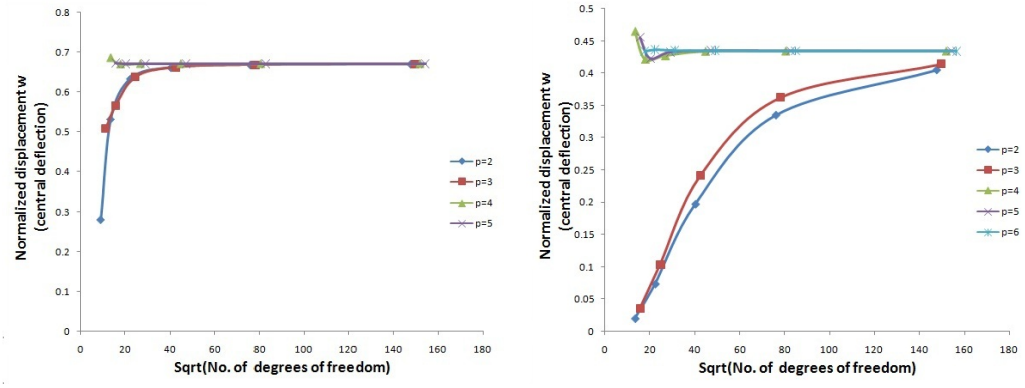
$$R_{\phi_x}^\Omega = B_{11}\frac{\partial^2 u_0^h}{\partial x^2} + 2B_{16}\frac{\partial^2 u_0^h}{\partial x\partial y} + B_{66}\frac{\partial^2 u_0^h}{\partial y^2} + B_{12}\frac{\partial^2 v_0^h}{\partial x\partial y} + B_{16}\frac{\partial^2 v_0^h}{\partial x^2} + B_{26}\frac{\partial^2 v_0^h}{\partial y^2} + B_{66}\frac{\partial^2 v_0^h}{\partial x\partial y} + D_{11}\frac{\partial^2 \phi_x^h}{\partial x^2} + 2D_{16}\frac{\partial^2 \phi_x^h}{\partial x\partial y} + D_{66}\frac{\partial^2 \phi_x^h}{\partial y^2} + D_{12}\frac{\partial^2 \phi_y^h}{\partial x\partial y} + D_{16}\frac{\partial^2 \phi_y^h}{\partial x^2} + D_{26}\frac{\partial^2 \phi_y^h}{\partial y^2} + D_{66}\frac{\partial^2 \phi_y^h}{\partial x\partial y} - KA_{55}\left(\frac{\partial w_0^h}{\partial x} + \phi_x^h\right) - KA_{45}\left(\frac{\partial w_0^h}{\partial y} + \phi_y^h\right) \tag{16d}$$

$$R_{\phi_y}^\Omega = B_{16}\frac{\partial^2 u_0^h}{\partial x^2} + B_{66}\frac{\partial^2 u_0^h}{\partial x\partial y} + B_{12}\frac{\partial^2 u_0^h}{\partial x\partial y} + B_{26}\frac{\partial^2 u_0^h}{\partial y^2} + B_{26}\frac{\partial^2 v_0^h}{\partial x\partial y} + B_{66}\frac{\partial^2 v_0^h}{\partial x^2} + B_{22}\frac{\partial^2 v_0^h}{\partial y^2} + B_{26}\frac{\partial^2 v_0^h}{\partial x\partial y} + D_{16}\frac{\partial^2 \phi_x^h}{\partial x^2} + D_{66}\frac{\partial^2 \phi_x^h}{\partial x\partial y} + D_{12}\frac{\partial^2 \phi_x^h}{\partial y\partial x} + D_{26}\frac{\partial^2 \phi_x^h}{\partial y^2} + D_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y} + D_{66}\frac{\partial^2 \phi_y^h}{\partial x^2} + D_{22}\frac{\partial^2 \phi_y^h}{\partial y^2} + D_{26}\frac{\partial^2 \phi_y^h}{\partial x\partial y} - KA_{45}\left(\frac{\partial w_0^h}{\partial x} + \phi_x^h\right) - KA_{44}\left(\frac{\partial w_0^h}{\partial y} + \phi_y^h\right) \tag{16e}$$

Similarly, residuals of the equilibrium equations valid at Neumann boundaries R_u^Γ, R_v^Γ, R_w^Γ, $R_{\phi_x}^\Gamma$ & $R_{\phi_y}^\Gamma$ are obtained by substituting eq. (14) into eqs. (10). The order of polynomial and knot vector are the same for the NURBS functions specified in eqs. (14). The collocation scheme adopted is

$$\begin{aligned} &R_u^\Omega = 0 \quad R_v^\Omega = 0 \quad R_w^\Omega = 0 \quad R_{\phi_x}^\Omega = 0 \quad R_{\phi_y}^\Omega = 0 \quad on\ S\left(\left\{\hat{\xi}\right\}^\Omega\right) \\ &R_u^\Gamma = 0 \quad R_v^\Gamma = 0 \quad R_w^\Gamma = 0 \quad R_{\phi_x}^\Gamma = 0 \quad R_{\phi_y}^\Gamma = 0 \quad on\ S\left(\left\{\hat{\xi}\right\}^\Gamma\right) \end{aligned} \tag{17}$$

where $S\left(\left\{\hat{\xi}\right\}^\Gamma\right)$ is the mapping of Greville points from the parametric space to physical space.



(a) h = a/10 (b) h = a/100

Figure 1: Lateral displacement w^* versus $\sqrt{N_f}$

4 NUMERICAL RESULTS

4.1 Bending of all round simply supported square cross ply laminated plates

Consider an all round simply supported rectangular laminated plate. The laminate consists of four plies (0/90/90/0) of equal thickness. Let h be the total thickness of the plate. The material properties are

$$E_1 = 25E_2, \quad G_{12} = G_{13} = 0.5E_2, \quad G_{23} = 0.2E_2, \quad \nu_{12} = 0.25, \& \qquad K = 5/6 \tag{18}$$

where the origin of the co-ordinate system is taken at a corner of the plate, $0 \leq x \leq a \ \& \ 0 \leq y \leq b$. The lateral displacement and the stresses evaluated are non - dimensionalized using the following relationship,

$$\begin{aligned} w^* &= w_0(\frac{a}{2}, \frac{b}{2})\frac{E_2 h^3}{b^4 q_0}, & \sigma^*_{xx} &= \sigma_{xx}(\frac{a}{2}, \frac{b}{2}, \frac{h}{2})\frac{h^2}{b^2 q_0} \\ \sigma^*_{yy} &= \sigma_{yy}(\frac{a}{2}, \frac{b}{2}, \frac{h}{4})\frac{h^2}{b^2 q_0}, & \sigma^*_{xy} &= \sigma_{xy}(a, b, -\frac{h}{2})\frac{h^2}{b^2 q_0} \\ \sigma^*_{xz} &= \sigma_{xz}(a, \frac{b}{2}, k = 1, 3)\frac{h}{b q_0} \end{aligned} \tag{19}$$

The results obtained are listed in Table 1 and compared with methods existing in literature. Figure 1 shows the plot of normalized central displacement w^* versus N_f (square root of the number of degrees of freedom) for two plate width to thickness ratios of 10 & 100. Convergence of normalized central displacement w^* can be observed.

4.2 Bending of all round clamped square sandwich plates

Consider an all round clamped square sandwich plate. The sandwich plate consists of a central core, embedded with face sheets on top and bottom. The face sheets are

Table 1: Simply supported $[0/90/90/0]$ square laminated plate under sinusoidal load.

a/h	Method	w^*	σ^*_{xx}	σ^*_{yy}	σ^*_{xy}	τ^*_{xz}
100	HSDT [31]	0.4343	0.5507	0.2769	0.0217	0.2948
	HSDT [32]	0.4343	0.5387	0.2708	0.0213	0.2897
	FSDT [33]	0.4337	0.5382	0.2705	0.0213	0.178
	Elasticity ‘ [34]	0.4347	0.539	0.271	0.0214	0.339
	HSDT [35]	0.4365	0.5413	0.3359	0.0215	0.4106
	Layer-wise [36]	0.4374	0.542	0.2697	0.0216	0.3232
	RBF-PS [37]	0.432	0.5387	0.2697	0.0213	0.3154
	Wavelets [38]	0.4335	0.5381	0.2704	0.0213	0.339
	IGA- galerkin [39]	0.4354	0.5376	0.2702	0.0213	0.1907
	Quadratic	0.3339	0.4138	0.2079	0.182	1.5795
	Cubic	0.3622	0.4493	0.2255	0.0206	1.0969
	Quartic	0.4335	0.5379	0.27	0.0235	0.9867
	Quintic	0.4346	0.5392	0.2707	0.0223	0.8195
10	HSDT [31]	0.7149	0.5589	0.3974	0.0273	0.2697
	HSDT [32]	0.7147	0.5456	0.3888	0.0268	0.264
	FSDT [33]	0.6628	0.4989	0.3615	0.0241	0.1667
	Elasticity [34]	0.743	0.559	0.403	0.0276	0.301
	HSDT [35]	0.7153	0.5466	0.4383	0.0267	0.3347
	Layer-wise [36]	0.7309	0.5496	0.3956	0.0273	0.2888
	RBF-PS [37]	0.7204	0.5609	0.3911	0.0273	0.2843
	Wavelets [38]	0.6627	0.4989	0.3614	0.0241	0.3181
	IGA- galerkin [39]	0.6654	0.4983	0.361	0.0242	0.1669
	Quadratic	0.6670	0.5022	0.3623	0.0202	0.1742
	Cubic	0.6688	0.5038	0.3636	0.0204	0.1741
	Quartic	0.6693	0.5043	0.3639	0.0203	0.1744
	Quintic	0.6709	0.5055	0.3648	0.0203	0.1744
20	HSDT [31]	0.5061	0.5523	0.311	0.0233	0.2883
	HSDT [32]	0.506	0.5393	0.3043	0.0228	0.2825
	FSDT [33]	0.4912	0.5273	0.2957	0.0221	0.1749
	Elasticity [34]	0.517	0.543	0.309	0.023	0.328
	HSDT [35]	0.507	0.5405	0.3648	0.0228	0.3818
	Layer-wise [36]	0.5121	0.5417	0.3056	0.023	0.3248
	RBF-PS [37]	0.5078	0.5436	0.3052	0.023	0.3066
	Wavelets [38]	0.4912	0.5273	0.2956	0.0221	0.3332
	IGA- galerkin [39]	0.4931	0.5268	0.2953	0.0221	0.1755
	Quadratic	0.4882	0.5232	0.2923	0.0203	0.1818
	Cubic	0.4907	0.5264	0.2942	0.0204	0.1813
	Quartic	0.4940	0.5299	0.2962	0.0205	0.1827

Table 2: All round clamped square sandwich plate under uniform load.

a/h	Method	w^*	σ^*_{xx}	σ^*_{yy}	σ^*_{xy}	τ^*_{xz}
100	Reddy J N [40]	0.2785	0.5347	0.0094	0.003	0.24
	Quadratic	0.2350	0.4489	0.009	0.0114	0.3225
	Cubic	0.2482	0.4749	0.0092	0.0123	0.3098
	Quartic	0.2749	0.5264	0.0114	0.0151	0.3368
50	Reddy J N [40]	0.3111	0.5356	0.0108	0.0039	0.2406
	Quadratic	0.3044	0.5044	0.0109	0.0139	0.3342
	cubic	0.3085	0.5113	0.0112	0.0143	0.3299
	Quartic	0.3162	0.5241	0.0120	0.0152	0.3365
10	Reddy J N [40]	1.2654	0.5018	0.0550	0.0120	0.2318
	Quadratic	1.3569	0.4153	0.0361	0.0281	0.2893
	cubic	1.3591	0.4154	0.0362	0.0284	0.2894
	Quartic	1.3619	0.4162	0.0362	0.0285	0.2899

assumed to be orthotropic and the core is assumed to be transversely isotropic. Let h be the total thickness of the plate. Thickness of the core is 0.8 times the total thickness while the face sheets are one - tenth the total thickness of the sandwich plate. Material properties of the core are given below

$$E_1 = 25E_2, \quad E_2 = 10^6 \quad G_{12} = G_{13} = 0.5E_2, \quad G_{23} = 0.2E_2, \quad \nu_{12} = 0.25, \& \qquad K = 5/6 \tag{20}$$

Material properties of the face sheets are provided below

$$E_1 = E_2 = 10^6 \quad G_{13} = G_{23} = 0.06 * 10^6 psi, \quad G_{12} = 0.016 * 10^6 psi, \quad \nu_{12} = 0.2 \tag{21}$$

The transverse load acting on the rectangular plate is given by $q(x, y) = q_0$, where the origin of the co-ordinate system is taken at a corner of the plate, $0 \leq x \leq a \,\&\, 0 \leq y \leq b$. The lateral displacement and the stresses evaluated are non - dimensionalized using the following relations.

$$\begin{aligned} w^* &= w_0(\frac{a}{2}, \frac{b}{2})\frac{E_2 h^3}{b^4 q_0}, & \sigma^*_{xx} &= \sigma_{xx}(\frac{a}{2}, \frac{b}{2}, \frac{h}{2})\frac{h^2}{b^2 q_0} \\ \sigma^*_{yy} &= \sigma_{yy}(\frac{a}{2}, \frac{b}{2}, \frac{h}{2})\frac{h^2}{b^2 q_0}, & \sigma^*_{xy} &= \sigma_{xy}(a, b, -\frac{h}{2})\frac{h^2}{b^2 q_0} \\ \sigma^*_{xz} &= \sigma_{xz}(\frac{a}{2}, b, 0)\frac{h}{b q_0} \end{aligned} \tag{22}$$

The results obtained are listed in Table 2 and compared with solution existing in literature.

5 CONCLUSIONS

In this paper, analysis of laminated composite plates has been performed using Isogeometric collocation method. The method has been assessed for both laminated plate and sandwich plates under all round simply supported and all round clamped conditions respectively. The accuracy obtained by this method is comparable with the methods existing in literature. Shear locking was observed under the thin plate limit when lower order approximations were adopted. Mixed formulation may be considered in order to alleviate shear locking. Further research direction would be to study free vibration analysis, buckling analysis and extension of present formulation under large deformations.

References

[1] T. J. Hughes, J. A. Cottrell, and Y. Bazilevs, "Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement," *Computer methods in applied mechanics and engineering*, vol. 194, no. 39, pp. 4135–4195, 2005.

[2] J. Caseiro, R. Valente, A. Reali, J. Kiendl, F. Auricchio, and R. A. De Sousa, "On the assumed natural strain method to alleviate locking in solid-shell nurbs-based finite elements," *Computational Mechanics*, vol. 53, no. 6, pp. 1341–1353, 2014.

[3] J. Cottrell, T. Hughes, and A. Reali, "Studies of refinement and continuity in isogeometric structural analysis," *Computer methods in applied mechanics and engineering*, vol. 196, no. 41, pp. 4160–4183, 2007.

[4] J. A. Cottrell, A. Reali, Y. Bazilevs, and T. J. Hughes, "Isogeometric analysis of structural vibrations," *Computer methods in applied mechanics and engineering*, vol. 195, no. 41, pp. 5257–5296, 2006.

[5] C. De Falco, A. Reali, and R. Vázquez, "Geopdes: a research tool for isogeometric analysis of pdes," *Advances in Engineering Software*, vol. 42, no. 12, pp. 1020–1034, 2011.

[6] T. Elguedj, Y. Bazilevs, V. M. Calo, and T. J. Hughes, "B and f projection methods for nearly incompressible linear and non-linear elasticity and plasticity using higher-order nurbs elements," *Computer methods in applied mechanics and engineering*, vol. 197, no. 33, pp. 2732–2762, 2008.

[7] T. J. Hughes, J. A. Evans, and A. Reali, "Finite element and nurbs approximations of eigenvalue, boundary-value, and initial-value problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 272, pp. 290–320, 2014.

[8] S. Morganti, F. Auricchio, D. Benson, F. Gambarin, S. Hartmann, T. Hughes, and A. Reali, "Patient-specific isogeometric structural analysis of aortic valve closure,"

Computer Methods in Applied Mechanics and Engineering, vol. 284, pp. 508–520, 2015.

[9] S. Lipton, J. A. Evans, Y. Bazilevs, T. Elguedj, and T. J. Hughes, "Robustness of isogeometric structural discretizations under severe mesh distortion," *Computer Methods in Applied Mechanics and Engineering*, vol. 199, no. 5, pp. 357–373, 2010.

[10] A. Reali, "An isogeometric analysis approach for the study of structural vibrations," *Journal of Earthquake Engineering*, vol. 10, no. spec01, pp. 1–30, 2006.

[11] I. Akkerman, Y. Bazilevs, V. Calo, T. Hughes, and S. Hulshoff, "The role of continuity in residual-based variational multiscale modeling of turbulence," *Computational Mechanics*, vol. 41, no. 3, pp. 371–378, 2008.

[12] Y. Bazilevs, V. Calo, J. Cottrell, T. Hughes, A. Reali, and G. Scovazzi, "Variational multiscale residual-based turbulence modeling for large eddy simulation of incompressible flows," *Computer Methods in Applied Mechanics and Engineering*, vol. 197, no. 1, pp. 173–201, 2007.

[13] Y. Bazilevs and T. Hughes, "Nurbs-based isogeometric analysis for the computation of flows about rotating components," *Computational Mechanics*, vol. 43, no. 1, pp. 143–150, 2008.

[14] A. Buffa, C. De Falco, and G. Sangalli, "Isogeometric analysis: Stable elements for the 2d stokes equation," *International Journal for Numerical Methods in Fluids*, vol. 65, no. 11-12, pp. 1407–1422, 2011.

[15] J. Liu, H. Gomez, J. A. Evans, T. J. Hughes, and C. M. Landis, "Functional entropy variables: a new methodology for deriving thermodynamically consistent algorithms for complex fluids, with particular reference to the isothermal navier–stokes–korteweg equations," *Journal of Computational Physics*, vol. 248, pp. 47–86, 2013.

[16] H. Gomez, T. J. Hughes, X. Nogueira, and V. M. Calo, "Isogeometric analysis of the isothermal navier–stokes–korteweg equations," *Computer Methods in Applied Mechanics and Engineering*, vol. 199, no. 25, pp. 1828–1840, 2010.

[17] H. Gomez and T. J. Hughes, "Provably unconditionally stable, second-order time-accurate, mixed variational methods for phase-field models," *Journal of Computational Physics*, vol. 230, no. 13, pp. 5310–5327, 2011.

[18] H. Gómez, V. M. Calo, Y. Bazilevs, and T. J. Hughes, "Isogeometric analysis of the cahn–hilliard phase-field model," *Computer methods in applied mechanics and engineering*, vol. 197, no. 49, pp. 4333–4352, 2008.

[19] Y. Bazilevs, V. Calo, Y. Zhang, and T. J. Hughes, "Isogeometric fluid–structure interaction analysis with applications to arterial blood flow," *Computational Mechanics*, vol. 38, no. 4-5, pp. 310–322, 2006.

[20] Y. Bazilevs, V. M. Calo, T. J. Hughes, and Y. Zhang, "Isogeometric fluid-structure interaction: theory, algorithms, and computations," *Computational mechanics*, vol. 43, no. 1, pp. 3–37, 2008.

[21] Y. Zhang, Y. Bazilevs, S. Goswami, C. L. Bajaj, and T. J. Hughes, "Patient-specific vascular nurbs modeling for isogeometric analysis of blood flow," *Computer methods in applied mechanics and engineering*, vol. 196, no. 29, pp. 2943–2959, 2007.

[22] F. Auricchio, L. B. Da Veiga, T. Hughes, A. Reali, and G. Sangalli, "Isogeometric collocation methods," *Mathematical Models and Methods in Applied Sciences*, vol. 20, no. 11, pp. 2075–2107, 2010.

[23] D. Schillinger, J. A. Evans, A. Reali, M. A. Scott, and T. J. Hughes, "Isogeometric collocation: Cost comparison with galerkin methods and extension to adaptive hierarchical nurbs discretizations," *Computer Methods in Applied Mechanics and Engineering*, vol. 267, pp. 170–232, 2013.

[24] F. Auricchio, L. B. da Veiga, T. Hughes, A. Reali, and G. Sangalli, "Isogeometric collocation for elastostatics and explicit dynamics," *Computer Methods in Applied Mechanics and Engineering*, vol. 249, pp. 2–14, 2012.

[25] H. Gomez, A. Reali, and G. Sangalli, "Accurate, efficient, and (iso) geometrically flexible collocation methods for phase-field models," *Journal of Computational Physics*, vol. 262, pp. 153–171, 2014.

[26] L. De Lorenzis, J. Evans, T. Hughes, and A. Reali, "Isogeometric collocation: Neumann boundary conditions and contact," *Computer Methods in Applied Mechanics and Engineering*, vol. 284, pp. 21–54, 2015.

[27] A. Reali and H. Gomez, "An isogeometric collocation approach for bernoulli–euler beams and kirchhoff plates," *Computer Methods in Applied Mechanics and Engineering*, vol. 284, pp. 623–636, 2015.

[28] L. B. da Veiga, C. Lovadina, and A. Reali, "Avoiding shear locking for the timoshenko beam problem via isogeometric collocation methods," *Computer Methods in Applied Mechanics and Engineering*, vol. 241, pp. 38–51, 2012.

[29] F. Auricchio, L. B. da Veiga, J. Kiendl, C. Lovadina, and A. Reali, "Locking-free isogeometric collocation methods for spatial timoshenko rods," *Computer Methods in Applied Mechanics and Engineering*, vol. 263, pp. 113–126, 2013.

[30] J. Kiendl, F. Auricchio, L. B. da Veiga, C. Lovadina, and A. Reali, "Isogeometric collocation methods for the reissner–mindlin plate problem," *Computer Methods in Applied Mechanics and Engineering*, vol. 284, pp. 489–507, 2015.

[31] J. N. Reddy, "A simple higher-order theory for laminated composite plates," *Journal of applied mechanics*, vol. 51, no. 4, pp. 745–752, 1984.

[32] G. Akhras, M. Cheung, and W. Li, "Finite strip analysis of anisotropic laminated composite plates using higher-order shear deformation theory," *Computers & Structures*, vol. 52, no. 3, pp. 471–477, 1994.

[33] G. Akhras, M. Cheung, and W. Li, "Static and vibration analysis of anisotropic composite laminates by finite strip method," *International Journal of Solids and Structures*, vol. 30, no. 22, pp. 3129–3137, 1993.

[34] N. Pagano, "Exact solutions for rectangular bidirectional composites and sandwich plates," *Journal of composite materials*, vol. 4, no. 1, pp. 20–34, 1970.

[35] A. Ferreira, C. Roque, and P. Martins, "Analysis of composite plates using higher-order shear deformation theory and a finite point formulation based on the multiquadric radial basis function method," *Composites Part B: Engineering*, vol. 34, no. 7, pp. 627–636, 2003.

[36] A. Ferreira, "Analysis of composite plates using a layerwise theory and multiquadrics discretization," *Mechanics of Advanced Materials and Structures*, vol. 12, no. 2, pp. 99–112, 2005.

[37] A. Ferreira and G. Fasshauer, "Computation of natural frequencies of shear deformable beams and plates by an rbf-pseudospectral method," *Computer Methods in Applied Mechanics and Engineering*, vol. 196, no. 1, pp. 134–146, 2006.

[38] A. Ferreira, L. M. Castro, and S. Bertoluzza, "A high order collocation method for the static and vibration analysis of composite plates using a first-order theory," *Composite Structures*, vol. 89, no. 3, pp. 424–432, 2009.

[39] C. H. Thai, H. Nguyen-Xuan, N. Nguyen-Thanh, T.-H. Le, T. Nguyen-Thoi, and T. Rabczuk, "Static, free vibration, and buckling analysis of laminated composite reissner–mindlin plates using nurbs-based isogeometric approach," *International Journal for Numerical Methods in Engineering*, vol. 91, no. 6, pp. 571–603, 2012.

[40] J. N. Reddy, *Mechanics of laminated composite plates and shells: theory and analysis*. CRC press, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MODELLING STIFFENED LIGHTWEIGHT STRUCTURES WITH ISOGEOMETRIC ANALYSIS VIA MORTAR METHODS

Malte Woidt[1], **Kay Sommerwerk**[1], **Matthias C. Haupt**[1], **Peter Horst**[1]

[1]Institute of Aircraft Design and Lightweight Structures
Technische Universität Braunschweig
Hermann.Blenk-Str. 35, 38108 Braunschweig, Germany
e-mail: m.woidt,k.sommerwerk,m.haupt,p.horst@tu-braunschweig.de

Keywords: isogeometric analysis, mortar methods, lagrange multipliers, stiffened structures

Abstract. *A method to model stiffened panels with NURBS-based isogeometric analysis is presented. Stiffeners and panel are modeled as separate patches and coupled by Mortar methods. The stiffeners and panel meshes can be non-conforming. Therefore modeling of the geometry is simple even with the NURB's limitation to structured meshes. The feasibility of the method is shown for static and modal analyses. The method is validated against results obtained from standard finite element approaches with conforming meshes for a panel with one stiffener and a panel with multiple stiffeners. For all cases good accordance to the FEM results is observed with a difference of less than* 1%.

1 INTRODUCTION

Isogeomeric analysis (IGA) is a method intended to bridge the gap between finite element analysis and computer aided design (CAD). It uses the methods that are used in CAD to describe geometries for numerical analysis with the finite element method (FEM). The original formulation from Hughes et. al. [1] uses Non-Uniform Rational B-Splines (NURBS) basis functions as they are most commonly used in CAD. Lately also T-Splines, a generalization of NURBS, are used as they enable local h-refinement for IGA. The Idea behind IGA is that standard FEM approximate geometries with Lagrange polynomials where an exact geometry description is already available by CAD.

Compared to standard FEM the application of NURBS basis functions has several advantages. Due to the mathematical properties of NURBS higher polynomial orders p can be applied what may lead to higher accuracy per degree of freedom and higher convergence rates [2]. Furthermore higher inter-element continuities can achieved. While Lagrange polynomials are limited to C^0 inter-element continuity NURBS can be constructed with up to C^{p-1} continuity. This results in smooth surface representation and enables the usage of governing equations, that need higher order derivatives. Among other fields of application this property is appealing for analysis of thin-walled structures as many shell theories include second order derivatives [3].

This potential cannot be fully used as NURBS are limited to structured meshes. Many lightweight structures are stiffened structures and therefore consists of a base panel with several stiffeners connected to it which cannot be modeled as one structured mesh. One possibility to model this kind of structures is to model stiffener and base panel as separate domains and connect them afterwards. In the simplest case the common border of the two parts are conforming and therefore can simply be connected by sharing control points. This may be a possible solution for classical FEM but if the limitation to structured meshes is taken into account this may lead to massive mesh distortion in the base plate or the need of splitting the base-plate into many parts. Therefore coupling methods are needed that enables the coupling of the stiffener's foot onto the base plate in an arbitrary position. Mortar methods are frequently used for coupling of non-conforming edges in IGA [4]. They are also used in contact formulations [5]. Therefore it is not only possible to couple two boundary edges of NURBS patches but also to couple displacements on a shell surface e.g. example in problems involving contact of shells. Depending on the structural behavior that should be achieved only transversal or transversal and rotational displacements can be coupled.

The aim of the paper is to use coupling methods for connecting a stiffener on a base panel without having to adapt the baseplate mesh to the stiffener position or splitting it into sub-regions. This is desirable because each coupling of subregions with Lagrange multipliers introduce extra degrees of freedom and breaks the high inter-element continuity locally. At least the here proposed methods can only enforce a C^0 continuity of the displacement field over the coupling edge. This is not problematic in this context of Mindlin-Reissner which are only requiring C^0 continuous transversal and rotational displacements.

This gives not only great flexibility in modeling structures with many stiffeners but also enables the movement of stiffeners without having to recompute big parts of the system. Therefore the proposed method is especially interesting for optimization tasks. The feasibility of the method is demonstrated by comparing IGA results for stiffened panels with results obtained by classic FEM-models with conforming meshes.

2 Fundamentals

The examples presented in this paper are based on standard Mindlin-Reissner shell kinematic as also used in FEM. These type of shell theory represents the geometry as its mid-surface and a thickness parameter t, describing the thickness of the shell normal to the mid-surface. It accounts for the mid-surface curvature. Mindlin-Reissner shells often have 6 displacement variables. Three of them describing the mid-surface displacement in physical space u_x, u_y, u_z and three describing the rotation of a material fiber that was normal to the mid-surface in the undeformed state $\theta_x, \theta_y, \theta_z$ as described in [3]. As the three rotation parameters correspond to the Euler angels, which have the well-known non-uniqueness problem a drilling penalty method is used. In difference to shell elements used in classical FEM the mid-surface normal, also denoted as the shell director, is not interpolated from nodal values but directly computed what is possible due to the higher inter-element continuity. Even there exist more evolved shell models for IGA as locking-free 5-Parameter Mindlin-Reissner shells in this paper 6-parameter shell theory is used because the presented examples contain geometries with C^0 continuous locations which are complicated to handle with 5-Parameter shells. It should be noted that there are also more elaborated 5-parameter theories that can handle such regions (e.g. [7])

2.1 B-Splines and NURBS

This work uses NURBS based geometry description which is based on B-Splines. Following the isoparametric approach these are also used for interpolation of displacements. As the examples shown in this paper are based on shell theory the geometries consist of surfaces and curves in three dimensional space. B-splines describe geometries in a parametric way. A curve is therefore defined over one parametric coordinate ξ as the product of geometric control points $\mathbf{P}_i$ and basis functions $N_{i,p}$ of order p as

$$\mathbf{C}(\xi) = \sum_{i=1}^{n} N_{i,p}(\xi)\mathbf{P}_i \quad . \tag{1}$$

These n basis functions form a basis of the space of all piecewise polynomials of degree p on an interval defined by a knot vector

$$\Xi = [\xi_1, \xi_2, \ldots, \xi_{n+p+1}], \quad \xi_i \leq \xi_{i+1} \quad . \tag{2}$$

The positions ξ_i are called knots where each knot may appear several times consecutively. The knots denote positions in parametric space where two polynomials are connected in a C^{p-m} continuous way where m is called multiplicity and denotes how often a knot appears in the knot vector. Therefore continuity of the B-Spline curve and of the B-Spline basis can be controlled by the knot vector. The non-zero knot-spans are something close to what is referred to as elements in classic FEM. Therefore they are often called elements even though there exist some differences in principle. The basis functions can be computed recursively from the knot vector with the Cox-de Boor formula [1].
A B-Spline surface is simply constructed by the tensorial product of two B-Spline curves

$$\mathbf{S}(\xi, \eta) = \sum_i \sum_j N_{i,p}(\xi)\, M_{j,q}(\eta)\, \mathbf{P}_{ij} \quad . \tag{3}$$

Both curves may have different orders p and q. The control points $\mathbf{P}_{ij}$ can be understood as a structured control mesh in the two dimensional case. The Surface spans over a rectangular parametric region with the parameters $[\xi, \eta]$. Therefore B-Splines and NURBS are limited to structured control meshes and regions that can be parametrized by a rectangular parameter space.

The limitation to structured control meshes can be overcome using T-Splines what enables the possibility of local refinement [8] whereas the limitation to regions that can be parametrized by a rectangular parameter space cannot.
The NURBS basis functions are computed from the B-Spline basis functions as

$$R_{i,p}(\xi) = \frac{N_{i,p}(\xi) w_i}{\sum_{k=1}^{n} N_{k,p}(\xi) w_k} \quad . \tag{4}$$

For each geometric control point an additional parameter w_i is introduced called the weight of the control point.
NURBS are a generalization of B-Splines. They share many properties as the inter-element continuity of basis functions, partition of unity and non-negativity. In contrast to B-splines they allow the exact representation of conic sections (e.g. circles) and therefore the geometric continuity can be higher than the continuity of the basis functions. The geometric region that is described by one set of NURBS basis functions and control points is called a NURBS-patch.

2.2 Multipatch coupling with Lagrange multipliers

The Lagrange multiplier approach introduces boundary conditions as a weak formulation so that they do not need to be fulfilled implicitly by the base functions. The approach is e.g. a useful method to couple two sub-domains with incompatible displacement discretization on an common boundary. It allows the unavoidable deviations between the two displacement fields minimizing the error resulting from the incompatible displacement fields in means of the energy norm [6]. To couple the displacements $\mathbf{u}^{(1)}$ of sub-domain Ω_1 with the displacements $\mathbf{u}^{(2)}$ of sub-domain Ω_2 at the common boundary Γ_λ a new vector field λ is introduced on this boundary. The coupling conditions are then incorporated into the weak formulation as an additional term Π_λ that is added to the governing equations $\Pi^{(1)}$ and $\Pi^{(2)}$ of the two sub-domains

$$\Pi = \Pi^{(1)} + \Pi^{(2)} + \Pi_\lambda \qquad \Pi_\lambda = \int_{\Gamma_\lambda} (\mathbf{u}^{(1)} - \mathbf{u}^{(2)}) \cdot \lambda d\Gamma \tag{5}$$

For equations of elasticity λ is the traction field that needs to be applied on the border to couple the two subregions. Basically there are two different methods of discretizing Π_λ. The first one is to compute λ from the stress field of the subregions and therefore indirectly from the displacement field. This approach is called Nitsche-Method and has the advantage that no additional degrees of freedom (DOF) are introduced [9]. The drawback of this method is that the formulation heavily depends on the governing differential equations on the subregions. Therefore the more general method is used in this work where the Lagrange multipliers are interpolated by basis functions and add new DOF to the system. This leads to a saddle point problem and must be treated by appropriate solving algorithms. The advantage of this approach is that the governing equations of the sub-domains do not matter as long as they use the displacements as primary variables. As basis functions for the Lagrange multipliers the basis functions that interpolate the displacements on the boundary of one of the adjacent sub-domains are used. This choice is quite common even there may be more elaborate solutions available. For example dual Mortar methods, which use orthogonal basis functions to the ones of the coupled regions are subject to ongoing research as they greatly increase efficiency when solving the resulting system of linear equations. An open problem is the construction of such a dual-basis for higher order basis functions that do not compromise the overall convergence rate even there are promising approaches to this problem [5].

As the boundary edge of a stiffener is coupled to a panel the basis functions of the stiffener are used for the Lagrange multipliers λ. The stiffeners basis functions are chosen because the boundary that should be coupled is also a boundary of the parameter space. This makes the extraction of the basis functions easier. The Integral in eq. 5 is evaluated by a gauss type quadrature with the evaluation points defined in the parametric space of the stiffener. Here again the evaluation points lie on the boundary of the stiffeners parameter space and can be evaluated more easily. The corresponding positions on the panel are located by a projection algorithm that uses a gradient descent method. This evaluation method of the integral is used as stiffener and plate are refined uniformly and therefore both have similar element size. For refinements where this is not the case more complex evaluation methods have to be chosen in order not to bypass base functions on the more refined edge as described in [10].

3 Investigation of a simple stiffened panel

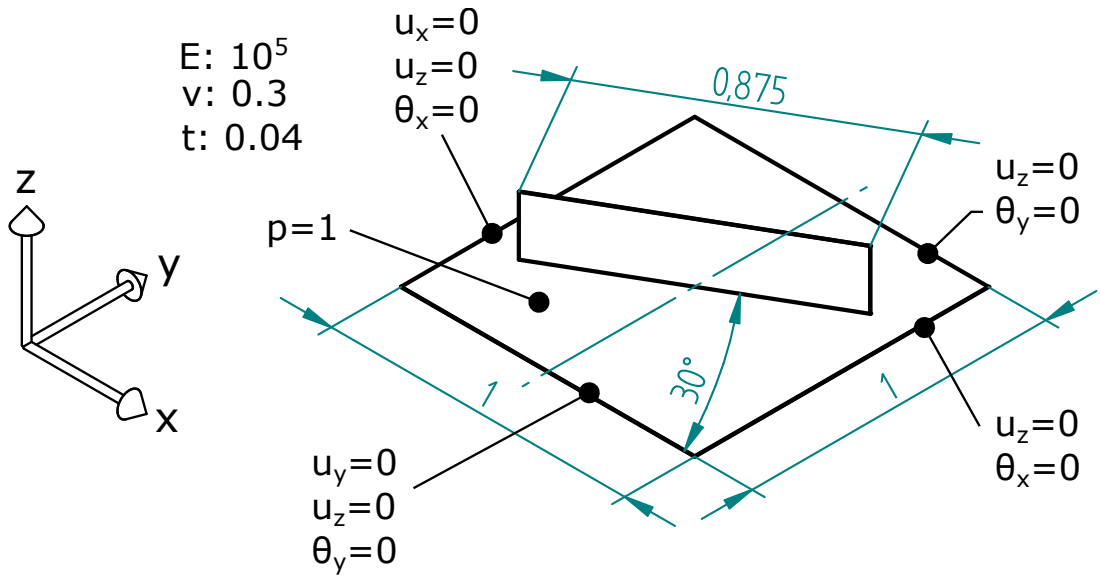


Figure 1: Model description of the test problem

To test the method a simple example is investigated where a rectangular stiffener is coupled to a square panel as shown in fig. 1. The panel is loaded by a pressure in positive z-Direction. The angle between the stiffener and the panel edge is $30°$ so that the foot of the stiffener is crossing elements of the panel at different locations. At the coarsest IGA-mesh panel and stiffener are modeled as one element each of degree $p = 1$. To investigate performance they are refined uniformly. The deformation and the von Mises stresses for one refinement step are shown in fig. 2.

To investigate the overall performance of the method the results are compared to a FEM-model with conforming meshes of quadratic shell elements. The model uses a structured mesh of $n_{elem} = 20 \times 87 = 1740$ elements for the stiffener and $n_{elem} = 12157$ elements for the base panel with a total of $n_{DOF} = 250848$ DOF. The error is measured as the relative difference of the total strain energy Δe.

$$\Delta e = \frac{0.5\mathbf{U}^{FE} \cdot \mathbf{F}^{FE} - 0.5\mathbf{U}^{IGA} \cdot \mathbf{F}^{IGA}}{0.5\mathbf{U}^{FE} \cdot \mathbf{F}^{FE}} \tag{6}$$

with $\mathbf{U}$ being the displacement vector and $\mathbf{F}$ the force vector of the discrete models.
To investigate how well the displacement coupling conditions are fulfilled the L_2-norm of the relative differences between the displacements of both subregions at the common boundary Γ_λ is used.

$$||e_U||_{L_2} = \int_{\Gamma_\lambda} \frac{|\mathbf{u}^{(1)} - \mathbf{u}^{(2)}|}{0.5(|\mathbf{u}^{(1)} + \mathbf{u}^{(2)}|)} d\Gamma \tag{7}$$

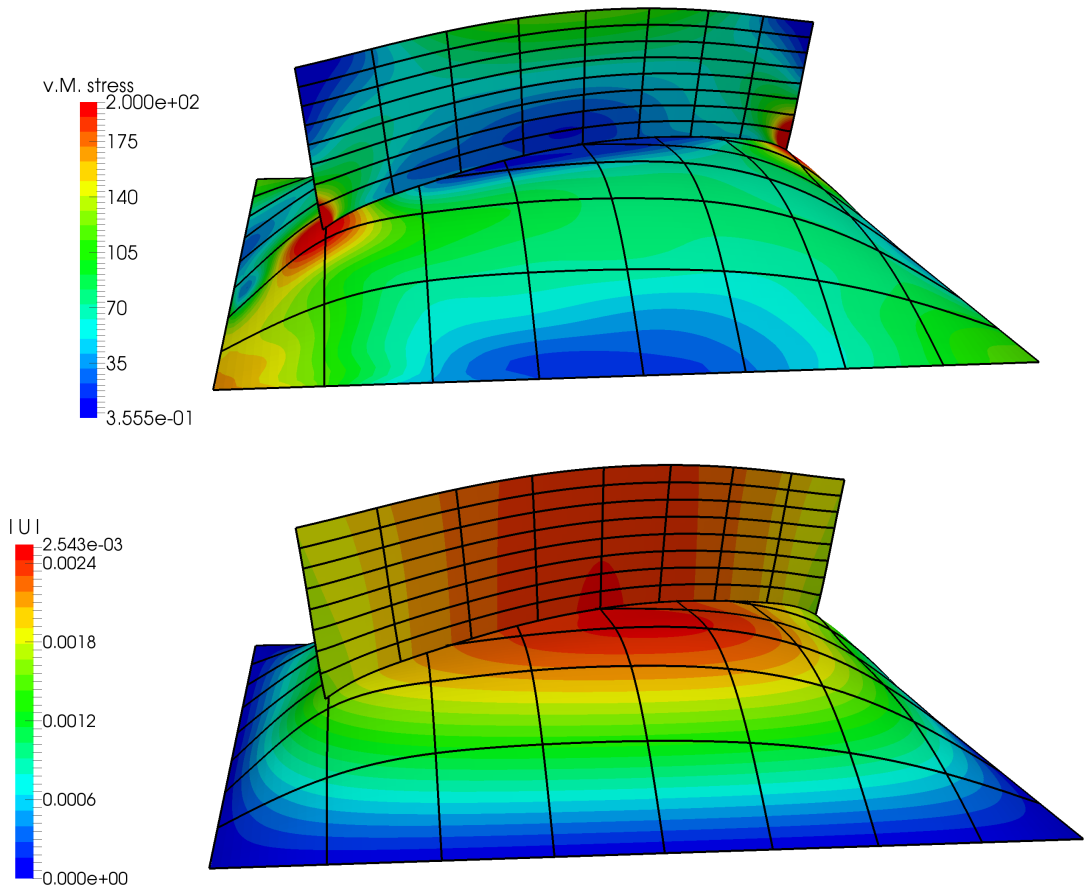


Figure 2: Von Mises stress and displacement in deformed state (warped by factor 70. Stresses clipped at 200). Panel and stiffener have $n_{elem} = 64$ each of order $p = 7$

The results are shown in fig. 3. The difference between the IGA and FEM solution by means of strain energy is smaller than 1% for the finest IGA-mesh. The results show that the convergence rate does not improve with higher approximation orders. Degrees of $p > 3$ do not improve the results. As seen in the stress plot the stiffener introduces two stress singularities at the end-points as these are reentrant corners. Furthermore the Lagrange multipliers introduce a line load to the panel that is modeled as a shell. This line load results in a discontinuity of the σ_{xz} and σ_{yz} stress components of the panel. The component $\sigma_{zz} = 0$ as the shell is based on a plane stress assumption. Therefore the tractions in z-direction are coupled into the transversal shear stresses resulting in steep gradients in the discretized model. As seen in fig. 2 this does not lead to oscillation phenomenas in the stress field even with high order basis functions but impairs the convergence speed at least locally around the stiffener foot. The displacement coupling condition can be fulfilled very accurate even in the end-points with the stress singularities.

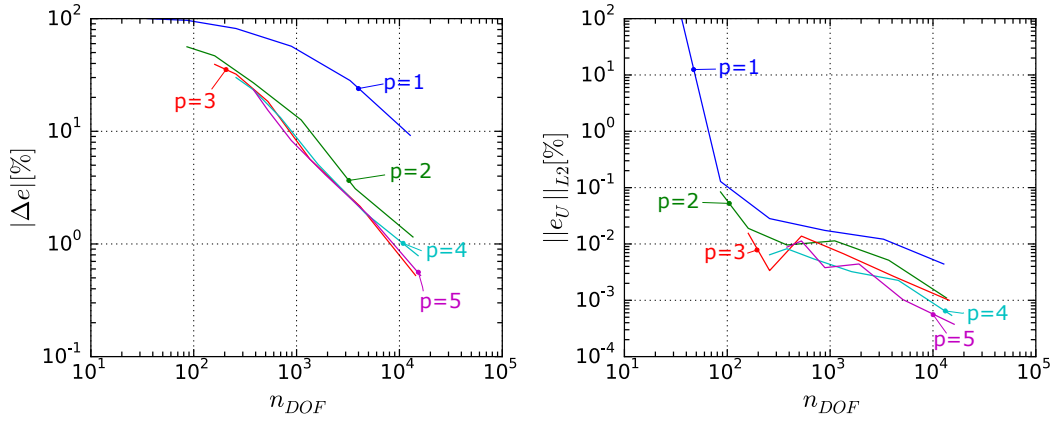


Figure 3: Convergence against FEM-solution and relative error of displacement coupling condition in L_2-norm

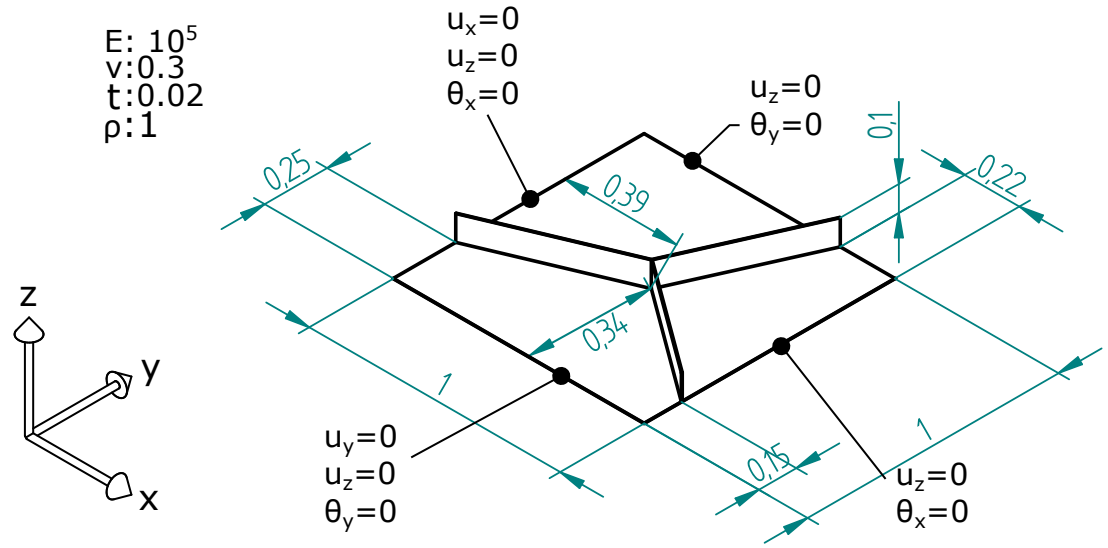


Figure 4: Geometry for test problem with multiple stiffeners

4 Results for multiple stiffeners

As a second example a panel with multiple stiffeners is investigated. The test geometry is depicted in fig. 4. It is modeled with 3 Patches. One for the base panel and two for the stiffeners where one stiffener has a kink. The two stiffeners are connected by Lagrange multipliers in the same way as the stiffeners and the base panel. The results are again validated against a FEM-model with conforming meshes by comparing the strain energy. The FE-model consists of quadratic shell elements with $n_{elem} = 1730$ elements for the stiffeners and $n_{elem} = 12127$ elements for the panel with a total of $n_{DOF} = 250848$ DOF. The base panel is loaded by a pressure of $p = 1$ for the static test case. Beside the strain energy of the static case the first 4 eigenfrequencies are compared where a density of $\rho = 1$ is assumed. The results are given in fig. 6. Fig. 7 shows the 4th eigenform obtained by the IGA method. All eigenfrequencies can be obtained with less than 1% difference to the FE-results. In the static and the modal analyses the higher orders basis functions tend to better accuracy with same number of DOF but again the convergence speed is limited by the stress singularities.

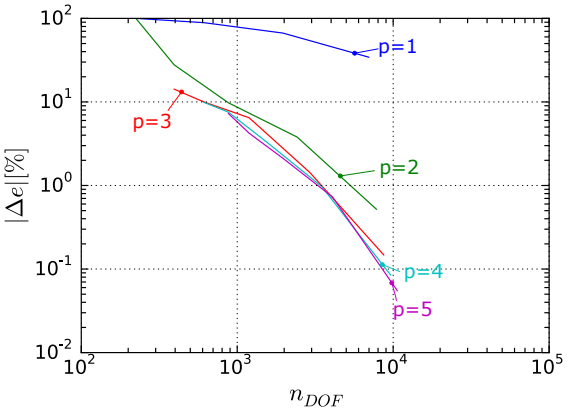


Figure 5: Convergence against FEM-solution for test case 2

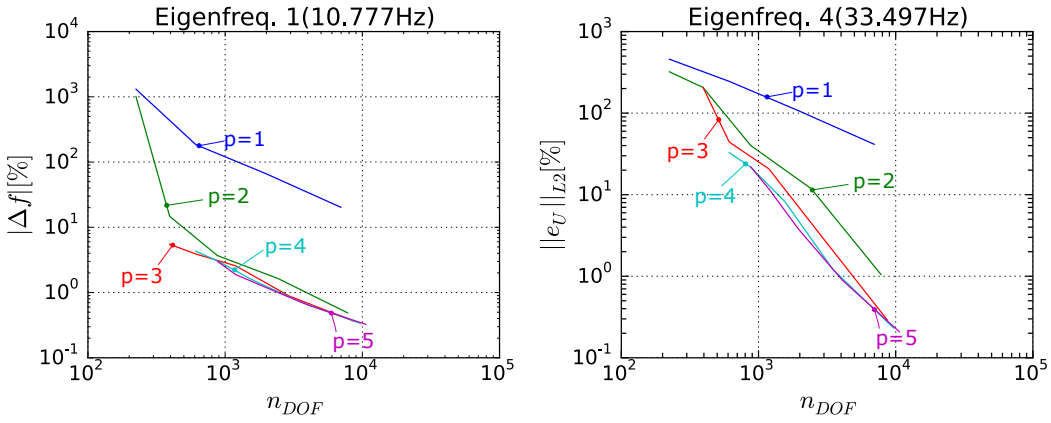


Figure 6: Convergence of eigenfrequencies no. 1 and 4 against FEM-solution

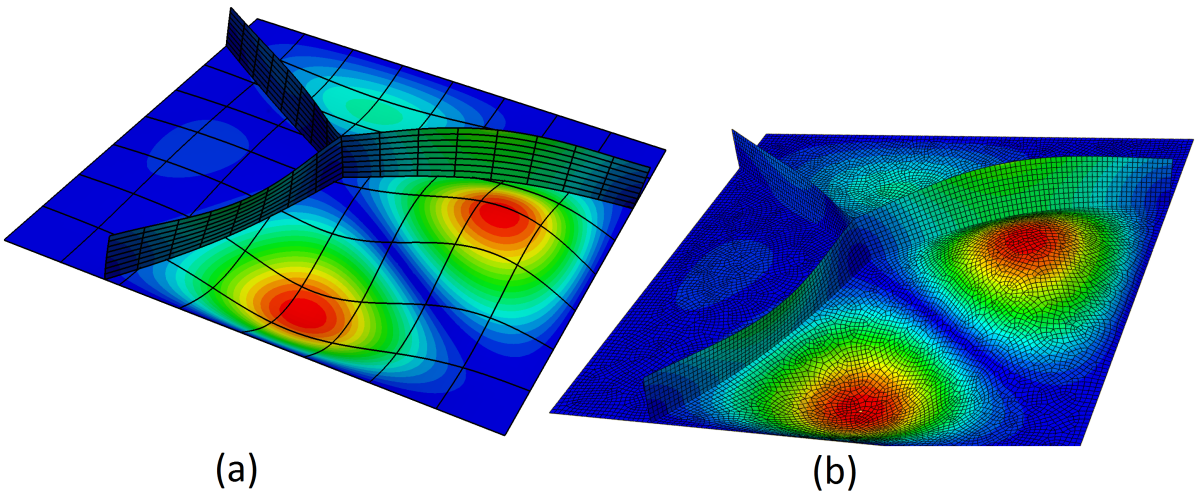


Figure 7: Eigenmode no. 4 obtained by IGA-model of order $p = 7$ and $n_{elem} = 64$ for panel and short stiffener and $n_{elem} = 128$ for long stiffener (a) and obtained by FEM-reference model (b)

5 Conclusion

A method to model stiffened structures in IGA based on Lagrange multipliers is implemented. With the Lagrange multiplier approach stiffeners and base panel can be modeled as separate domains and coupled afterwards. Even though NURBS are limited to structured meshes arbitrary shaped stiffeners can be coupled to a base panel without subdividing it into multiple parts with the proposed method. It could be shown that the results are of similar accuracy compared to standard FEM approaches with conforming meshes for static and modal analysis. The displacement coupling conditions can be fulfilled very accurate especially by higher order base functions. Stress singularities at the end of stiffeners does not lead to stress oscillations even in cases with high order base functions.
The convergence speed for higher order base functions is limited by the stress discontinuities at the stiffeners foot. This problem is not introduced by the Lagrange multiplier approach but by the way of modeling the stiffener and also occurs on conforming meshes. A way to evade the problem is local refinement what can be accomplished by T-Splines or adaptive IGA methods as described in [11] where the latter even the order p can be lowered locally. Both approaches seem promising for future improvements of the method.

REFERENCES

[1] T.J.R. Hughes, J.A. Cottrell, Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Comput. Methods Appl. Mech. and Engrg.*, **194**, 4135-4195, 2005.

[2] J.A. Evans, Y. Bazilevs, I. Babuška, T.J. Hughes, n-widths, sup-infs, and optimality ratios for the k-version of the isogeometric finite element method. *Comput. Methods Appl. Mech. and Engrg.*, **198**, 1726-1741, 2009.

[3] D. Chapelle, K.J. Bathe, *The Finite Element Analysis of Shells - Fundamentals, 2nd Edition*. Springer, 2011.

[4] E. Brivadis, A. Buffa, B. Wohlmuth, L. Wunderlich, Isogeometric mortar methods. *Comput. Methods Appl. Mech. and Engrg.*, **284**, 292-319, 2015.

[5] A. Seitz, P. Farah, J. Kremheller, B. Wohlmuth, W.A. Wall, A. Popp, Isogeometric dual mortar methods for computational contact mechanics. *Comput. Methods Appl. Mech. and Engrg.*, **301**, 259-280, 2016.

[6] O.C. Zienkiewicz, R.C. Taylor, *The finite element method, Vol. I, 4th Edition*. McGraw Hill, 1989.

[7] W. Dornisch, S. Klinkel, Treatment of Reissner-Mindlin shells with kinks without the need for drilling rotation stabilization in an isogeometric framework. *Comput. Methods Appl. Mech. and Engrg.*, **276**, 35-66, 2014.

[8] Y. Bazilevs, V.M. Calo, J.A. Cottrell, J.A. Evans, T.J.R. Hughes, S. Lipton, M.A. Scott, T.W. Sederberg, Isogeometric analysis using T-splines. *Comput. Methods Appl. Mech. Engrg.*, **199**, 229-263, 2010.

[9] A. Apostolatos, R. Schmidt, R. Wüchner, K.-U. Beltzinger, A Nitsche-type formulation and comparison of the most common domain decomposition methods in isogeometric analysis. *Int. J. Numer. Meth. Engng.*, **97**, 473-504, 2014.

[10] M.A. Ouso, T.A. Laursen, A mortar segment-to-segment contact method for large deformation solid mechanics. *Comput. Methods Appl. Mech. and Engrg.*, **193**, 601-629, 2004.

[11] Z. Wu, Z. Huang, Q. Liu, B. Zuo, A local solution approach for adaptive hierarchical refinement in isogeometric analysis. *Comput. Methods Appl. Mech. and Engrg.*, **283**, 1467-1492, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DOMAIN DECOMPOSITION SOLUTION SCHEMES FOR LARGE-SCALE IGA PROBLEMS

George Stavroulakis[1], **Dimitris Tsapetis**[1] **and Manolis Papadrakakis**[1]

[1] Institute of Structural Analysis and Seismic Research, National Technical University of Athens 9, Iroon Polytechneiou, Zografou Campus, GR-15780 Athens, Greece
e-mail: stavroulakis@nessos.gr

Keywords: Isogeometric Analysis, Refinement, PCG, IETI, Solutions scheme, Refinement, Mapping.

Abstract. *Isogeometric Analysis (IGA) [5] is a promising concept that establishes a close link between the technologies of CAD (computer aided design) and numerical simulation via finite element analysis (FEA). In the IGA framework, the same function spaces, which are used for the geometric representation of the computational domain, are used for the approximation of the problem unknowns. There are several computational geometry technologies that could serve as a basis for IGA with Non-Uniform Rational B-Splines (NURBS) being the most widely adopted due to their popularity in CAD software. In contrast with FEA where there is a very broad spectrum of solution techniques for the fast and efficient solution of the linear or linearized systems that occur [1-4], IGA solution schemes are still an open issue. With respect to domain decomposition techniques, the Isogeometric Tearing and Interconnecting (IETI) [8] method combines the advanced solver design of dual domain decomposition methods with the exact geometry representation of IGA, relying on patches for the subdivision of the domain. In this work, an innovative solution scheme is presented, showcasing greatly enhanced performance when compared to the established and tested solution schemes for IGA and its numerical performance is exhibited in numerical examples.*

1 SOLUTION TECHNIQUES

Iterative methods to solve a linear system such as Conjugate Gradient (CG) method [1, 2] are commonly used with the aid of a preconditioner [3, 4] to ensure fast convergence of the method. In our case, we employ a Preconditioned Conjugate Gradient (PCG) method in which the preconditioner is not a matrix as commonly thought of in numerical analysis, but a transformation in the wider mathematical sense. In the following sections, we will explain the need that led us to the proposal of a transformation, compare two different approaches of the new "preconditioner" and explain thoroughly the optimal among the two.

1.1 Large – scale problems in IGA

IGA is a method that was initially proposed by Hughes et al. [5] and envelops both the geometric information of the structures as well as the analysis mesh and data in one single model. Since 2005, that first appeared, IGA has gained a wide acceptance by the scientific community as it gradually establishes its background and grows rapidly. This raises the need of researchers to solve constantly large and more complex problems utilizing this promising method. These problems are of the form:

$$Ku = f \tag{1}$$

where K is the stiffness matrix, f is the applied force field and u is the displacement field. Unfortunately, the major advantage of IGA i.e.: higher Continuity Shape Functions, becomes a disadvantage. For example, NURBS Shape Functions span through several Isogeometric Elements which inevitably leads to a denser Stiffness Matrix with a greater bandwidth in comparison to simple Finite Elements.

Since complexity of a solver increases with increasing continuity [6, 7] it is crucial for the further evolution of IGA, the development of an algorithm that will provide both high accuracy and minimized time complexity compared to the existing techniques. IETI that was proposed in [8], is a powerful method that divides the problem into a multitude of independent problems and thus reduces the complexity of a large scale system. The major drawback of this approach is that the number of subdomains must be identical to the number of patches present to an IGA model. This means that subdivisioning is mesh-dependent which makes both load balancing and scalability to suffer in high performance computing environments. On the other hand, inducing discontinuities in a model by subdividing it into a number of subdomains, increases the error and discontinuities in second-order characteristics like stress and strain.

1.2 The PCG-IETI method

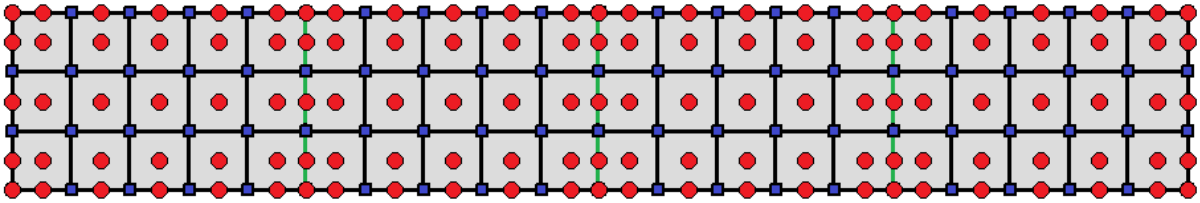
In order to circumvent this drawback, a new model with stiffness matrix K_p is introduced which we will refer to as "patched model" and has the geometry of the original model with appropriately induced discontinuities in the form of patches, in order to subdivide the original model appropriately. Due to the fact that there are errors introduced due to these discontinuities but the geometry of the problem is identical, the original and subdivided models constitute near-by problems with the stiffness matrix of the original model being equal to $K = K_p + \Delta K$.

Let's consider the PCG algorithm, equipped with a preconditioner following the rationale of incomplete Cholesky preconditionings which feature an error matrix E_i. This matrix is usually defined by the computed positions of small elements in the lower triangular matrix produced by the incomplete Cholesky factorization procedure, which do not satisfy a specified magnitude criterion and therefore are discarded [15]. Considering the near-by problem of the form $\left(K_p + \Delta K\right)u = f$, if matrix E_i is taken as ΔK, the preconditioning matrix becomes the initial matrix $\tilde{A} = K_p$. The repeated solutions required for the evaluation of the preconditioned residual vector $z^k = \tilde{A}^{-1} r^k$ are performed using IETI and are treated as problems with multiple right-hand sides.

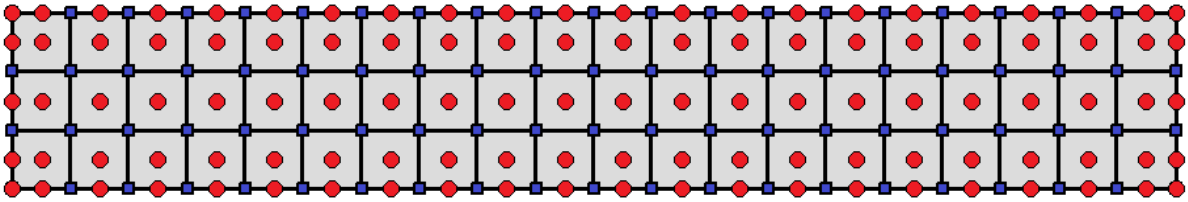
In contrast to models generated using FEA, IGA-generated models with the same geometry but different patches have different number of degrees of freedom (dof). This means that the residual vector r^k cannot be used in order to perform the evaluation of the preconditioned re-

sidual vector using matrix K_p. Even if that was possible, this evaluation would produce a vector with a different size compared to the vectors produced using matrix *K*. In order to circumvent this problem, a transformation between these two IGA meshes is employed. Specifically, with the aid of double transformation, the residual of the original model is being mapped to the patched model space and after the preconditioned residual vector evaluation, the solution is being mapped back to the original model space.

The PCG algorithm, equipped with the latter double-transformation preconditioner throughout the entire solution process, constitutes the PCG-IETI method providing both load balancing and scalability properties.



(a) Example of a Simple Cantilever split into four patches (patched model).



(b) Example of the same Cantilever modeled as a single patch (original model).

Figure 1: Cantilever examples. Black continuous lines illustrate the patch border and squares the Control Points in each case. Dotted lines depict the Control Net.

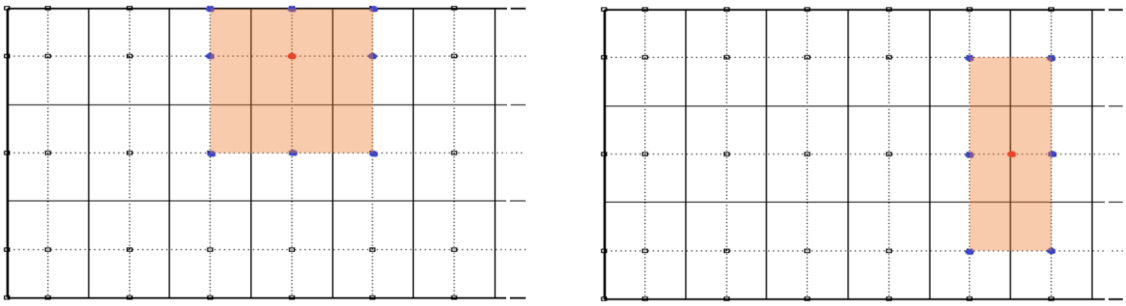
2 MAPPING

2.1 Shape Functions Mapping

Since the implementation of PCG-IETI requires a transformation between two different meshes, we utilized the intergrid transfer operator ***R*** - Restriction operator for the transfer of the fine grid information to the coarse one and the prolongation or interpolation Operator ***I*** for the transverse transfer from the coarse to the fine grid [10]. In our case, we use the full weighting Operator ***R*** [11] which is derives from interpolation Operator using:

$$\mathrm{R} = \mathrm{c} \cdot \mathrm{I}^{\mathrm{T}} \tag{2}$$

So by defining the interpolation operator ***I*** the transformation of the two meshes will be completed for application in the iterative solver. For finite elements purposes, a simple linear interpolation between the nodal values is adequate for the intergrid transfer, but since the application will be in the field of Isogeometric Analysis, the choice was to use a more complex interpolation that would provide better results taking into account the increased continuity of the Isogeometric method. Our choice is to use NURBS shape functions. The interpolation between the two control meshes is performed considering the Control Point parametric coordinates when the isoparametric domain in each direction spans in [0, 1]. Given the same parametric domain both coarse and fine mesh can transfer information between each other with simple interpolation.



(a) Approximation of an existing Control Point with the aid of NURBS interpolation. We observe that the point's value is defined with the aid of nine Control Points.

(b) Approximation of a non-existing Control Point with the aid of NURBS interpolation. We observe that the point's value is defined with the aid of six Control Points.

Figure 2: Domain of influence using NURBS interpolation. Rhinoceros [9] software was used.

In the next subsections we provide a brief description of Shape Function in Isogeometric Analysis.

2.1.1. B-Splines

Given a degree $p \geq 0$ of the shape functions and an ascending sequence $\Xi = \{\xi_1, \xi_2, ..., \xi_{n+p+1}\}$ of n+p+1 real values called Knot Vector, where n is the number of Control Points the B-Splines can be calculated using the recursive Cox-de Boor algorithm. Note that we only use open Knot Value Vectors which implies that the first and last Knot Value has multiplicity of p+1 and the internal values multiplicity m lower or equal to p. The recursive formulation of B-Splines starts with constant shape functions of degree p=0.

$$N_{i,0}(\xi) = \begin{cases} 1, & \text{if } \xi_i \leq \xi < \xi_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

For degree p>0 B-Splines are defined using p-1 degree according to the following formula:

$$N_{i,p}(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} \cdot N_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} \cdot N_{i+1,p-1}(\xi) \tag{4}$$

It must be noted that in the fractions of equation (3), when a denominator equals zero then the fraction's value is considered as zero.

2.1.2. NURBS

NURBS are produced using both B-Spline functions and the weights of the Control Points. NURBS curves are given by:

$$R_i^p(\xi) = \frac{N_{i,p}(\xi) \cdot w_i}{\sum_{\hat{i}=1}^{n} N_{\hat{i},p}(\xi) \cdot w_{\hat{i}}}$$
$$C(\xi) = \sum_{i=1}^{n} R_i^p(\xi) \cdot B_i \tag{5}$$

In a similar way we can define two and three dimensional NURBS Shape Function.

$$R_{i,j}^{p,q}(\xi,\eta)=\frac{N_{i,p}(\xi)\cdot M_{j,q}(\eta)\cdot w_{i,j}}{\sum_{\hat{i}=1}^{n}\sum_{\hat{j}=1}^{m}N_{\hat{i},p}(\xi)\cdot M_{\hat{j},q}(\eta)\cdot w_{\hat{i},\hat{j}}}$$
$$R_{i,j,k}^{p,q,r}(\xi,\eta,\zeta)=\frac{N_{i,p}(\xi)\cdot M_{j,q}(\eta)\cdot L_{k,r}(\eta)\cdot w_{i,j,k}}{\sum_{\hat{i}=1}^{n}\sum_{\hat{j}=1}^{m}\sum_{\hat{k}}^{n}N_{\hat{i},p}(\xi)\cdot M_{\hat{j},q}(\eta)\cdot L_{\hat{k},r}\cdot w_{\hat{i},\hat{j},\hat{k}}} \quad (6)$$

The resulting univariate, bivariate or trivariate NURBS Shape Function are used for interpolation of the new Control Points with known parametric coordinates. Interpolation is used for approximation of both displacements and forces of the new Control Net using the following:

$$u_x=\sum N_i\cdot u_{x,i} \qquad u_y=\sum N_i\cdot u_{y,i} \qquad u_z=\sum N_i\cdot u_{z,i} \quad (7)$$

As eq. (6) shows that the approximation of the displacements of e.g. the patched Control Net is performed using the displacements and Shape Functions of the unpatched one and vice versa. In the case of forces, the transpose of the Shape Functions is used. Specifically, for calculation of the forces of the patched model, the forces of the unpatched are used with the transpose of the patched shape functions and vice versa which sums up to:

$$f_x=\sum N_i^T\cdot f_{x,i} \qquad f_y=\sum N_i^T\cdot f_{y,i} \qquad f_z=\sum N_i^T\cdot f_{z,i} \quad (8)$$

2.1.3. Cook Cantilever Example

In order to illustrate the performance of the NURBS Shape Function transformation, a cook cantilever was examined. The transformation had to transfer the information from the single-patch cantilever with 11x3 Control Points to a Cantilever subdivided into 9 patches with resulting 19x3 Control Points as shown in Fig. 1. By applying the transformation to this model we approximate displacements and forces. In Table 1 we can see how the approximation works by using norms of the resulting values compared to the initial ones. Even thought at first the approximated displacements seem close enough based on the total norms, the third and norm clearly depicts that an error up to 8% is induced using this technique.

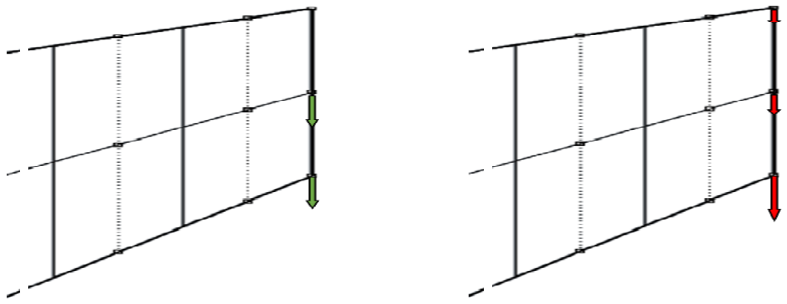
	Unpatched	Patched
$\text{norm}(U_{\text{exact}})$	0.1283	0.1561
$\text{norm}(U_{\text{approximate}})$	0.1294	0.1571
Error norm	7.15%	7.94%

Table 1: Performance of transformation on displacements using Shape Functions.

In case of forces, the error as shown in Table 2 dramatically increases. Unfortunately, this approach does not only create an error of the value but also changes the loading as clearly illustrated in Fig.3. This causes the algorithm either to converge real slow or in some cases even diverge due to the large error created in every intergrid approximation. In case of non-linear analysis, wrong estimation of both internal and external forces, can result to much greater error induced in the solution procedure and thus wrong results.

	Unpatched	Patched
$\text{norm}\left(F_{\text{exact}}\right)$	141.4214	141.4214
$\text{norm}\left(F_{\text{approximate}}\right)$	136.9306	136.9306
$\frac{\text{norm}\left(F_{\text{exact}} - F_{\text{approximate}}\right)}{\text{norm}\left(F_{\text{exact}}\right)}$	43.3%	43.3%

Table 2: Performance of transformation on forces using Shape Functions.



(a) Loads on a Cantilever before transformation

(b) Load on the Cantilever after the transformation

Figure 3: Changes of the Cantilever Loading due to the transformation. Rhinoceros [9] software was used.

2.2 Isogeometric Refinement Mapping

The second approach of the mapping was based on Isogeometric Refinement. Since our goal is to transfer information between two meshes that are based on the same structure, refinement as proposed in [5] can be considered the best tool applicable. It is used to calculate coordinates and weights of the new Control Points keeping the geometry unchanged, so it seems logical to approximate the data needed using this method. In [12] it is clearly illustrated how a structure can be analyzed using multiple separate patches, yet it is very important to explain the link with Isogeometric refinement in our case. Following the inverse procedure of [12] we want to segment one single patch into multiple individual subdomains for the sake of our iterative solver.

2.2.1. h-Refinement

The refinement algorithm that will be used in the aforementioned transformation is h-refinement or knot insertion. For a known knot value vector $\Xi=\{\xi_1, \xi_{2,}\ldots,\xi_{n+p+1}\}$ we want to introduce new knot values on increase the multiplicity of already known ones. For each one of the new inserted knot values the number of control points is increased by one. Their coordinates and weights $\bar{B}$ are computed

$$\bar{B} = T^p \cdot B \tag{9}$$

Where

$$T_{ij}^0(\xi) = \begin{cases} 1, & \text{if } \xi_j \le \bar{\xi}_i < \xi_{j+1} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

and

$$T_{ij}^{q+1} = \frac{\bar{\xi}_{i+q} - \xi_j}{\xi_{j+q} - \xi_j} \cdot T_{ij}^q + \frac{\xi_{j+q+1} - \bar{\xi}_{i+q}}{\xi_{j+p+1} - \xi_{j+1}} \cdot T_{ij+1}^q \tag{11}$$

2.2.2. Multiple patches segmentation

In Fig.4 we observe the 1D shape functions of a structure model as one patch. It is known from[13]that NURBS [13] Shape functions have a continuity of C^{p-m} over knots, where p is the polynomial degree of the function and m is the multiplicity of the knot. Fig. 4 shows NURBS of degree p=2 with C^1 continuity on knots. Using Knot Insertion algorithm of Isogeometric Refinement we raise the multiplicity of selected knots creating a C^0 continuity resulting in Fig. 5.

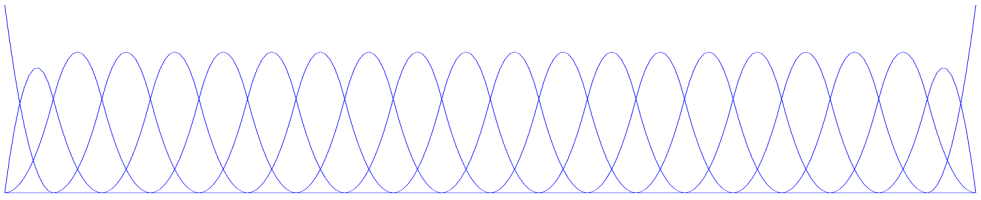


Figure 4: 1D NURBS of one single patch.

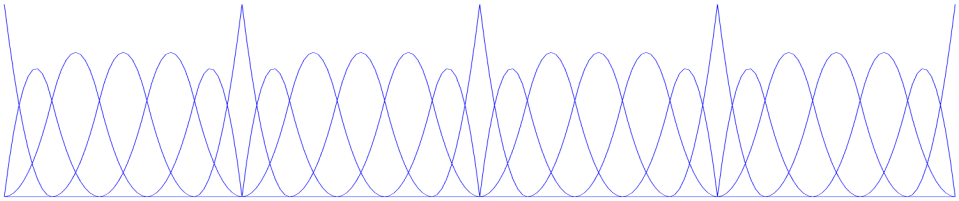


Figure 5: 1D NURBS of one single patch with C^0 continuity over selected knots.

In Fig.6. we observe the domain after the knot value insertion resembles to several independent subdomains united as one by merging their C^{-1} boundary shape function into a new C^0 one. In addition to the shape function the boundary Control Points between the two subdomains coincide. Considering that they have an independent role for each subdomain we result in Fig.7. where now every patches are independent and when combined with its data can be processed as a separate entity like in Fig.8. only take into account the boundary interconnection with other subdomains.

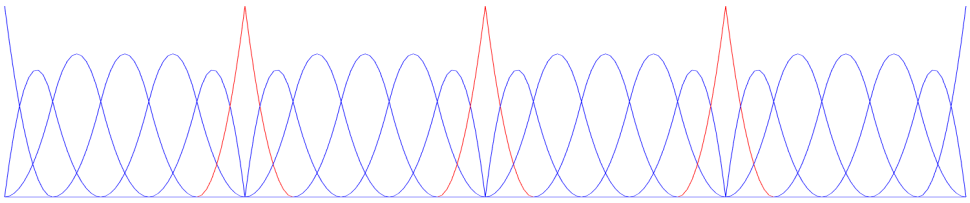


Figure 6: 1D NURBS of one single patch with C^0 continuity over selected knots.

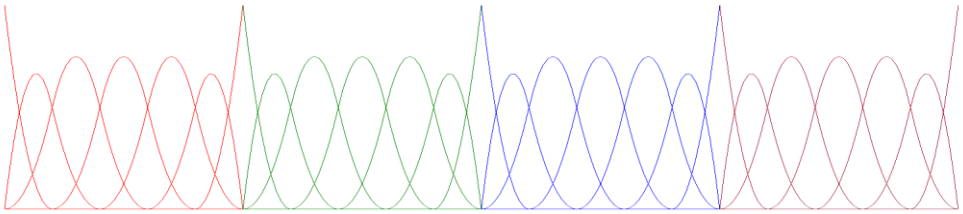


Figure 7: 1D NURBS of one single patch with C^0 continuity over selected knots.

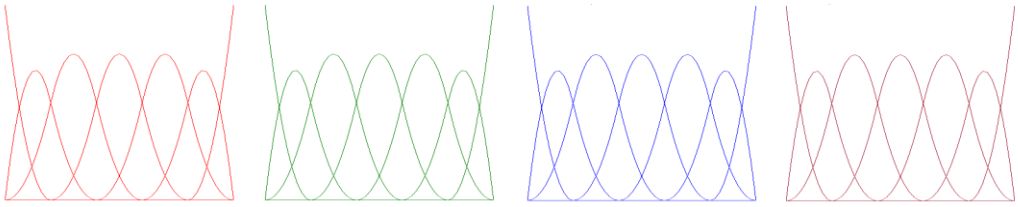


Figure 8: 1D NURBS of four discrete patches

2.2.3. Refinement transformation

As described above the Control Points of the segmented structure are given by h-refinement of the single-patched domain. The transformation matrix in this 1D case as described in [5, 12, and 13] is:

$$\underset{(mx3)}{\left\{P^{Multipatch}_{(\xi)}\right\}} = \underset{(mxn)}{\left[T^{MS}\right]} \cdot \underset{(nx3)}{\left\{P^{Singlepatch}_{(\xi)}\right\}} \tag{12}$$

In order to create the inverse transformation, we need to remove the inserted knot values. Since they are removable without altering the geometry a matrix is needed that will provide us with the inverse transformation.

$$\underset{(nx3)}{\left\{P^{Singlepatch}_{(\xi)}\right\}} = \underset{(nxm)}{\left[T^{SM}\right]} \cdot \underset{(mx3)}{\left\{P^{Multipatch}_{(\xi)}\right\}} \tag{13}$$

Using eq.11 and since knot values are removable we have:

$$\underset{(mx3)}{\left\{P^{Multipatch}_{(\xi)}\right\}} = \underset{(mxn)}{\left[T^{MS}\right]} \cdot \underset{(nx3)}{\left\{P^{Singlepatch}_{(\xi)}\right\}} \Rightarrow$$

$$\underset{(nxm)}{\left[T^{MS}\right]^{T}} \cdot \underset{(mx3)}{\left\{P^{Multipatch}_{(\xi)}\right\}} = \underset{(nxn)}{\left(\underset{(nxm)}{\left[T^{MS}\right]^{T}} \cdot \underset{(mxn)}{\left[T^{MS}\right]}\right)} \cdot \underset{(nx3)}{\left\{P^{Singlepatch}_{(\xi)}\right\}} \Rightarrow \tag{14}$$

$$\underset{(nx3)}{\left\{P^{Singlepatch}_{(\xi)}\right\}} = \underset{(nxn)}{\left(\underset{(nxm)}{\left[T^{MS}\right]^{T}} \cdot \underset{(mxn)}{\left[T^{MS}\right]}\right)^{-1}} \cdot \underset{(nxm)}{\left[T^{MS}\right]^{T}} \cdot \underset{(mx3)}{\left\{P^{Multipatch}_{(\xi)}\right\}}$$

So the inverse transformation can be created using the initial one by:

$$\underset{(nxm)}{\left[T^{SM}\right]} = \underset{(nxn)}{\left(\underset{(nxm)}{\left[T^{MS}\right]^{T}} \cdot \underset{(mxn)}{\left[T^{MS}\right]}\right)^{-1}} \cdot \underset{(nxm)}{\left[T^{MS}\right]^{T}} \tag{15}$$

This transformation matrix is one dimensional so it can be applied to each set of Control Points on axis Ξ. In case, we want to apply a refinement in more than one axes, a new equivalent transformation matrix for each refined axis must be created and consecutively applied in order to create the final Control Net. It is important to note that the null difference of the loads is a property that derives from knot insertion in IGA and thus it not an approximation but an exact transformation. Specifically, for the Shape Functions we know:

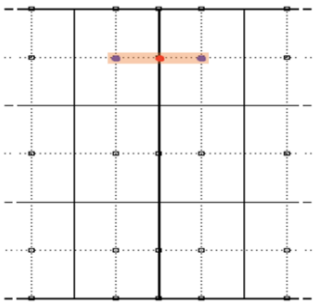
$$\underset{(mx1)}{\left\{P^{Multipatch}_{(\xi,\eta,\zeta)}\right\}} = \underset{(mxn)}{\left[T^{MS}\right]^{T}} \cdot \underset{(nx1)}{\left\{P^{Singlepatch}_{(\xi,\eta,\zeta)}\right\}} \tag{16}$$

The calculation of loads on the multipatch domain is performed by:

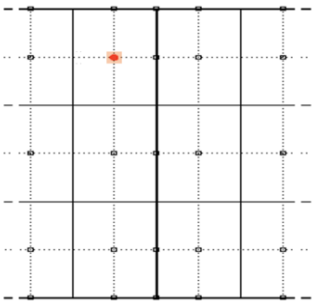
$$\begin{aligned}
&\underset{(mx3)}{\left\{L^{Multipatch}\right\}} \int_{\Omega} \underset{(mx1)}{\left\{R^{Multipatch}_{(\xi,\eta,\zeta)}\right\}} \underset{(1x3)}{f\left(\xi,\eta,\zeta\right)} \det\left[J\right] d\Omega = \\
&\int_{\Omega} \underset{(mxn)}{\left[T^{MS}\right]^{T}} \underset{(nx1)}{\left\{R^{Single-patch}_{(\xi,\eta,\zeta)}\right\}} \underset{(1x3)}{f\left(\xi,\eta,\zeta\right)} \det\left[J\right] d\Omega = \\
&\underset{(mxn)}{\left[T^{MS}\right]^{T}} \int_{\Omega} \underset{(nx1)}{\left\{R^{Single-patch}_{(\xi,\eta,\zeta)}\right\}} \underset{(1x3)}{f\left(\xi,\eta,\zeta\right)} \det\left[J\right] d\Omega \Rightarrow \\
&\underset{(mx3)}{\left\{L^{Multipatch}\right\}} = \underset{(mxn)}{\left[T^{MS}\right]^{T}} \cdot \underset{(nx3)}{\left\{L^{Single-patch}\right\}}
\end{aligned} \tag{17}$$

Finally, Fig. 8 shows the influence to the approximation of the new mesh. In case of a new Control Point its value is an interpolation of the Control Points closest to the inserted knot value and only on the axis knot insertion is performed. On the other hand, when a Control Point is not in proximity to an h-refinement its value and position is identical to its previous state. Fig.9 and 10 illustrate how NURBS affect the data continuity. For instance, in Fig. 9 examining the yellow element separation line of axis Ksi, we observe that full continuity allows NURBS to spans smoothly across many elements. On the other hand, when the same yellow line is distorted in Fig.10 by raising the multiplicity of the knot value, only one shape function now connects different subdomains. In addition, as expected the insertion of a knot

value resulted to the insertion of a single shape function in axis Ksi. Distortion of the single patch mesh only occurred on the subdomain interface while all other Control Points remained intact. In case we had a larger degree of shape functions this change would have also affected Control Points close to the knot inserted by changing their coordinates and weights.



(a) *Approximation of a non-existing Control Point with the aid of Refinement Transformation. We observe that the point's value is defined with the aid of two Control Points.*



(b) Approximation of an existing Control Point with the aid of Refinement Transformation. We observe that the point's value is retained to the identical Control Point of the previous mesh.

Figure 8: Domain of influence using Refinement Transformation.

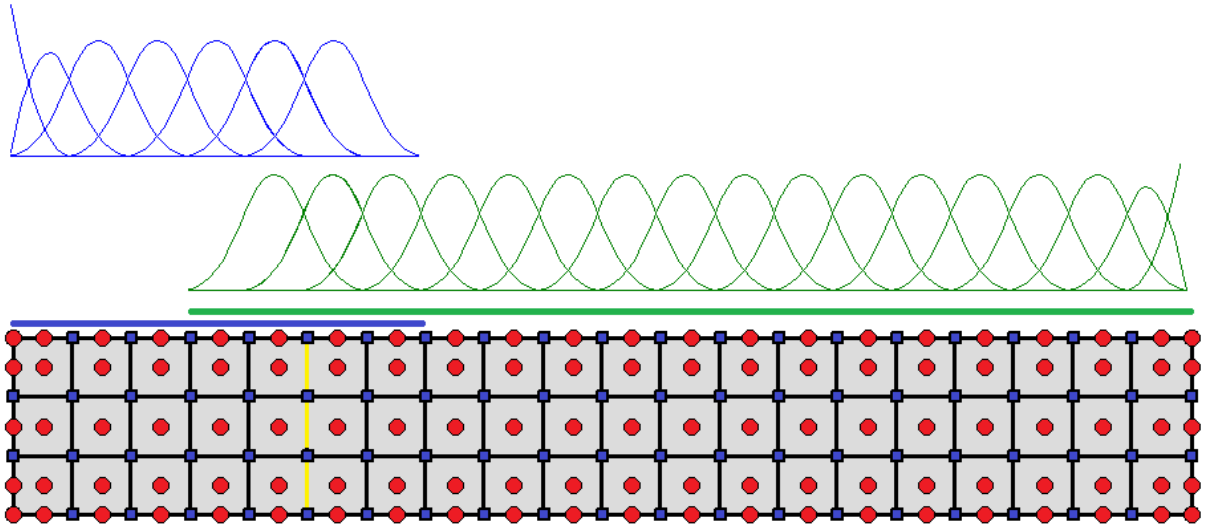


Figure 9: Shape functions and continuity of a single patch.

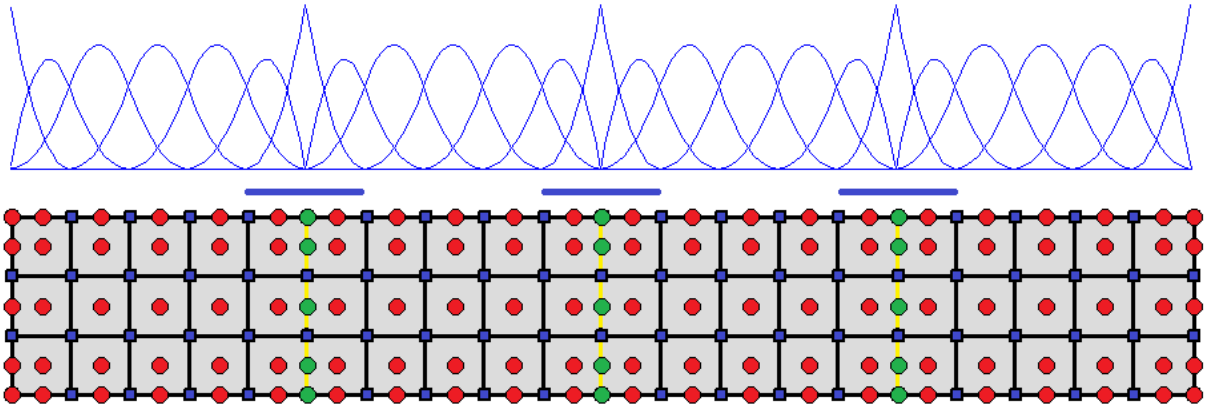


Figure 10: Shape functions and continuity of multiple patched structure.

2.2.4. Cook Cantilever Example

The same cantilever example of 1.2.3 will be used here in order to compare the results of the two approaches.

	Unpatched	Patched
$\text{norm}\left(U_{\text{exact}}\right)$	0.1283	0.1561
$\text{norm}\left(U_{\text{approximate}}\right)$	0.1284	0.1560
Error norm	0.23%	0.34%

Table 3: Performance of transformation on displacements using Refinement Transformation Matrix

	Unpatched	Patched
$\text{norm}\left(F_{\text{exact}}\right)$	141.4214	141.4214
$\text{norm}\left(F_{\text{approximate}}\right)$	141.4214	141.4214
Error norm	0%	0%

Table 4: Performance of transformation on forces using Refinement Transformation Matrix

It is clear given the two comparison matrices that the error of the intergrid transformation is now minimized. The transformation between the meshes can now act as a preconditioner that will not induce error and approximate with increased accuracy the single patch solution. As result the refinement transformation considered ideal and it will be implemented and tested in three dimensional examples.

3 NUMERICAL EXAMPLES

Since the comparison of the two proposed transformation methods resulted in advantage of the Isogeometric Refinement method, two numerical examples will be provided that will illustrate the performance of the method in real life three dimensional problems. Note that, our proposal was tested on the simple cantilever beams that was described above providing satisfactory results, yet the numerical examples chosen will be of different degrees of freedom magnitudes in order to illustrate the wide range of applications for the proposed method.

3.1 Pipe

The first application of the PCG-IETI method is a pipe and this method is tested on two models with 1k and 10k dof respectively.

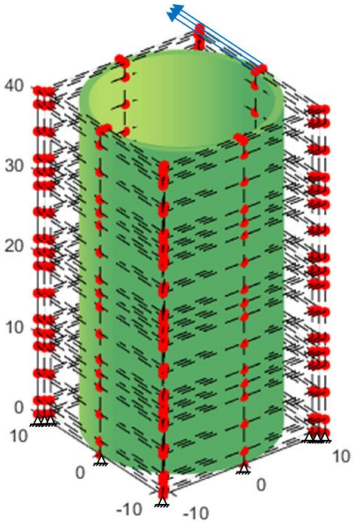


Figure 11: Patched Pipe with 459 Control Points.

In Figure 11 we can see the patched model counterpart of the pipe model. It is created with three dimensional NURBS using NURBS Toolbox [14]. The object is created by 9 Control Points per axis ξ that create the circumference of the circle, 3 per axis η that create the radius of the annulus and 17 per axis η that give height to the pipe. The degree of NURBS Shape functions is considered to be consistent in all parametric directions and equal to 2 for the sake of simplicity. The aforementioned Control Points result to (9x3x17)x3=1377 degrees of freedom. As depicted in Fig.11, all Control Points at the base are clamped, in order to constrain the base displacements. In addition, concentrated loads are applied to Control Points interpolatory to the geometry, introducing both a bending and torsional strain to the structure. The Pipe is segmented into four equal patches of (9x3x5)x3=405 degrees of freedom for the implementation of the IETI part of the implemented method. This patched Pipe is used to approximate a single-patched pipe of similar load and boundary conditions. Both geometries are depicted in Fig.12. The single patched structure is composed now by 9x3x14 Control Points resulting to 1134 degrees of freedom, which are of similar magnitude to the ones of the patched structure.

(a) (b)

Figure 12: (a) Pipe designed as a single patch. (b) Pipe designed as four separate patches.

In similar fashion, the Pipe structure is analyzed in the case of 10k Dof magnitude. Fig.13 shows the new geometries as created with NURBS Toolbox both patched and unpatched.

(a) (b)

Figure 13: (a) Pipe designed as a single patch. (b) Pipe designed as ten separate patches.

The new structures are considered now to examine a greater order of magnitude. Specifically, the single patch Pipe has now 17 Control Points per axis ξ, 3 Control Points per axis η and 62 Control Points per axis ζ. This results to (17x3x62)x3=9486 degrees of freedom while all other properties, such as degree of Shape Functions, Loads and Boundary Condition are retained. The patched structure is now defined by 17x3x71 Control Points resulting to 10863 degrees of freedom.

	Pipe 1K	Pipe 10K
PCG-IETI iterations	10	6
Total IETI iterations	22	21
Mean IETI iterations	2,2	3,5

Table 5: Performance metrics of the PCG-IETI method for the two pipe models.

In table 5, the performance metrics of the PCG-IETI method are presented, considering an error tolerance of 10^{-5}. The total iterations are depicted on the first row where it can be seen that the 1K model needs more iterations to converge, compared to the 10K one. This is due to the fact that the errors introduced from the patched model are larger for the 1K model, compared to the 10K model. Moreover, a mean of 2.2 and 3.5 IETI iterations where needed for each PCG-IETI iteration. The error for each iteration is depicted on Fig. 14

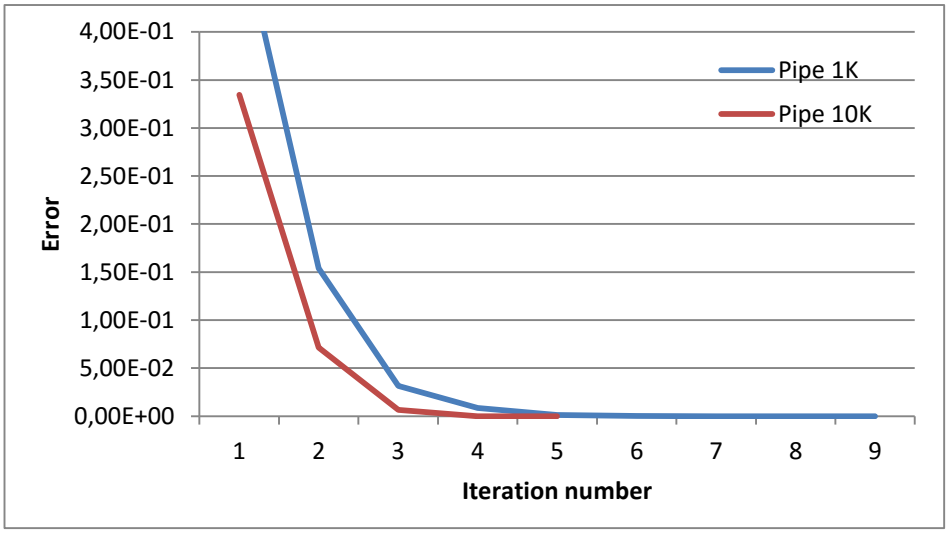


Figure 14: Iterations vs. error for the two pipe models.

3.2 Shell

Second example for the proposed methodology is a Shell once more tested in the same orders of magnitude as the Pipe example.

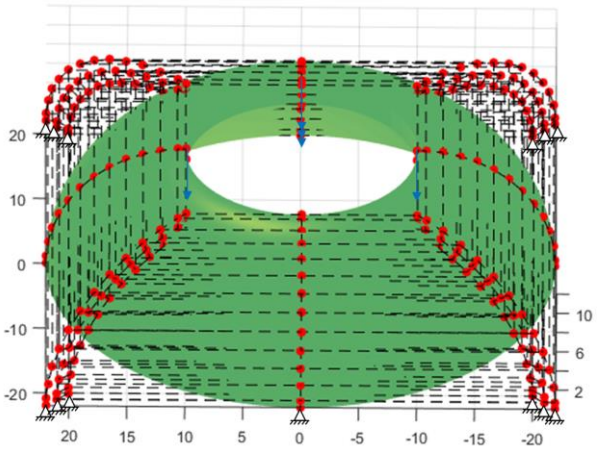
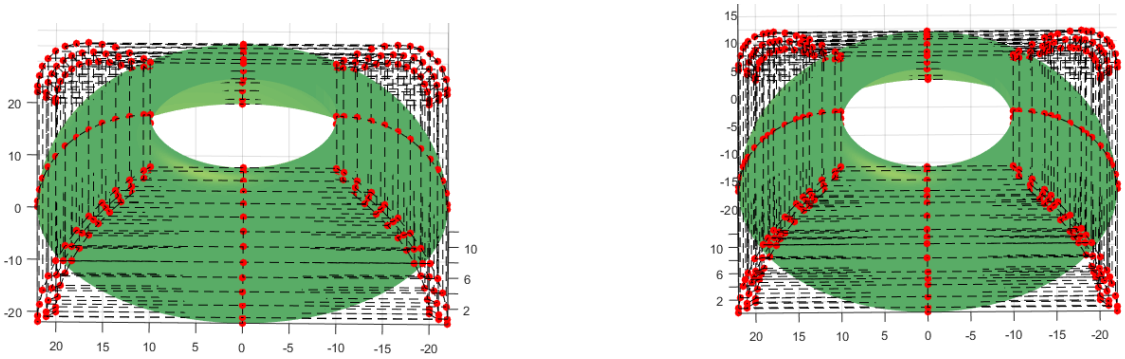


Figure 15: 3D Shell designed with NURBS Toolbox.

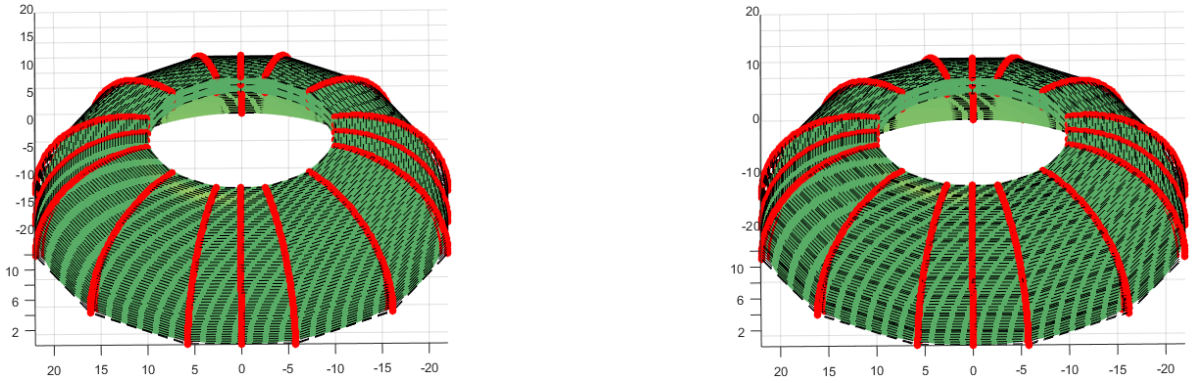
The unpatched shell structure is created by using 9 Control Points per axis ξ, 14 per axis η and 3 per axis ζ, producing 1134 degrees of freedom. As shown in Fig.15, the all the base Control Points are clamped to constrain the base displacements. Loads are once again applied to interpolatory Control Points of the Shell as blue arrows illustrates. Four patches are again used to approximate the exact solution. Both patched and unpatched geometries are shown in Fig.16. Each patch consist of (9x5x3)x3=405 degrees of freedom creating a 1377 Dofs structure.



(a) (b)

Figure 16: (a) Shell designed as a single patch. (b) Shell designed as four separate patches.

Finally, two Shells are analyzed for the 10k Dof case. The degrees of freedom are now, 9486 for the single patch and 10863 for the patched one. Ten patches are used now with 1224 Control Points each for the IETI part of the method. Fig.17 illustrated these two structures and their Control Nets.



(a) (b)

Figure 17: (a) Shell designed as a single patch. (b) Shell designed as ten separate patches.

	Shell 1K	Shell 10K
PCG-IETI iterations	8	5
Total IETI iterations	20	20
Mean IETI iterations	2,2	4

Table 6: Performance metrics of the PCG-IETI method for the two shell models.

In table 6, the performance metrics of the PCG-IETI method are presented, considering an error tolerance of 10^{-5}. The total iterations are depicted on the first row where it can be seen that the 1K model needs more iterations to converge, compared to the 10K one as for the case of the pipes. Once again, this is due to the fact that the errors introduced from the patched model are larger for the 1K model, compared to the 10K model. Moreover, a mean of 2.2 and 4 IETI iterations were needed for each PCG-IETI iteration. The error for each iteration is depicted on Fig. 18

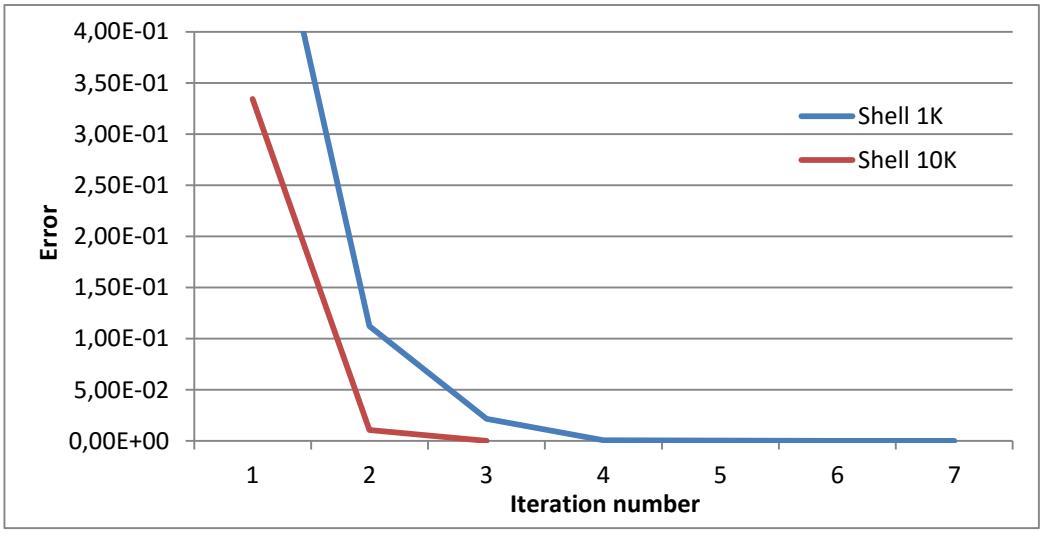


Figure 18: Iterations vs. error for the two pipe models.

REFERENCES

[1] M. Hestenes, E. Stiefel, Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards 49*, 1952.

[2] Y. Saad, Iterative methods for sparse linear systems (2nd Ed.). Philadelphia, *Society for Industrial and Applied Mathematics*, 2003

[3] Axelsson, Owe, Iterative Solution Methods, *Cambridge University Press*, 1996

[4] Y. Saad, H. van der Vorst, Iterative solution of linear systems in the 20th century, §8 Preconditioning methods, pp 193–8 of Numerical Analysis: Historical Developments in the 20th Century, *Elsevier Science Publishers*, 2001.

[5] T.J.R. Hughes, J.A. Cottrell, Y. Bazilevs, Isogeometric analysis: CAD finite elements NURBS exact geometry and mesh refinement, *Comput. Methods Appl. Mech. Engrg*. 194 (2005) 4135–4195

[6] N. Collier, D. Pardo, L. Dalcin, M. Paszynski, V.M. Calo, The cost of continuity: A study of the performance of isogeometric finite elements using direct solvers, *Comput. Methods Appl. Mech. Engrg*, 2011

[7] N. Collier, L. Dalcin, D. Pardo, V.M. Calo, The cost of continuity: Performance of Iterative Solvers on Isogeometric Finite Elements, *Society for Industrial and Applied Mathematics*, 2013

[8] Stefan K. Kleiss, C. Pechstein, B. Jüttler, S. Tomar, IETI – Isogeometric Tearing and Interconnecting, *Comput. Methods Appl. Mech. Engrg, 2012*

[9] Rhino, CAD Modeling and Design Toolkit. < www.rhino3d.com >

[10] Y. Zhu; A. C. Cangellaris, Multigrid finite element methods for electromagnetic field modeling, *Wiley*, 2006.

[11] W. L. Briggs, Van Emden Henson, S. F. McCormick, A Multigrid Tutorial, *SIAM*, 2000

[12] J. A. Cottrell, T. J. R Hughes, Y. Bazilevs, Isogeometric Analysis: Toward Integration of CAD and FEA, *Wiley*, 2009

[13] L. A. Piegl, W. Tiller, The NURBS Book, *Springer*, 1997

[14] D.M.Spink, NURBS Toolbox

[15] M. Papadrakakis, Solving large-scale linear problems in solid and structural mechanics, in: M. Papadrakakis (Ed.), Solving Large-Scale Problems in Mechanics, John Wiley & Sons, Chichester, 1993, pp. 1–37.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

NUMERICAL PROPERTIES OF A DISCONTINUOUS GALERKIN FOMULATION FOR ELECTRO-THERMAL COUPLED PROBLEMS

L. Homsi[1], C. Geuzaine[2], and L. Noels[1]

[1]Computational & Multiscale Mechanical of Materials (CM3),
Department of Aerospace and Mechanical Engineering,
University of Liège,
Quartier Polytech 1, Allée de la Découverte 9, B-4000 Liège, Belgium
lina.homsi@student.ulg.ac.be, L.Noels@ulg.ac.be

[2] Department of Electrical Engineering and Computer Science,
University of Liège,
Quartier Polytech 1, campus Allée de la Découverte 10, B-4000 Liège, Belgium
cgeuzaine@ulg.ac.be

Keywords: Multi-scale, Electro-thermal coupling, Discontinuous Galerkin method, Nonlinear elliptic problem.

Abstract. *Discontinuous Galerkin (DG) methods are attractive tools to integrate several PDEs in engineering sciences, due to their high order accuracy and their high scalability in parallel simulations. The main interest of this work is to derive a constant and stable Discontinuous Galerkin method for two-way electro-thermal coupling analyses.*
A fully coupled nonlinear weak formulation for electro-thermal problems is developed based on continuum mechanics equations which are discretized using the Discontinuous Galerkin method. Toward this end, the weak form is written in terms of energetically conjugated fields gradients and fluxes.
In order to validate the effectiveness of the formulation and illustrate the algorithmic properties, a numerical test for composite materials is performed.

1 INTRODUCTION

There has been increased a recent interest in electro-thermal coupling due to its application in energy conversion [3, 4, 7, 8].
The objective of this paper is to extend the DG method to study steady state and response of electro-thermal material model taking into account the Peltier and Seebeck effects.
Some of the features of DG methods include the use of structured grid and high order approximation, hp adaptivity as there is no need for continuity requirement across element interfaces, and the flexibility in terms of mesh design. However, for practitioners, it is important to have methods available which yield reliable results for a wide variety of problems. By using stabilization techniques and inter element flux definitions, the shortcomings of non-stabilized DG methods can be overcome.
Recently, DG has been used to solve coupled problems. For instance a primal DG method with interior penalty (IP) terms has been proposed in [10] to solve coupled reactive transport in porous media. A DG method has also been used in [9] to solve the thermo-elastic coupling problems due to temperature and pressure dependent thermal contact resistance.
We have recently extended the DG method to solve electro-thermal coupled problems in terms of energetically conjugated fields gradients and fluxes [1], which to the authors knowledge, has not been introduced yet. In [1], the numerical properties of the nonlinear elliptic problem, following the method proposed in [2, 5], have been derived. In particular, the convergence rates of the error in both the energy and L^2-norms have been shown to be optimal with respect to the mesh size in terms of the polynomial degree approximation k (respectively in order $k-1$ and k).
The new formulation proposed in this paper is able to capture the electro-thermal effects, which describe the direct conversion of electric potential difference into temperature difference and vice versa. In addition it is able to effectively capture the electro-thermal behavior of the composite materials. This paper is structured as follows. In Section 2, the electro-thermal governing equations are discussed. In Section 3, the DG weak form of electro-thermal coupling is presented. Numerical results in the response of composite materials are shown in Section 4. Section 5 is dedicated to the concluding remarks.

2 Governing equations

Let us consider a volume Ω and let the boundary of the domain $\partial\Omega$ be the union of two disjoint sets: the Dirichlet boundary, $\partial_{\mathrm{D}}\Omega$ and Neumann boundary, $\partial_{\mathrm{N}}\Omega$.
The first balance equation is the electrical charge conservation equation

$$\nabla \cdot \mathbf{j}_{\mathrm{e}} = 0 \qquad \forall\, \mathbf{x} \in \Omega, \tag{1}$$

where $\mathbf{j}_{\mathrm{e}}$ denotes the flow of electrical current density vector, which is defined as

$$\mathbf{j}_{\mathrm{e}} \;= \mathbf{l} \cdot (-\nabla \mathrm{V}) + \alpha \mathbf{l} \cdot (-\nabla \mathrm{T}). \tag{2}$$

In this equation α is the Seebeck coefficient, and $\mathbf{l}$ is the electric conductivity .
The second balance equation is the conservation of the energy flux

$$\nabla \cdot \mathbf{j}_{\mathrm{y}} \;= -\partial_{\mathrm{t}} \mathrm{y} \qquad \forall\, \mathbf{x} \in \Omega. \tag{3}$$

The right hand side of this equilibrium equation is the time derivative of the internal energy density y

$$\mathrm{y} \;= \mathrm{y}_0 + \mathrm{c}_{\mathrm{v}}\, \mathrm{T}, \tag{4}$$

which consists of the constant y_0 independent of the temperature and of the electric potential, and of the volumetric heat capacity c_v multiplied by the absolute temperature T.
Moreover the energy flux $\mathbf{j}_y$, which is a combination of the inter exchanges between the thermal and electric energies, is defined as

$$\mathbf{j}_y \ = \mathbf{q} + \ V\mathbf{j}_e, \tag{5}$$

where, $\mathbf{q}$ is the heat flux, defined as

$$\begin{aligned} \mathbf{q} \ &= \mathbf{k}\cdot(-\nabla T) + \beta_\alpha \mathbf{j}_e = \mathbf{k}\cdot(-\nabla T) + \alpha T\mathbf{j}_e && (6)\\ &= (\mathbf{k} + \alpha^2\, T\mathbf{l})\cdot(-\nabla T) + \alpha T\mathbf{l}\cdot(-\nabla V). && (7)\end{aligned}$$

In this equation $\mathbf{k}$ denotes the thermal conductivity and $\beta_\alpha = \alpha T$ is Peltier coefficient.
First we rewrite the equations (2, 5, 6) under the form

$$\mathbf{j} = \begin{pmatrix} \mathbf{j}_e \\ \mathbf{j}_y \end{pmatrix} = \begin{pmatrix} \mathbf{l} & \alpha\mathbf{l} \\ V\mathbf{l} + \alpha T\mathbf{l} & \mathbf{k} + \alpha V\mathbf{l} + \alpha^2 T\mathbf{l} \end{pmatrix} \begin{pmatrix} -\nabla V \\ -\nabla T \end{pmatrix}. \tag{8}$$

The conservation laws of this problem can then be formulated as finding $V,\, T \in H^2(\Omega) \times H^{2^+}(\Omega)$ such that

$$\begin{aligned} \nabla\cdot\mathbf{j}_e = 0 \qquad & \forall\, \mathbf{x}\in\Omega, && (9)\\ \nabla\cdot\mathbf{j}_y = \nabla\cdot\mathbf{q} + \mathbf{j}_e\cdot\nabla V = -\partial_t y \qquad & \forall\, \mathbf{x}\in\Omega, && (10)\\ T = \bar{T} > 0, \quad V = \bar{V} \qquad & \forall\, \mathbf{x}\in\partial_D\Omega, && (11)\\ \mathbf{q}\cdot\mathbf{n} = \bar{q}, \quad \mathbf{j}_e\cdot\mathbf{n} = \bar{j}_e, \quad \mathbf{j}_y\cdot\mathbf{n} = \bar{j}_y \qquad & \forall\, \mathbf{x}\in\partial_N\Omega, && (12)\end{aligned}$$

where $\bar{T}$ and $\bar{V}$ are the prescribed temperature and electric potential respectively, and $\mathbf{n}$ is the outward unit normal to the boundary $\partial\Omega$.
It should be noted that $H^{2^+}(\Omega)$ is the manifold to which T belongs, which is always positive.
The set of equations (9, 10) can be rewritten under a matrix form using Eq. (8)

$$\mathrm{div}\,(\mathbf{j}) = \begin{pmatrix} 0 \\ -\partial_t y \end{pmatrix} = \mathbf{i}. \tag{13}$$

Let us define the vector of the unknown fields $\mathbf{M}(2\times 1) = \begin{pmatrix} f_V \\ f_T \end{pmatrix}$, with $f_V = -\frac{V}{T}$ and $f_T = \frac{1}{T}$, [1, 3]. Indeed it can be shown that the fluxes $\mathbf{j}_e$, $\mathbf{j}_y$ and the fields gradients $\nabla(-\frac{V}{T})$, $\nabla(\frac{1}{T})$ are conjugated pairs. Therefore the gradients of the fields vector $\nabla\mathbf{M}(6\times 1)$ in terms of $(\nabla f_V, \nabla f_T)$ is defined by

$$(\ \nabla\mathbf{M}\ ) = \begin{pmatrix} \nabla f_V \\ \nabla f_T \end{pmatrix} = \begin{pmatrix} \nabla(-\frac{V}{T}) \\ \nabla(\frac{1}{T}) \end{pmatrix} = \begin{pmatrix} \frac{-1}{T} & \frac{V}{T^2} \\ 0 & \frac{-1}{T^2} \end{pmatrix} \begin{pmatrix} \nabla V \\ \nabla T \end{pmatrix}. \tag{14}$$

Hence, the fluxes defined by Eq. (8) can be expressed in terms of f_V, f_T, yielding

$$\mathbf{j} \ = \begin{pmatrix} \mathbf{l}T & V T\mathbf{l} + \alpha T^2\mathbf{l} \\ V T\mathbf{l} + \alpha T^2\mathbf{l} & T^2\mathbf{k} + 2\alpha T^2 V\mathbf{l} + \alpha^2 T^3\mathbf{l} + T V^2\mathbf{l} \end{pmatrix} \begin{pmatrix} \nabla f_V \\ \nabla f_T \end{pmatrix} = \ \mathbf{Z} \ \ \nabla\mathbf{M}. \tag{15}$$

The fluxes vector $\mathbf{j}(6\times 1)$ is the product of the gradients fields vector $\nabla\mathbf{M}$, which derive from the state variables (f_V, f_T), by a coefficients matrix $\mathbf{Z}(6\times 6)$ which is temperature and electric

potential dependent.
The symmetric coefficients matrix $\mathbf{Z}(\mathrm{V},\mathrm{T})$ in Eq. (15), which is positive definite if $\mathrm{T} > 0$, can be rewritten in terms of $(\mathrm{f_V},\mathrm{f_T}) = (-\frac{\mathrm{V}}{\mathrm{T}},\frac{1}{\mathrm{T}})$, as $\mathrm{T} = \frac{1}{\mathrm{f_T}}, \mathrm{V} = -\frac{\mathrm{f_V}}{\mathrm{f_T}}$:

$$\mathbf{Z}(\mathrm{f_V},\mathrm{f_T}) = \begin{pmatrix} \frac{1}{\mathrm{f_T}}\mathbf{1} & -\frac{\mathrm{f_V}}{\mathrm{f_T^2}}\mathbf{1} + \alpha\frac{1}{\mathrm{f_T^2}}\mathbf{1} \\ -\frac{\mathrm{f_V}}{\mathrm{f_T^2}}\mathbf{1} + \alpha\frac{1}{\mathrm{f_T^2}}\mathbf{1} & \frac{\mathbf{k}}{\mathrm{f_T^2}} - 2\alpha\frac{\mathrm{f_V}}{\mathrm{f_T^3}}\mathbf{1} + \alpha^2\frac{1}{\mathrm{f_T^3}}\mathbf{1} + \frac{\mathrm{f_V^2}}{\mathrm{f_T^3}}\mathbf{1} \end{pmatrix}. \tag{16}$$

Therefore, the strong form (8, 13) can be expressed as

$$\begin{aligned} \mathrm{div}(\mathbf{j}) &= \mathbf{i} && \forall\, \mathbf{x} \in \Omega, \\ \mathbf{M} &= \bar{\mathbf{M}} && \forall\, \mathbf{x} \in \partial_{\mathrm{D}}\Omega, \\ \bar{\mathbf{n}}\,\mathbf{j} &= \bar{\mathbf{j}} && \forall\, \mathbf{x} \in \partial_{\mathrm{N}}\Omega, \end{aligned} \tag{17}$$

where $\bar{\mathbf{n}} = \begin{pmatrix} \mathbf{n} & 0 \\ 0 & \mathbf{n} \end{pmatrix}, \bar{\mathbf{M}} = \begin{pmatrix} \bar{\mathrm{f}}_{\mathrm{V}} \\ \bar{\mathrm{f}}_{\mathrm{T}} \end{pmatrix} \in \mathrm{L}^2(\partial_{\mathrm{D}}\Omega) \times \mathrm{L}^{2^+}(\partial_{\mathrm{D}}\Omega)$, meaning that $\bar{\mathrm{f}}_{\mathrm{T}}$ belongs to the space which is always positive, and $\bar{\mathbf{j}} = \begin{pmatrix} \bar{\mathrm{j}}_{\mathrm{e}} \\ \bar{\mathrm{j}}_{\mathrm{y}} \end{pmatrix}$.

3 Weak Discontinuous Galerkin (DG) form for electro-thermal coupled problems

Let us introduce the finite element space $\Omega_{\mathrm{h}} = \cup_{\mathrm{e}}\Omega^{\mathrm{e}}$, associated with the triangulation of the domain Ω. If subscript I denotes the boundary between two elements, $\partial\Omega_{\mathrm{e}} = \partial_{\mathrm{N}}\Omega^{\mathrm{e}} \cup \partial_{\mathrm{D}}\Omega^{\mathrm{e}} \cup \partial_{\mathrm{I}}\Omega^{\mathrm{e}}$, and $\partial_{\mathrm{I}}\Omega_{\mathrm{h}} = \cup_{\mathrm{e}}\partial_{\mathrm{I}}\Omega^{\mathrm{e}} \setminus \partial\Omega_{\mathrm{h}}$, where $\partial_{\mathrm{I}}\Omega_{\mathrm{h}}$ is the intersecting boundary of the finite elements. The finite discontinuous polynomial approximation $\mathbf{M}_{\mathrm{h}} = \begin{pmatrix} \mathrm{f}_{\mathrm{V_h}} \\ \mathrm{f}_{\mathrm{T_h}} \end{pmatrix} \in \mathrm{X}^{\mathrm{k}}$ of the solution is thus defined in the space

$$\mathrm{X}^{\mathrm{k}} = \left\{ \mathbf{M}_{\mathrm{h}} \in \mathrm{L}^2(\Omega_{\mathrm{h}}) \times \mathrm{L}^{2^+}(\Omega_{\mathrm{h}}) \mid_{\mathbf{M}_{\mathrm{h}}|_{\Omega^{\mathrm{e}}} \in \mathbb{P}^{\mathrm{k}}(\Omega^{\mathrm{e}}) \times \mathbb{P}^{\mathrm{k}^+}(\Omega^{\mathrm{e}})\ \forall \Omega^{\mathrm{e}} \in \Omega_{\mathrm{h}}} \right\}, \tag{18}$$

where $\mathbb{P}^{\mathrm{k}}(\Omega^{\mathrm{e}})$ is the space of polynomial functions of order up to k and $\mathbb{P}^{\mathrm{k}^+}$ means that the polynomial approximation remains positive. Moreover we can define the kinematically admissible counterpart:

$$\mathrm{X}_{\mathrm{c}}^{\mathrm{k}} = \left\{ \delta\mathbf{M}_{\mathrm{h}} \in \mathrm{X}^{\mathrm{k}} \mid_{\delta\mathbf{M}_{\mathrm{h}} = \mathrm{I_h}\bar{\mathbf{M}}|_{\partial_{\mathrm{D}}\Omega^{\mathrm{e}}}} \right\}, \tag{19}$$

where $\mathrm{I_h}\bar{\mathbf{M}}$ is the interpolation of $\bar{\mathbf{M}}$ on $\partial_{\mathrm{D}}\Omega_{\mathrm{h}}$.
Let us derive the weak form of the governing equations by multiplying the first equation by a kinematically admissible test function $\delta\mathbf{M} \in \mathrm{X}_{\mathrm{c}}^{\mathrm{k}}$. Then performing a volume integral and using the divergence theorem on each element Ω^{e}, lead to state the problem as finding $\mathbf{M}_{\mathrm{h}} = \begin{pmatrix} \mathrm{f}_{\mathrm{V_h}} \\ \mathrm{f}_{\mathrm{T_h}} \end{pmatrix} \in \mathrm{X}^{\mathrm{k}}$ such that

$$-\sum_{\mathrm{e}} \int_{\Omega^{\mathrm{e}}} \nabla\delta\mathbf{M}_{\mathrm{h}}^{\mathrm{T}}\mathbf{j}(\mathbf{M}_{\mathrm{h}}, \nabla\mathbf{M}_{\mathrm{h}})\, \mathrm{d}\Omega + \sum_{\mathrm{e}} \int_{\partial\Omega^{\mathrm{e}}} \mathbf{n}^{\mathrm{T}}\delta\mathbf{M}_{\mathrm{h}}^{\mathrm{T}}\mathbf{j}(\mathbf{M}_{\mathrm{h}}, \nabla\mathbf{M}_{\mathrm{h}})\, \mathrm{dS} = 0 \quad \forall\delta\mathbf{M}_{\mathrm{h}} \in \mathrm{X}_{\mathrm{c}}^{\mathrm{k}}. \tag{20}$$

Let us define suitable face operators which are required for the definition of the proceeding DG. For two adjacent elements $+$, $-$, who share a common face $\partial_{\mathrm{I}}\Omega^{\mathrm{s}}$, the averages of $\mathbf{M}$, $\mathbf{j}_{\mathrm{e}}$ are given respectively by

$$\langle\mathbf{M}\rangle = \frac{1}{2}(\mathbf{M}^+ + \mathbf{M}^-), \qquad \langle\mathbf{j}_{\mathrm{e}}\rangle = \frac{1}{2}(\mathbf{j}_{\mathrm{e}}^+ + \mathbf{j}_{\mathrm{e}}^-),$$

and the jumps are defined respectively as

$$[\![\mathbf{M}]\!] = \frac{1}{2}(\mathbf{M}^+ - \mathbf{M}^-), \qquad [\![\mathbf{j}_e]\!] = \frac{1}{2}(\mathbf{j}_e^+ - \mathbf{j}_e^-).$$

For simplicity, we introduce the vector $\mathbf{M}_{\mathbf{n}} = \begin{pmatrix} \mathbf{n}^- & 0 \\ 0 & \mathbf{n}^- \end{pmatrix} \mathbf{M}$, where $\mathbf{n}^-$ is the unit normal vector pointing from element $-$ toward element $+$ and $\bar{\mathbf{M}}_{h_{\mathbf{n}}} = \bar{\mathbf{n}} I_h \bar{\mathbf{M}}$.
Then the surface integral of Eq. (20) is rewritten using the boundary condition stated in Eq. (17) and the notation introduced above, leading to

$$\int_{\partial_N\Omega_h} \delta\mathbf{M}_h^T \bar{\mathbf{j}}\, dS + \int_{\partial_D\Omega_h} \bar{\mathbf{M}}_{h_{\mathbf{n}}}^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h)\, dS - \int_{\Omega_h} \mathbf{i}\delta\mathbf{M}_h^T d\Omega = \int_{\Omega_h} \nabla\delta\mathbf{M}_h^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h) d\Omega + \int_{\partial_I\Omega_h} [\![\delta\mathbf{M}_{h_{\mathbf{n}}}^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h)]\!]\, dS \quad \forall\delta\mathbf{M}_h \in X_c^k. \tag{21}$$

Applying the mathematical identity $[\![ab]\!] = [\![a]\!]\langle b\rangle + [\![b]\!]\langle a\rangle$, and neglecting the second term because only consistency in $\delta\mathbf{M}_h^T$ needs to be enforced, the interface flux related to Eq. (21) becomes$[\![\delta\mathbf{M}_{h_{\mathbf{n}}}]\!]\langle\mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h)\rangle$.
Due to the discontinuous nature of the approximation functions in the DG finite element, interelement discontinuity is allowed, so the continuity of unknown variables in the DG formulation is enforced weakly by adding symmetrization and stabilization terms at the interior element boundary $\partial_I\Omega_h$. Using Eq. (15), the stabilized form can be stated as finding $\mathbf{M}_h \in X^k$ such that

$$\begin{aligned} &\int_{\partial_N\Omega_h} \delta\mathbf{M}_h^T \bar{\mathbf{j}} d\Omega + \int_{\partial_D\Omega_h} \bar{\mathbf{M}}_{h_{\mathbf{n}}}^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h) ds = \int_{\Omega_h} \nabla\delta\mathbf{M}^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h) d\Omega \\ &+ \int_{\partial_I\Omega_h} [\![\delta\mathbf{M}_{h_{\mathbf{n}}}^T]\!] \langle\mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h)\rangle\, dS + \int_{\partial_I\Omega_h} [\![\mathbf{M}_{h_{\mathbf{n}}}^T]\!] \langle\mathbf{Z}(\mathbf{M}_h)\nabla\delta\mathbf{M}_h\rangle\, dS \\ &+ \int_{\partial_I\Omega_h} [\![\delta\mathbf{M}_{h_{\mathbf{n}}}^T]\!] \left\langle \frac{\mathcal{B}}{h_s}\mathbf{Z}(\mathbf{M}_h)\right\rangle [\![\mathbf{M}_{h_{\mathbf{n}}}]\!]\, dS \quad \forall\delta\mathbf{M}_h \in X_c^k. \end{aligned} \tag{22}$$

In this equation $\mathcal{B}$ is the stability parameter, which has to be sufficiently high to guarantee stability, and h_s is the mesh size.
[1].
Then, the IP discontinuous Galerkin method corresponding to problem Eq. (17) is defined as: find $\mathbf{M}_h \in X^k$ such that

$$a_3(\mathbf{M}_h, \delta\mathbf{M}_h) = b_3(\delta\mathbf{M}_h) - \int_{\Omega_h} \mathbf{i}\delta\mathbf{M}_h^T d\Omega + c_3(\mathbf{M}_h; \bar{\mathbf{M}}_h) \quad \forall\delta\mathbf{M}_h \in X_c^k, \tag{23}$$

$$\begin{aligned} a_3(\mathbf{M}_h, \delta\mathbf{M}_h) \quad &= \int_{\Omega_h} \nabla\delta\mathbf{M}_h^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h) d\Omega + \int_{\partial_I\Omega_h} [\![\delta\mathbf{M}_{h_{\mathbf{n}}}^T]\!] \langle\mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h)\rangle\, dS \\ &+ \int_{\partial_I\Omega_h} [\![\mathbf{M}_{h_{\mathbf{n}}}^T]\!] \langle\mathbf{Z}(\mathbf{M}_h)\nabla\delta\mathbf{M}_h\rangle\, dS + \int_{\partial_I\Omega_h} [\![\delta\mathbf{M}_{h_{\mathbf{n}}}^T]\!] \left\langle \frac{\mathcal{B}}{h_s}\mathbf{Z}(\mathbf{M}_h)\right\rangle [\![\mathbf{M}_{h_{\mathbf{n}}}]\!]\, dS \\ &\forall\delta\mathbf{M}_h \in X_c^k, \end{aligned} \tag{24}$$

$$c_3(\mathbf{M}_h; \bar{\mathbf{M}}_h) = \int_{\partial_D\Omega_h} \bar{\mathbf{M}}_{h_{\mathbf{n}}}^T \mathbf{j}(\mathbf{M}_h, \nabla\mathbf{M}_h) ds, \text{ and} \tag{25}$$

$$b_3(\delta\mathbf{M}_h) = \int_{\partial_N\Omega_h} \delta\mathbf{M}_h^T \bar{\mathbf{j}} d\Omega. \tag{26}$$

The demonstration of numerical properties such as the optimal error estimate, stability of the formulation and uniqueness of the solution for β high enough is reported in [1].

4 Numerical Results

The studied problem involves the production of heat by supplying electric power. The boundary conditions shown in Fig. 1 are applied on a composite material consisting of a combination of two materials: matrix (e.g., polymer) which is a non-conductive material and conductive fillers (e.g., carbon fiber).

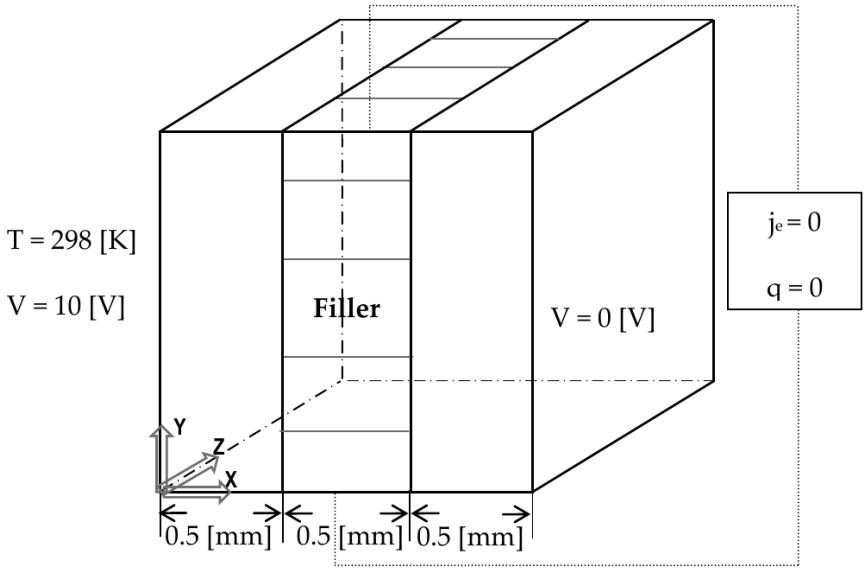


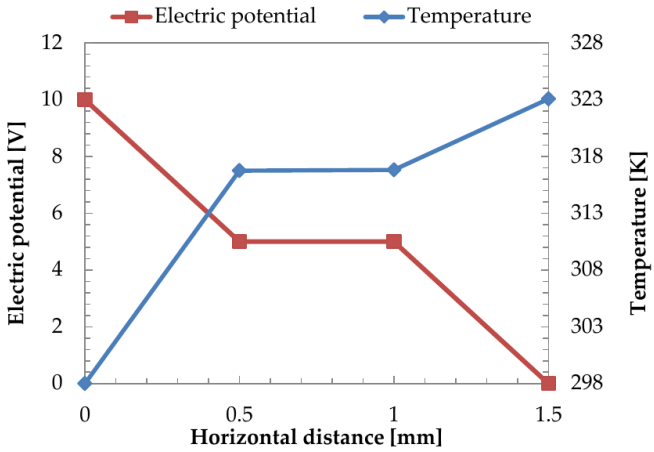
Figure 1: Electrothermal composite domain and the boundary conditions.

A finite element mesh consisting of 114 quadratic elements is used, and a stabilization parameter $\beta = 100$ is considered, to solve the DG discretization. The different material parameters used in this test are listed in Table 1, where all the material properties are assumed to be independent of temperature including the Seebeck coefficient.

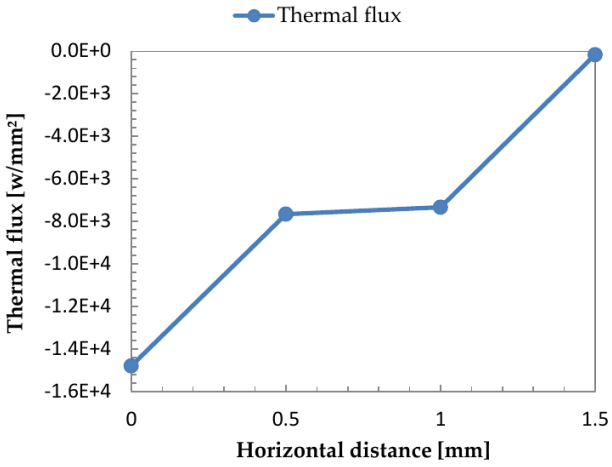
Material	$\mathbf{l}[\mathrm{S/m}]$	$\mathbf{k}[\mathrm{W/(K \cdot m)}]$	$\alpha[\mathrm{V/K}]$
Carbon fiber	diag(100000)	diag(40)	3×10^{-6}
Polymer	diag(0.1)	diag(0.2)	3×10^{-7}

Table 1: Composite material phases parameters.

Fig. 2 presents the distributions of the temperature and the of electric potential. When an electric potential of 10 [V] is applied on one side, the temperature on the other side increases from 298 [K] to 323 [K], We have also noticed that the temperature and electric potential, see Fig. 2(a), as well as the thermal flux, see Fig. 2(b), are almost constant in the fiber, as its electrical conductivity is high, and transfer gradually in the polymer matrix which is a non conductive material. Also a constant electric current density has been obtained with a value of $1.5 \times 10^3\,[\mathrm{A/m^2}]$.



(a)



(b)

Figure 2: (a) Distribution of the electrical potential and temperature in the electrothermal domain for composite materials, (b) The distribution of the thermal flux in the electrothermal domain for composite materials.

5 CONCLUSIONS

In this work

- A DG method was developed for electro-thermal coupling using energetically conjugated fluxes and fields.
- The result formulation was then applied to predict behavior of electro-thermal composite, and was able to compute the temperature and electric potential distributions as well as their fluxes, as a function of spatial coordinates.
- In the future, this formulation will be used for Electo-thermo-mechanical coupling in composite materials.

REFERENCES

[1] L. Homsi, C. Geuzaine, L. Noels, A coupled electro-thermal discontinuous Galerkin method. *Journal of Computational Physics* , submitted

[2] T. Gudi, A.K. Pani, N. Nataraj, hp-Discontinuous Galerkin methods for strongly nonlinear elliptic boundary value problems. *Numerische Mathematik*, **109**, 233-268, 2008

[3] L. Liu, A continuum theory of thermoelectric bodies and effective properties of thermoelectric composites. *International Journal of Engineering Science*, **55**, 35-53, 2015.

[4] J. L. Perez-Aparicio, R.L. Taylor, D. Gavela, Finite element analysis of nonlinear fully coupled thermoelectric materials. *Computational Mechanics*, **40**, 35-45, 2007.

[5] M.F. Wheeler, An elliptic collocation-finite element method with interior penalties. *SIAM J.Numer.Anal*, **15**, 152-161, 1978.

[6] S. Yadav, A. Pani and E. J.Park, Superconvergent discontinuous Galerkin methods for nonlinear elliptic equations. *Mathematics of Computation*, **82**, 1297-1335, 2013

[7] J. L. Pérez–Aparicio, R. Palma and R. L. Taylor, Finite element analysis and material sensitivity of Peltier thermoelectric cells coolers. *International Journal of Heat and Mass Transfer*, **55** ,1363-1374, 2012.

[8] G. D. Mahan, Density variations in thermoelectrics. *Journal of Applied Physics*,**87**, 7326-7332, 2000

[9] X. P. Zheng, D. H. Liu, and Y. Liu, Thermoelastic coupling problems caused by thermal contact resistance: A discontinuous Galerkin finite element approach. *Science China Physics, Mechanics and Astronomy*, **54**, 666-674, 2011

[10] S. Sun and M. F. Wheeler, Discontinuous Galerkin methods for coupled flow and reactive transport problems. *Applied Numerical Mathematics*,**52**, 273-298, 2005

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOWARDS LATTICE-BOLTZMANN ON DYNAMICALLY ADAPTIVE GRIDS — MINIMALLY-INVASIVE GRID EXCHANGE IN ESPRESSO

Michael Lahnert[1], Carsten Burstedde[2], Christian Holm[3], Miriam Mehl[1], Georg Rempfer[3], and Florian Weik[3]

[1]Institute of Parallel and Distributed Systems, University of Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany
email: {michael.lahnert, miriam.mehl}@ipvs.uni-stuttgart.de
[2]Institute for Numerical Simulation and Hausdorff Center for Mathematics, University of Bonn
Wegelerstraße 6, 53115 Bonn, Germany
email: burstedde@ins.uni-bonn.de
[3]Institute for Computational Physics, University of Stuttgart
Allmandring 3, 70569 Stuttgart, Germany
email: {christian.holm, georg.rempfer, florian.weik}@icp.uni-stuttgart.de

Keywords: Octree Grids, Lattice-Boltzmann, Dynamical Adaptivity, Fluid Dynamics, Coupled Soft-Matter Simulations.

Abstract. *We present the minimally-invasive exchange of the regular Cartesian grid in the lattice-Boltzmann solver of ESPResSo by a dynamically-adaptive octree grid. Octree grids are favoured by computer scientists over other grid types as they are very memory-efficient. In addition, they represent a natural generalisation of regular Cartesian grids, such that most discretisation details of a regular grid solver can be maintained. Optimised codes, however, require a special tree-oriented grid traversal, which typically conflicts with existing simulation codes using various iterators, some for only parts of the grid, e.g., boundaries. ESPResSo is a large software package developed for soft-matter simulations involving fluid flow, electrostatic, and electrokinetic effects, and molecular dynamics. The currently used regular Cartesian grid hinders the simulation of realistic domain sizes and significant time periods, a problem that can be solved using grid adaptivity. In a first step, we focus on the lattice-Boltzmann flow solver in ESPResSo.*

`p4est` is a grid framework, that already provides dynamically adaptive quadtree and octree grids together with high-level interfaces for flexible grid traversals with direct neighbour access in all grid components. In this paper, we first describe extensions of `p4est` that were necessary to fulfill certain application requirements. The second part of our work consists of the minimally-invasive changes in ESPResSo preserving the expertise accumulated in the software's implementation over years. Our numerical results demonstrate physical correctness of the implementation, good parallel scalability and low overhead of the dynamical grid adaptivity. These are prerequisites to actually profit from grid adaptivity in terms of being able to simulate larger domains over longer time periods with limited computational resources. Thus, the current status forms a solid basis for further steps such as the development of refinement criteria, the setup of more realistic application scenarios, and a GPU implementation.

1 INTRODUCTION

Many scientifically and industrially relevant soft matter systems contain hydrodynamic flow with strong gradients in small subvolumes. Most biomolecules, including DNA and proteins, are charged when solvated in aqueous solution. Any charged object in electrolytic solution is surrounded by a so-called double layer – a thin layer, typically on the scale of one to a few hundred nanometres. This double layer is the only part of the volume that is not net neutral, a state maintained by the attraction of counterions and the repulsion of coions from the charged surface. Applying an external electric field causes fluid stress in the double layer and perturbs the ionic distributions. The fluid flow resulting from this complex coupling of ion and fluid dynamics is called electroosmotic flow (EOF) in the case of a charged surface at rest, while the resulting motion of a charged particle is called electrophoresis.

Both EOF and electrophoresis are important in applications such as the separation of biomolecules and other charged particles, DNA profiling, and pumping and sample transport in microfluidic devices, to name just a few.

Simulations of EOF and electrophoresis prove challenging, due to the large discrepancy between the nanometre scale of the double layer and the system sizes, which often range up to several hundred microns. Existing numerical schemes that work on unstructured grids like the finite-element method (FEM) or the finite-volume method (FVM) are great tools to investigate stationary electrophoresis and EOF systems, but the overhead incurred through remeshing makes these methods infeasible for systems containing mobile particles.

In this paper, we present our approach to moving a large simulation application from a regular Cartesian mesh to fully adaptive tree-structured Cartesian meshes in a minimally-invasive way. The application is the molecular and continuum dynamics package ESPResSo [1, 2], where we particularly focus on the included lattice-Boltzmann solver. For scalable parallel adaptive mesh refinement, we base our work on the software library `p4est` [3, 4].

Many large simulation codes use regular Cartesian meshes due to their cheap representation in memory, the simple data layout, the easy domain partitioning, and the straight-forward discretisations. However, the missing grid adaptivity in time and space limits the size of the domain and the simulation time accessible with reasonable computation time on available machines. Grid adaptivity would allow for coarsening of the mesh in regions of minor importance and provide the possibility for space-time adaptivity, i.e., to also apply larger time steps in spatially coarser domains. Two major types of adaptive meshes are widely used in simulation codes. The first type, unstructured meshes, offers a very high flexibility in terms of local mesh resolution, shape and even type of mesh elements. Thus, even very complex geometries can be represented accurately. However, unstructured meshes come with high memory requirements, and structures and element types completely different from regular Cartesian meshes. The second type of adaptive meshes consists of tree-structured Cartesian meshes, with octrees and quadtrees being the most famous examples. These meshes form a natural generalisation of regularly refined Cartesian meshes. They are constructed by recursively refining mesh elements into a fixed number of equally sized and equally shaped child elements. They are very efficient, in particular in terms of memory requirements and memory access patterns, if represented in a linearised form according to a depth-first traversal of the underlying tree and an ordering of the children prescribed by a space-filling curve [5, 6, 7]. Depending on the aggressiveness of the encoding, however, this may impede traversal of the mesh or parts thereof in an arbitrary order. In contrast, existing application codes on regularly refined Cartesian meshes often rely on access to parts of the mesh (e.g., boundaries), particular mesh regions, or neighbours of

mesh elements. Thus, integrating a tree-structured Cartesian mesh with an existing simulation code based on regular Cartesian meshes in a minimally-invasive way calls for a mesh interface that allows for the easy implementation of user-defined iterators and direct neighbour access. The challenge in this lies in finding the optimal trade-off between minimally-invasiveness, and memory and run-time efficiency.

Our idea for making spatial adaptivity more accessible to researchers from other fields is to turn the integration of an octree-based Cartesian grid into a three-step process. In a first step the regular grid of the existing application is replaced with an octree-based one, specialised for uniform refinement. In this way we are able to reproduce results of the existing implementation using a regular grid, without having to worry about spatial adaptivity, yet. This step preserves all of the regular grid's data-access patterns, up to a permutation of the mesh elements. This permutation is induced by the octree refinement, which replaces a traditional lexicographic ordering with one governed by the space-filling curve. Reordering this way benefits data locality and cache performance of large scale simulations, while the mathematical problem being solved remains the same. In case of a traversal of subregions of the grid, e.g. for setting boundary values, we may traverse the whole grid along the space-filling curve and only take action on the cells that meet a selection criterion, possibly caching the subset indices for later reuse. Alternatively, we may access the position of an initial cell and traverse the local part of the mesh by hopping to neighbours, independently of the space-filling curve.

The next step of our three-step process consists of solving numerical issues that arise when actually making use of the spatial adaptivity. An example for such an issue is the speed of sound in the lattice-Boltzmann fluid, which depends on the element size and the time step, and is therefore no longer uniform for a locally refined mesh with a constant time step. A uniform speed of sound can be recovered by using different time steps for differently refined parts of the mesh. There are several ways to consistently implement this, e.g., using a finite-volume approach [8], compact interpolation [9], or volumetric formulation [10].

As the final step, the application should be tuned by adapting both data access patterns and grid-traversal to make full use of the locality provided by the data storage scheme.

The need to be able to traverse subregions in our grid efficiently in step one introduces two requirements for the implementation of the grid library. First, we need random access to cells, and secondly, we need to be able to compute cell positions in our linearised array (as it is given by the space-filling curve) based on geometric coordinates. We will show that those two criteria are met by the `p4est_mesh` interface.

In the remainder of the paper, we present the software package ESPResSo together with a short introduction to the implemented lattice-Boltzmann solver on regular Cartesian meshes (Sect. 2), give an overview of the adaptive mesh software `p4est` (Sect. 3), a more detailed description of our contributions to both software packages establishing the first lattice-Boltzmann prototype on tree-structured meshes with minimal changes in ESPResSo (Sect. 4), followed by first results (Sect. 5) and a conclusion (Sect. 6).

2 ESPRESSO AND THE LATTICE-BOLTZMANN METHOD

2.1 ESPResSo

ESPResSo [1, 2], the extensible simulation package for research on soft matter systems, is a molecular dynamics (MD) program that focuses on coarse grained soft matter applications. While originally focused on the simulation of charged and uncharged polymers and colloids, today it implements a wide range of algorithms [11] for electrostatics [12, 13, 14, 15, 16], di-

electrics [17, 18, 19], magnetostatics [20] and hydrodynamics [11, 21], both for periodic and non-periodic systems. Applications include, amongst others, engineering problems such as soot aggregation [22], hydrogels [23], biological membranes [24], DNA like-charge attraction [25] and translocation [26], and ionic liquids [27]. Aside from the plethora of implemented methods, the major strength of ESPResSo lies in its ability to flexibly combine different methods and particle interactions to encompass new applications and systems quickly. ESPResSo is MPI parallelized and can employ GPU acceleration for hydrodynamics [21] and electrostatics calculations.

The regular mesh refinement currently used for continuum methods in ESPResSo hinders the efficient exploitation of different scales in different regions of the simulation volume, such that the generalisation to adaptively refined meshes is crucial for accurate simulations encompassing experimentally relevant time scales.

2.2 Lattice-Boltzmann

Lattice-Boltzmann (LB) is a method for solving the Navier-Stokes equations and similar partial differential equations, which was first proposed in the context of lattice gas automata. It was later discovered that the lattice-Boltzmann equation can also formally be derived as a discrete lattice approximation to the continuum Boltzmann equation for fluids, providing a sound mathematical basis for the method [28]. Further analysis based on the Chapman-Enskog expansion shows that the LB equation is equivalent to the full Navier-Stokes equations in the limit of low Mach numbers. Dünweg and Ladd gave a comprehensive analysis of the LB method and many of its extensions relevant to soft matter simulations in 2009 [29]. A good overview of the state of the development and its possible applications were given by Succi, Sbragaglia, and Ubertini in 2010 [30].

The LB equation is a discretisation of the Boltzmann equation. The Boltzmann equation is defined in the seven-dimensional space with three spatial coordinates, three velocity or momentum components and time as independent variables and describes the time evolution of the microscopic particle distribution. The lattice-Boltzmann method discretises the velocity space with a finite number of discrete velocities $\boldsymbol{c}_i$ and finite differences on a cubic grid in the three-dimensional coordinate space. This results in the lattice-Boltzmann equation (LBE)

$$f_i(\boldsymbol{r} + \boldsymbol{c}_i \Delta t, t + \Delta t) = f_i(\boldsymbol{r}, t) + \sum_k L_{ik} \left\{ f_k^{\mathrm{eq}} - f_k(\boldsymbol{r}, t) \right\} + \Psi_i \,, \tag{1}$$

where f_i denotes the densities of virtual particle populations assuming the velocity $\boldsymbol{c}_i$, and Ψ_i models the density change of the population f_i due to external forces. The discrete velocities are chosen such that they transport particles from one lattice node to one of its neighbours in one time step. Fig. 1 shows the widely used D3Q19 discretisation for the velocities with 19 velocity components associated to the centre of a three-dimensional cubic grid cell. We assume linear relaxation towards the equilibrium distributions f_i^{eq}, according to the linear operator L_{ik}.

The macroscopic values for the density ρ, momentum $\boldsymbol{p}$, and stress $\overline{\boldsymbol{\pi}}$ can be recovered from the microscopic particle populations through

$$\rho = \sum_i f_i, \qquad \boldsymbol{p} = \sum_i f_i \boldsymbol{c}_i, \qquad \overline{\boldsymbol{\pi}} = \sum_i f_i \boldsymbol{c}_i \otimes \boldsymbol{c}_i, \tag{2}$$

where $\otimes$ denotes the tensor product. Using those, the lattice equilibrium distribution is given

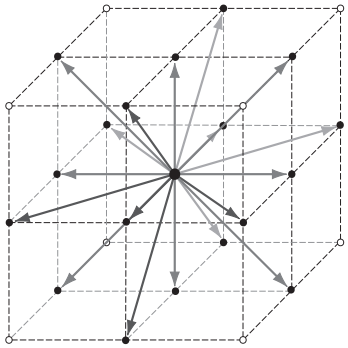


Figure 1: One cell from a D3Q19 LB grid, depicting the 19 discrete velocities, connecting the node with itself, its nearest, and next nearest neighbours. Image adapted from [31].

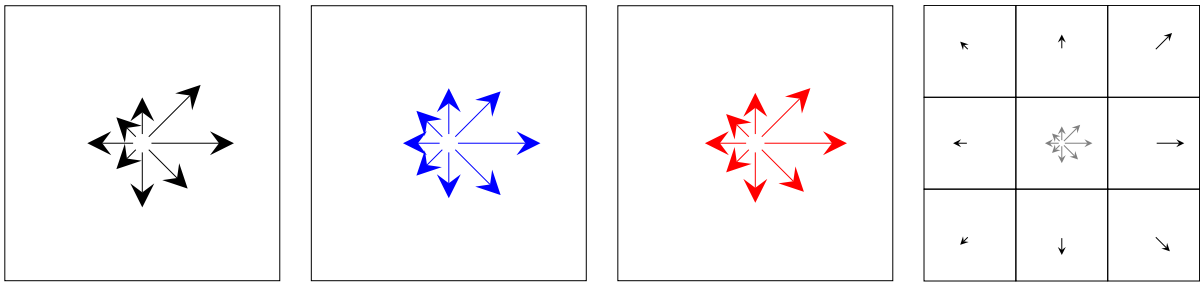


Figure 2: Schematic view of the collision and the streaming step. The arrows indicate the discrete velocities $\boldsymbol{c}_i$ in a two-dimensional quadratic grid cell scaled with the respective densities f_i. (a) – (c): collision step. Left to right: pre-collision populations, virtual local equilibrium, post-collision populations. (d): streaming step. Grey arrows indicate the pre-streaming population in the centre cell, black arrows the same populations transported to the respective neighbour cells, i.e., after the streaming step.

by a second order expansion of the continuum Maxwell-Boltzmann equilibrium distribution

$$f_i^{\mathrm{eq}} = a_i \rho \left(1 + \frac{\boldsymbol{v} \cdot \boldsymbol{c}_i}{c_s^2} + \frac{(\boldsymbol{v} \cdot \boldsymbol{c}_i)^2}{2c_s^4} - \frac{v^2}{2c_s^2} \right) , \tag{3}$$

with prefactors a_i and the speed of sound c_s depending on the specific lattice geometry.

Algorithmically, performing a time step for the LBE (1) consists of two parts with different requirements in terms of data access: Whereas the collision is a local operation modelling the interaction of densities of different velocity within a local cell described by the operator L_{ik}, the streaming step transports densities into neighbouring cells according to the direction of the respective velocity $\boldsymbol{c}_i$. Fig. 2 schematically shows the collision step as a purely cell-local operation on population densities and the equilibrium densities. In addition, it displays the streaming as a transport of data from one cell to its neighbour cells.

Collision. ESPResSo implicitly implements different versions of the collision operator: The simplest choice for the collision operator $L_{ik} = \lambda \delta_{ik}$ is the so-called BGK collision operator (after the inventors Bhatnagar, Gross, and Krook) using only a single relaxation parameter λ for all populations, which can be chosen such that the fluid assumes the desired shear viscosity. Other fluid properties such as the bulk viscosity cannot be chosen independently. The full multi-relaxation time LB method (MRT-LB) results in less limitations. Here one relaxes different linear combinations m_i of the velocity populations f_i towards their equilibrium values using

different relaxation rates λ_i. These linear combinations correspond to the different macroscopic quantities. The MRT-LB forms a superset of the simpler collision operators, i.e. by choosing the λ_i accordingly, the MRT collision operator can reproduce the simpler BGK and TRT (two relaxation time) collision operators [32].

Streaming. In terms of streaming, it is important to note that, due to the requirement that virtual particle populations are transported exactly from one cell midpoint to the midpoint of the respective neighbour cell, the size of the time step Δt is proportional to the side length of the grid cells. This observation plays a major role in adaptively refined grids in Sect. 4.

By identifying nodes on the boundaries of the cubic grid with the corresponding nodes on the opposite side, one can trivially implement periodic boundary conditions in the streaming step. Other boundary conditions, such as fixed velocity or fixed pressure are more complex to implement. For simplicity, we will only discuss no-slip boundary conditions here. Further information can be found in [33]. One simple way to realise no-slip boundary conditions is by introducing so-called bounce-back nodes in the grid. Instead of streaming populations along their respective velocities, they reverse the velocities of the populations and transfer them back to the originating node. This procedure mimics a fluid flow on the opposite side of the boundary, which is antisymmetric to the real one. Due to the antisymmetry, the flow velocity on the boundary needs to vanish.

2.3 Lattice-Boltzmann in ESPResSo

While ESPResSo is primarily a molecular dynamics code, it includes an implementation of the multi-relaxation time D3Q19 lattice-Boltzmann method (MRT LB) [34], to solve the fluctuating Navier-Stokes equations. This lattice-Boltzmann fluid can be point-coupled to MD particles using the method published by Dünweg and Ladd [29]. The LB is MPI parallelized, which is realised by domain decomposition and halo communication at the node boundaries.

3 PARALLEL ADAPTIVE TREE-STRUCTURED GRIDS AND P4EST

3.1 Parallel Tree-Structured Grids

Tree-structured Cartesian grids are a natural generalisation of regular Cartesian grids that have two important advantages over unstructured meshes: 1) The shape of the grid elements is the same as in a regular Cartesian mesh. Thus, the underlying discretisation of a solver operating on a regular Cartesian mesh can be generalised to this type of adaptive grid by locally incorporating interpolation and projection on hanging nodes and/or faces [35, 36, 37]. 2) The grids are strictly structured, which allows for a very cheap representation in memory and for efficient grid partitioning. The construction principle of a tree-structured grid is the recursive refinement of grid cells into a fixed number of children of equal size and shape. If the refinement is performed globally, we get a regular Cartesian mesh, if we refine only locally in certain marked cells, we get an adaptively refined grid. Figure 3 shows a pair of two-dimensional quadtree grids (refinement of each quad into four subquads) and their respective cell trees.

In order to provide a unique ordering of grid cells on all levels, discrete iterates of space-filling-curves are widely used. There are space-filling curves for two-dimensional and for three-dimensional tree-structured grids [38] (in fact, many generalise to arbitrary dimension) and for different refinement strategies. Figure 4 shows the iterate of a Peano-curve in a two-dimensional grid with a recursive refinement of grid cells into nine children, a Hilbert-curve iterate [39] in a quadtree grid, and the Z-curve or Morton-order [40] in a quadtree grid.

The tree-structure in combination with a space-filling curve iterate allows for a very efficient

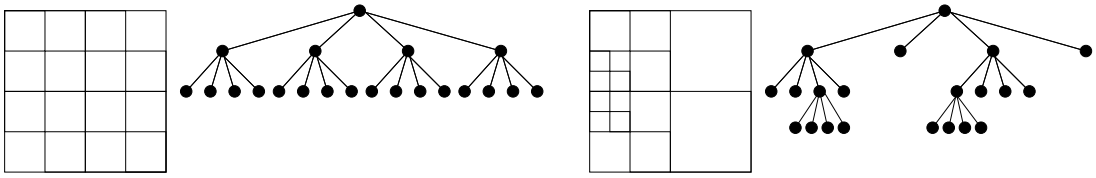


Figure 3: Left: a regularly refined quadtree grid of refinement level two. Right: an adaptively refined quadtree grid of minimal refinement level one and maximal refinement level three. In both cases, we also display the corresponding tree structure.

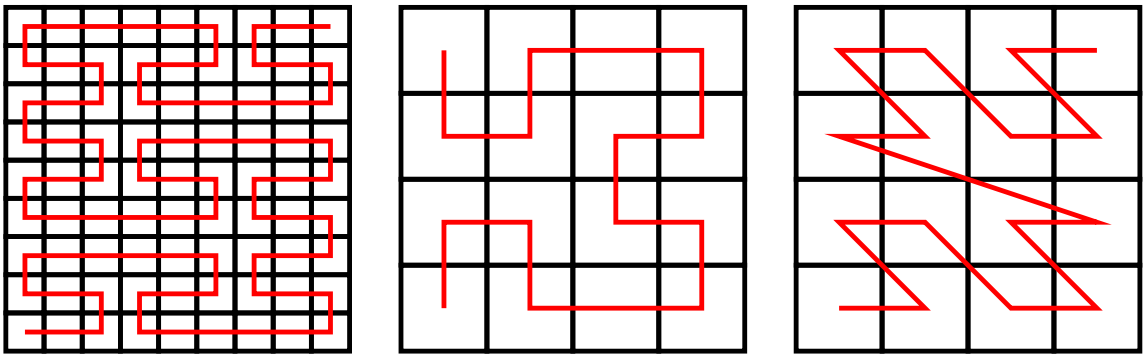


Figure 4: Some exemplary space-filling curve iterates on grids with regular refinement of level two. Left: Peano-curve; middle: Hilbert-curve; right: Z-curve or Morton-order.

linearised bit-code representation of the whole grid. This representation follows a depth-first grid traversal with an ordering of children according to the chosen space-filling curve. A '1' indicates a refined cell, whereas a '0' represents an unrefined cell. Figure 5 shows a two-dimensional adaptive quadtree mesh ordered according to the Z-curve together with the cell tree and the linearised representation.

In scientific simulation software, the combination of tree-structured grids and space-filling curves has been used in several ways, for example augmented by hashing [41], or for partial differential equation solvers with cache-optimised data administration [42, 5]. `Octor` [43] and `Dendro` [6] are two examples of parallel octree libraries that have been scaled to 62,000 [44] and 32,000 [45] cores, operating on parent-child pointers and a linearised octant storage, respectively.

3.2 The `p4est` Software

The `p4est` software library [3] extends the linearised storage strategy to a forest of interconnected octrees [46, 47]. Figure 6 shows a forest of two trees as an example. `p4est` provides

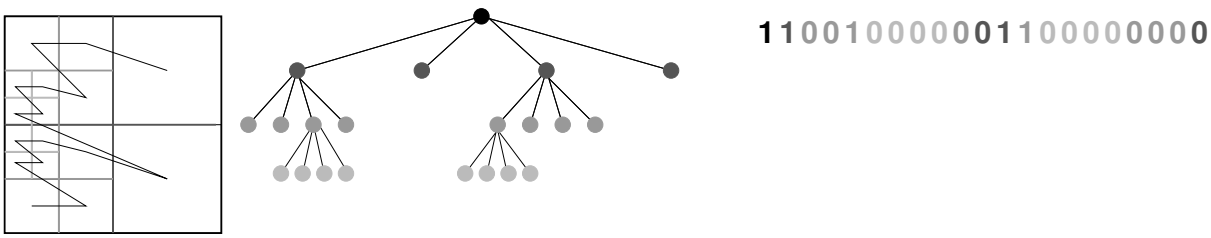


Figure 5: Example for an adaptively refined quadtree grid with cell-ordering according to the Z-curve, the corresponding cell-tree and the linearised bit-code.

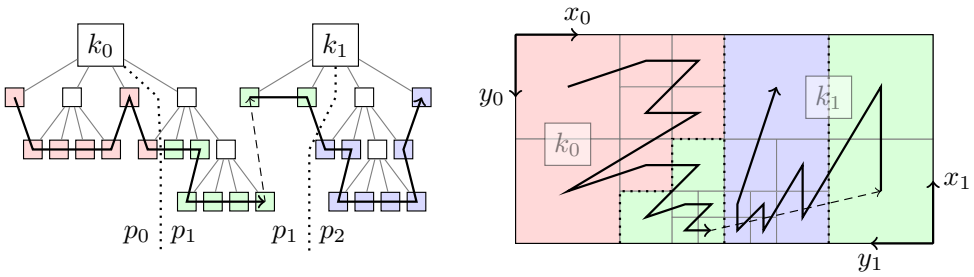


Figure 6: An example 2D forest of two adaptive quadtrees. It is partitioned into three processors (colour-coded). Note that the partition boundary does not always coincide with a tree boundary and that the coordinate systems of the trees can be rotated or mirrored relative to each other. This mesh is 2:1 balanced: Neighbour quadrants differ in size by at most a factor of two. We use the terms quadrant/octant, leaf, and element as synonyms, since they are logically equivalent.

a publicly available implementation of parallel algorithms required to create and dynamically refine, coarsen, and repartition an adaptive mesh. Furthermore, `p4est` includes an efficient 2:1 balancing algorithm [48]. In our context, 2:1 balancing means that the refinement is graded in such a way that it has a maximal level difference of one between adjacent grid cells. This feature limits the number of possible configurations of different-sized neighbours that can occur with arbitrary adaptation. Some applications are designed not to use this feature [49], while we exploit it in the lattice-Boltzmann code to maximise the simplicity of discretisation and time stepping.

`p4est` manages the dynamic-adaptive connectivity of the mesh elements. It is a C library using the MPI standard for parallelisation. The "root" or "coarse" mesh is a collection of tree roots, each identified with its cubical subvolume of the domain, and augmented by the connectivity relation between neighbouring trees. For many applications, it is sufficient to call builtin functions that create a one-tree unit cube or a rectangular brick of a small number of trees. The initial refinement of the forest is created by calling the New function with a given coarse mesh and an initial level to create a uniform distributed mesh. The refinement and coarsening functions can then be invoked any number of times to locally swap mesh elements with their children and vice versa.

`p4est` uses the Morton curve to order the elements in each tree; see Figures 4 (right) and 6. The code stores just the leaves of the forest by their lower left corner and refinement level and operates on the mesh by exploiting the mathematical structure implicit in the curve. Refinement and coarsening are local in the sense that a parent occupies the same segment in a curve as its four (2D) or eight (3D) child elements. The Partition algorithm exploits this property to shift the elements between processors without changing their order. Optionally, each element may be assigned a weight to account for a varying computational load throughout the mesh.

The basic grid traversal component of `p4est` is the `p4est_iterate` routine [50] that visits all interfaces of local mesh cells across all co-dimensions. The user may define callback functions separately for each co-dimension. Effectively, each mesh entity (cell, face, edge, or corner) is visited exactly once, and the callback function is informed with a list of the mesh cells adjacent to the visited entity. In this sense, the iterator provides a view of both the primal and the dual mesh local to a process. Calling `p4est_iterate` is the most general and lightweight way to expose the mesh connectivity to applications, formulated exclusively in `p4est`'s encoding of mesh entities and their relative orientations.

Existing simulation codes often require direct (random) access to neighbours of cells or cor-

ners, ideally in constant time. To meet this need, p4est provides the so-called p4est_mesh that we will shortly describe in more detail. In fact, we have chosen it as the central component for the integration of p4est with ESPResSo.

The p4est_mesh data structure consists of several lookup tables that store neighbourhood information for each grid cell as well as level-wise cell lists. It covers all process-local cells and the direct off-processor neighbour cells (ghosts). Thus, this module provides a way to traverse the local partition of the grid independently of the space-filling curve and allows for the evaluation of arbitrary stencils. However, this flexibility comes at the cost of additional run-time (it is created by calling p4est_iterate internally) and memory (the lookup tables are allocated), cf. section 5.3. This overhead can be amortised if the mesh structure is queried repeatedly, since each lookup only requires $\mathcal{O}(1)$ time. Information on a neighbouring cell is encoded in two integers: The first stores the index of the neighbour cell in the global cell list of the corresponding processor, the second stores the size and relative orientation of the neighbour cell. The latter is important if the neighbour cell belongs to a different quadtree or octree, which might have a different orientation. Due to the 2:1 balancing that we invoke, there are only three different cases: the neighbour of a cell can have the same size or be either twice or half as big. In the latter case, we store the indices of all smaller neighbours in the respective direction.

In summary, p4est provides a concise set of parallel AMR algorithms that are tailored to dynamic-adaptive PDE solution, but can also be used more generally whenever an efficient parallel space partition is required. p4est has been shown to scale to over 458k cores [50], with applications using it successfully on 1.57M cores [51] and 3.14M cores [52].

4 LBM ON SPATIALLY-ADAPTIVE GRIDS

In this section, we present the exchange of the regular Cartesian grid in ESPResSo by a dynamically-adaptive tree-structured grid provided by p4est. In Section 4.1, the underlying ideas of the lattice-Boltzmann solver on tree-structured grids are introduced. This includes, in particular, the realisation of collision and streaming, as well as the adaptive time stepping at boundaries between different spatial refinement levels of the octree grid. These discretisation details, together with the velocity discretisation scheme, define the neighbour cells that need to be made accessible in the p4est_mesh. The adaptations to p4est, including the generation of the p4est_mesh for our adaptive lattice-Boltzmann implementation is outlined in Section 4.2, whereas Section 4.3 sketches the changes in ESPResSo that were required for the replacement of the original grid implementation by the p4est grid.

4.1 Adaptive Formulation of the LBM

In order to cope with different mesh widths over the computational domain imposed by spatial adaptivity, the classical algorithm of the LBM needs to be extended. We use the collision operator independent scheme proposed by Rohde, Kandhai, Derksen, and van den Akker in 2006 [53]. The scheme proposes a volumetric formulation of the LBM, i.e., the nodes containing the populations are associated to the cell centres.

The algorithm consists of several steps that are shown schematically in Fig. 7 for a part of the mesh consisting of a coarse grid cell and a further refined neighbour cell with four children. Outward facing arrows indicate streaming, inward facing arrows collision. The steps of the algorithm result from two basic observations: 1) Due to the proportionality of the time step to the cell width, we have to perform two time steps in the smaller cells, while performing only one (twice as large) time step in the coarser cell. We will use the optional feature of

	coarse	fine
Mesh width Δx	Δx_c	$\Delta x_f = 0.5 \cdot \Delta x_c$
Time step Δt	Δt_c	$\Delta t_f = 0.5 \cdot \Delta t_c$
Lattice velocities c_i	$c_{i,c}$	$c_{i,f} = c_{i,c}$
Viscosity ν	ν_c	$\nu_f = \nu_c$
Relaxation parameter λ (BGK)	λ_c	$\lambda_f = 2.5 \cdot \lambda_c + 0.5$
Velocity u	u_c	u_f
External acceleration/force g	g_c	$g_f = 0.5 \cdot g_c$

Table 1: Conversion of parameters and observables between fine and coarse cells for 2:1 balanced grids according to [53].

`p4est_partition` to assign different weights to the leaves for load balancing. 2) Streaming populations from the finer grid cells into the coarse cell and vice versa requires populations associated to a common grid level. Thus, we have to interpolate populations of the coarse cell to virtual child cell populations before the streaming step and restrict the results back to the coarse cell after the second time step of the fine cells. Additionally, some fluid properties differ between different discretisation widths as shown in Table 1.

Step 1 in Fig. 7 depicts the collision on both coarse and fine cells for the discussed example situation. After this collision, we provide streaming partners for the finer cells by copying the coarse cell populations to four (2D) or eight (3D) virtual child cells (step 2). Then a streaming step can be performed on both the virtual and the original fine grid cells (step 3). Note that in the virtual fine cells, only populations associated to velocities pointing towards the original fine grid cells actually have to be streamed. In step 4a, the second collision step in the original fine cells is executed, followed by the second fine grid streaming step (step 4b). Finally, we transfer the virtual fine grid populations back to the coarse grid by taking the arithmetic mean, and finish the coarse grid time step with the streaming of the coarse grid populations whose velocities point towards other coarse cells. Note that, in a more general case without 2:1 balancing, we would have to perform more substeps on the fine grid cells before returning to the coarse level.

The generalisation of the presented adaptive lattice-Boltzmann algorithm to three-dimensional 2:1 balanced octree grids is, in principle, straight-forward. If we summarise the requirements for the octree mesh stemming from the adaptive LBM, we can state that the `p4est_mesh` has to provide direct access to the 18 neighbour cells corresponding to the 19 velocity directions minus the zero velocity in the D3Q19 LBM (compare Fig. 1). In addition, we have to implement the realisation of the virtual fine grid cells in `p4est`, a feature that was not available previously.

4.2 Preparing `p4est` for adaptive lattice-Boltzmann

As mentioned above, two aspects in `p4est` are crucial for a successful and minimally-invasive integration of the `p4est` grid in the lattice-Boltzmann solver in ESPResSo: the provision of the required neighbour cells in `p4est_mesh` and the implementation of virtual children of grid cells neighbouring a further refined region. In this section, we describe the generation of the `p4est_mesh` in more detail and briefly sketch the implementation of virtual refinement.

`p4est_mesh`. The first issue in generating the `p4est_mesh` (during a grid traversal using `p4est_iterate`) is to identify the neighbours actually required by the LBM among the information that can be generated with `p4est` callback functions for neighbour detection of faces, edges and corners. The iterator provides a top-view on each element of each co-dimension, i.e., two adjacent cells for faces, four for edges, and eight for corners. In an adaptive mesh, the in-

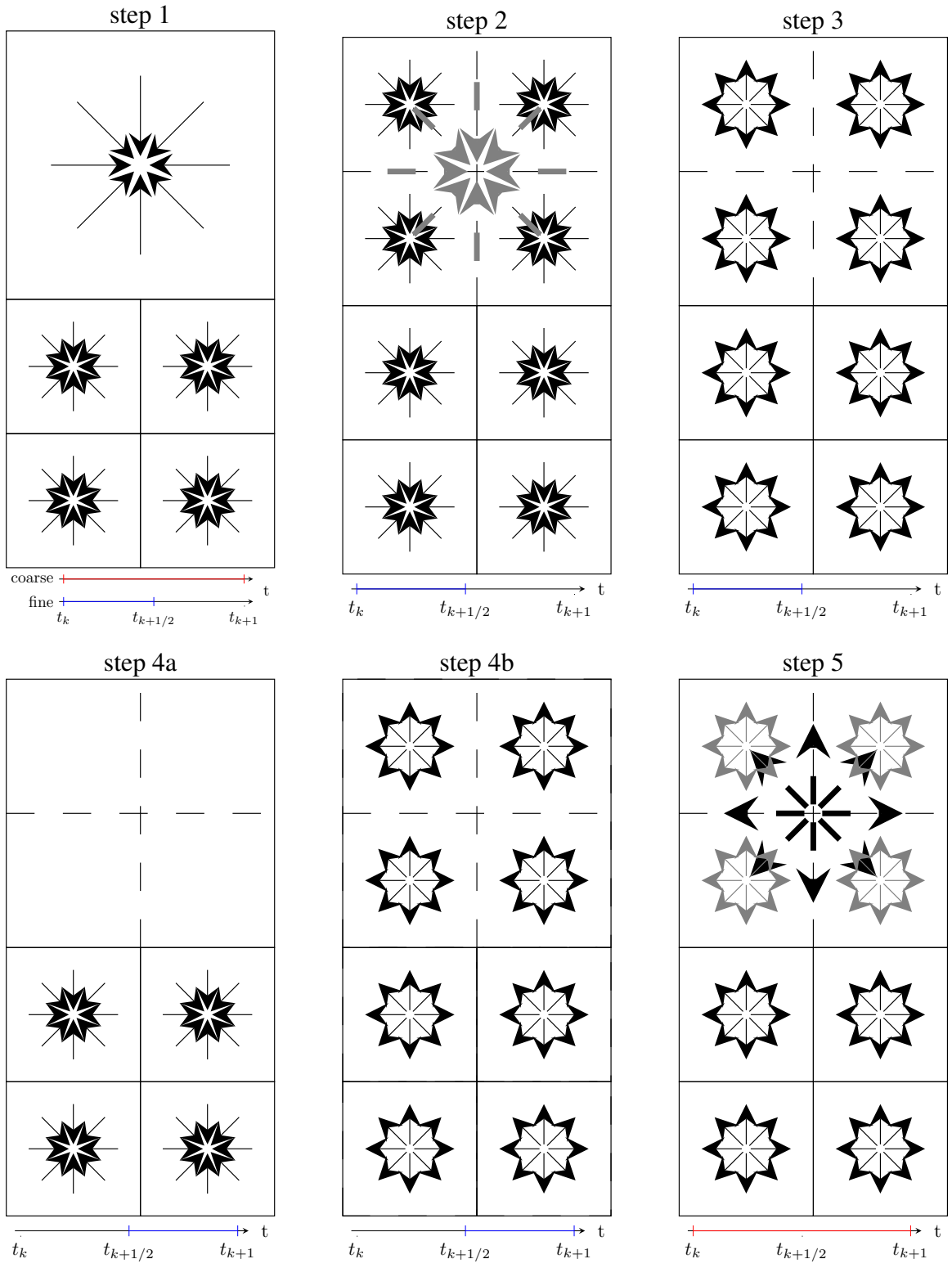


Figure 7: Collision operator independent extension to the algorithm of the LBM to spatially adaptive Cartesian grids as proposed by [53]. 1. Collision on coarse and fine grid. 2. Homogeneous redistribution of populations from coarse to fine grid. 3. Streaming step on virtual fine and fine grid. 4a. Collision step in fine grid. 4b. Streaming step on virtual fine and fine grid. 5. Homogeneous redistribution of populations from virtual fine to coarse grid. Note that steps 4a and 4b are only performed once due to the 2:1 balancing constraint.

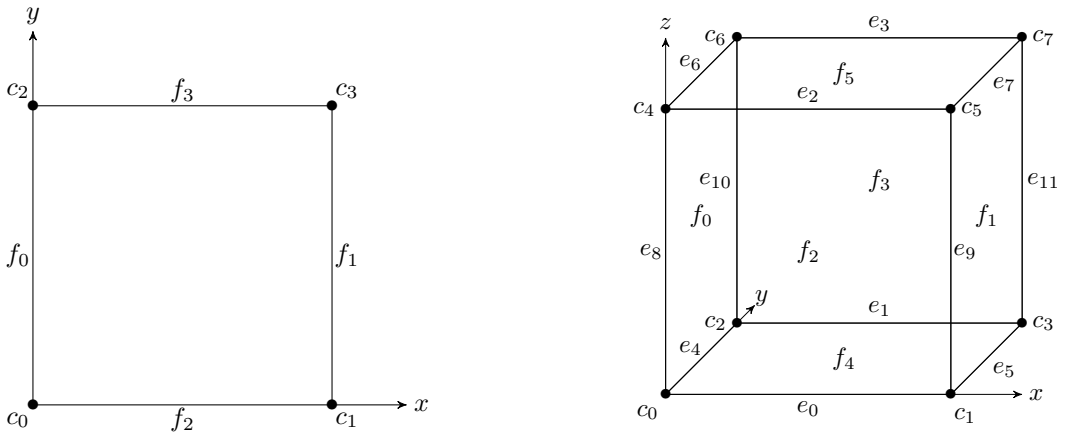


Figure 8: Naming scheme of faces, edges, and corners in `p4est`.

terface may be hanging and may have one side consisting of four small faces instead of one, or two small edges instead of one. We display two possible situations in Figure 9. We have to combine this information to end up with the list of the 18 neighbours required in the D3Q19 LBM, which means that we have to eliminate face neighbours from the edge neighbour list and face and edge neighbours from the corner neighbour list. The Z-curve allows for a very simple and efficient calculation of the Morton-index of neighbour cells. However, since the grid is adaptively refined, the Morton index is not the same as the position in the actual cell list. Therefore, neighbour cells are detected via a recursive divide-and-conquer algorithm operating on the cell list[1]. `p4est` uses a fixed naming scheme for faces, edges and corners of a cell as displayed in Figure 8. This helps to reduce the costs for the neighbour search by allowing us to fill the neighbour lookup tables pairwise for neighbour cells within the same tree using a simple mechanism based on XOR to determine the naming of the respective face, edge or corner in the considered partner cell.

For the detection of neighbours across boundaries between different octrees, the relative orientation of the involved octrees has to be taken into account. Figure 9 shows two examples for such situations. We observe that, in this case, even the number of neighbours of an edge or corner can take any value. However, for our LBM, only rectangular meshes are envisioned such that at least this latter issue does not occur. As different trees may have different local coordinate systems we cannot know beforehand which edge or corner index the diagonally opposite element will have. The orientation and the index of a neighbouring tree can be retrieved from the `p4est_connectivity`, which represents the macroscopic structure of the grid, i.e., the way that the different octrees are arranged relative to each other. With this, we calculate the edge or corner indices of neighbours using a transformation.

In order to transform the index of a corner from the perspective of one tree to the perspective of another across a face, a transformation was proposed in [3] that only needs the face indices of the common face from both perspectives (f_i, f_j), the relative orientation (r) of the two octree faces and the respective face corner ξ_i. A local Morton order is imposed on the corners of each face by enumerating all four in ascending order w.r.t. `p4est`'s naming scheme. Then the face

[1] `p4est` works with an optimised binary search that recursively splits the list of the leaf cells covered by a branch of the octree into eight sublists, one for each child branch, during the top-down depth-first tree traversal in `p4est_iterate` [50]. This function also identifies neighbours of different sizes created by adaptation.

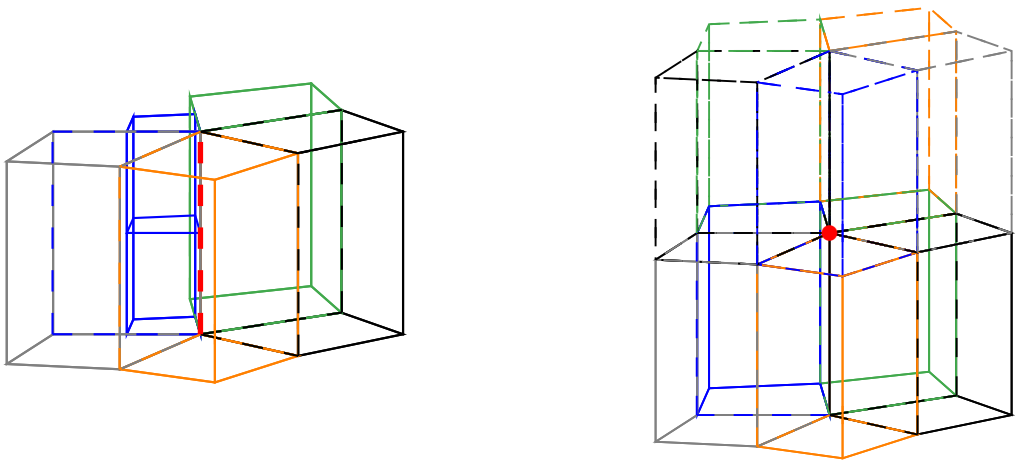


Figure 9: Exemplary multiple tree scenarios to find edge or corner neighbours for each cell. Each tree is drawn in a different colour. The element over which neighbours are to be found is marked in red.

corner number ξ_j viewed from the adjacent octree is given by the permutation

$$\xi_j = \mathcal{P}_c(\mathcal{Q}_c(\mathcal{R}_c(f_i, f_j), r), \xi_i) \equiv \xi_j(\xi_i) \tag{4}$$

with

$$\mathcal{R}_c = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & 2 & 2 & 0 & 0 & 1 \\ 0 & 2 & 2 & 0 & 0 & 1 \\ 2 & 0 & 0 & 2 & 2 & 0 \end{pmatrix} \quad \mathcal{Q}_c = \begin{pmatrix} 1 & 2 & 5 & 6 \\ 0 & 3 & 4 & 7 \\ 0 & 4 & 3 & 7 \end{pmatrix} \quad \mathcal{P}_c = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 2 & 1 & 3 \\ 1 & 0 & 3 & 2 \\ 1 & 3 & 0 & 2 \\ 2 & 0 & 3 & 1 \\ 2 & 3 & 0 & 1 \\ 3 & 1 & 2 & 0 \\ 3 & 2 & 1 & 0 \end{pmatrix}. \tag{5}$$

The permutations of the face corners listed in the rows of $\mathcal{P}_c$ correspond to the 2×4 elements of the dihedral group of face orientations. The matrices $\mathcal{R}_c$ and $\mathcal{Q}_c$ are designed to choose the correct one in a non-redundant way (using 0-based indexing).

To extend this logic to edges, we propose a similar transformation. In addition to the already defined face corners ξ_i, we define face edges ν_i in the same way, i.e. by enumerating the edges spanning a face in ascending (Morton-)order. Then each face edge is enclosed by two face corners that can be written as rows of

$$\Xi_i = \begin{pmatrix} 0 & 2 \\ 1 & 3 \\ 0 & 1 \\ 2 & 3 \end{pmatrix}. \tag{6}$$

We can thus infer the edge corners seen from the adjacent tree by the element-wise application of (4),

$$\Xi_j(\nu_i, k) = \xi_j\left(\Xi_i(\nu_i, k)\right), \quad k = 0, 1. \tag{7}$$

Using (6) in reverse, the resulting two face corners can be identified one-to-one with the face edge's number seen from the other tree. We can condense this process into the form used earlier by writing the edge transformation

$$\nu_j = \mathcal{P}_e(\mathcal{Q}_e(\mathcal{R}_e(f_i, f_j), r), \nu_i) \equiv \nu_j(\nu_i) \tag{8}$$

with

$$\mathcal{R}_e = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \quad \mathcal{Q}_e = \begin{pmatrix} 4 & 1 & 2 & 7 \\ 0 & 6 & 5 & 3 \end{pmatrix} \quad \mathcal{P}_e = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 3 & 2 \\ 1 & 0 & 2 & 3 \\ 1 & 0 & 3 & 2 \\ 2 & 3 & 0 & 1 \\ 2 & 3 & 1 & 0 \\ 3 & 2 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}. \tag{9}$$

The two permutations (4) and (11) allow excluding face neighbouring cells when creating edges and corner lookup tables and are sufficient for our D3Q19 LBM.

To complete `p4est_mesh` we additionally have to exclude edge neighbours, when the corner lookup table is created. We define edge corners μ_i, analog to face corners ξ_i and face edges ν_i. It has to be noted that in this case the relative orientation is no longer between two faces but between two edges. Thus, we will distinguish the relative orientation between two edges from the relative orientation between two faces by denoting the former by s. In case of two faces only two cases have to be distinguished instead of four, because edges can either face the same or the opposite direction. Again, this information can be derived from the face corner transformation. Inspecting (5) and (7), we see that some transformations will lead to two ascending edge corner numbers and others to descending ones. In the latter case the edge is flipped when going between the two trees and the edge corners must be permuted.

To determine s for each edge neighbour we consider all three edges that contain the corner across which we want to find neighbours separately. Beginning with the current cell, we set the relative orientation $s = 0$. Then we can iterate over the two faces adjacent to the respective edge into the next cell. One of the two faces we do not have to consider, because it is the face across which we entered the cell. The relative orientation of the edge s in the face neighbouring cell is determined using (10).

$$s_{\text{new}} = (s_{\text{old}} + (\mathcal{P}_s(\mathcal{Q}_s(\mathcal{R}_s(f_i, f_j), r), \nu_i))) \mod 2 \tag{10}$$

with

$$\mathcal{R}_s = \mathcal{R}_e \quad \mathcal{Q}_s = \begin{pmatrix} 0 & 2 & 1 & 3 \\ 0 & 1 & 2 & 3 \end{pmatrix} \quad \mathcal{P}_s = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \tag{11}$$

To map edge corners from one octree to the other we can apply the transformation

$$\mu_j = \mathcal{P}_{ec}(\mathcal{Q}_{ec}(r), \mu_i) \tag{12}$$

with

$$\mathcal{Q}_{ec} = \begin{pmatrix} 0 & 1 \end{pmatrix} \quad \mathcal{P}_{ec} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{13}$$

In the implementation, the relative orientation s does not need to be calculated separately for each edge neighbour. Instead, s is stored inside of `p4est_connectivity`. This redundancy facilitates creating `p4est_mesh`, because the above transformation can be applied directly.

Summarising, the first permutation from [3] allows removing the face neighbours when setting up the lookup table for corner neighbours, while the third permutation allows to omit edge

neighbours in that case. The second permutation is used to remove face neighbours when setting up edge neighbour lookup tables.

Virtual refinements. Our adaptive LBM algorithm requires to persistently store the populations within virtually embedded cells over multiple grid traversals to restrict their results back to the coarse cell after two smaller time steps on the fine grid. Therefore, we have to store their data additionally to the data of the original coarse cell. This leads to two obvious options considering the question which cells we store in the list of octree leaf quadrants. 1) We store the virtual fine cells as part of the leaf list and create the parent cell as a virtual cell or 2) we store the coarse cell and create four (2D) or eight (3D) virtual fine cells. We choose the latter solution and store the virtual cells in Morton order using a modified version of the `p4est_childrен` inside `p4est_mesh`, creating them simultaneously with the lookup tables. This algorithm differs from a standard refinement step only in the sense that the resulting quadrants are not stored within the list of leaf quadrants and that the parent quadrant is not discarded. This allows integrating virtual cells into an existing application in a completely transparent way, hiding the whole process from the application via abstraction in the neighbour lookup. Furthermore, we can account for the increased load of those cells that contain virtual children by modifying the weight passed to the partition function [3].

4.3 Minimally-Invasive Integration of `p4est` with ESPResSo

As a very general change that is not directly related to grid adaptivity, we wrote some additional callbacks such that the manager node can call the slave nodes to perform the respective distributed routines of `p4est` for initialising, traversing or deallocating a grid. This enables `p4est`, which is designed as a fully distributed library, to work with ESPResSo, which uses a manager/worker architecture. In case of a manager/worker architecture the system is controlled by one or more manager nodes which distribute the work among the worker nodes. Note, that the number of worker nodes is generally significantly bigger than the number of manager nodes. In case of a fully distributed architecture all nodes are equal and have to agree on distribution of the workload. Therefore, a manager/worker scheme avoids solving some (either implicit or explicit) consensus problems at the cost of a reduced fault tolerance.

Besides that, the actual changes that were made to ESPResSo for integrating `p4est` were fairly small. We changed the storage scheme and the grid-traversal from the classical lexicographic ordering to an ordering along the space-filling curve. Cell-local functions, a category in which in case of the LBM all algorithmic steps belong except from the streaming step, could be preserved the way they were. We only changed the function headers to make the respective functions callable as volume callback function using `p4est_iterate` and, for efficiency reasons, changed the data layout from a propagation optimised storage scheme to a collision optimised data scheme, cf. [54]. In a collision optimised data layout, all populations of a cell are grouped locally in memory, i.e., the memory contains $N \times Q$ populations. On the other hand, in a population optimised data layout all cells of a lattice velocity $\boldsymbol{c_i}$ are grouped locally such that $Q \times N$ populations are stored. Both data layout schemes are illustrated in Fig. 10. The reason for switching the data layout is that, in case of dynamic refining and coarsening, we have to shift data in a collision optimised scheme once to preserve data locality as it is given by the space-filling curve. In a propagation optimised scheme, we have to shift data Q times, i.e., in our case 19 times. The remainder of the code was preserved.

For functions requiring access to neighbouring cell data, which in case of the LBM is the streaming step, we changed the header such that the function is callable as a volume callback function using `p4est_iterate`. Neighbour data are accessed using the lookup tables as they

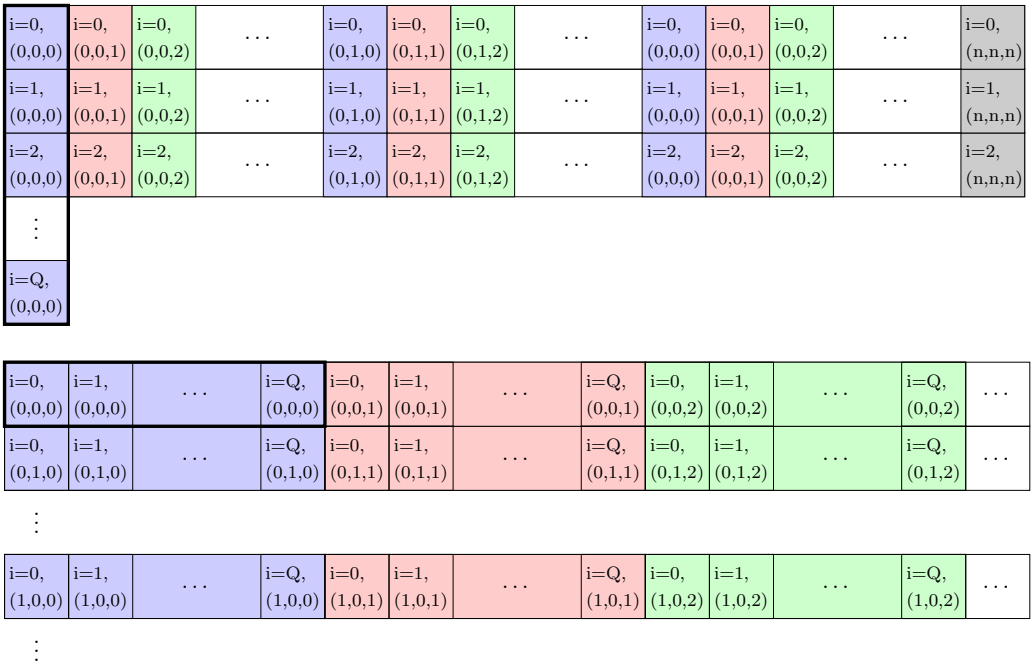


Figure 10: Comparison of different data layout schemes for the LBM. The highlighted black rectangle shows which data have to be loaded for the collision step in the first cell. The upper half shows the propagation optimised data layout which requires more cache lines to be read for the collision step. The lower half of the figure shows the collision optimised data layout. Image adapted from [33].

are provided by `p4est_mesh` instead of performing classical index calculations in the regular grid's (i, j, k)-indexing scheme. To hide the complexity of the different tables from the user, we extended the `p4est_mesh` module by a function returning a list of pointers to the respective neighbouring quadrants as well as a list of encodings specifying the meta information contained in the mesh.

Moreover, `p4est_iterate` does only visit local quadrants, specifically it does not visit ghost quadrants in the parallel domain partitioned version. That means we do not perform a collision step in the ghost layer but we perform an additional communication step after the collision step. This is motivated by a more rigorous understanding of ownership of quadrants: the data of a quadrant is only written by the process that logically owns this quadrant. Following to this idea and in order to perform correct streaming steps at process boundaries, the streaming and the bounce back step (at boundaries) have been slightly adapted as shown in Fig. 11. In case that a neighbouring cell is part of the ghost layer we mirror the streaming step. Instead of streaming populations from the current cell into the ghost cell, we do the inverse streaming operation (for the inverse lattice velocity) from the ghost cell into the current cell.

While the complete integration of spatial adaptivity is not yet finished, the mesh is already prepared for saving virtual quadrants. To achieve full dynamical adaptivity, we will change the grid traversal to level-wise traversals matching the time-adaptive algorithm outlined in 4.1 and Fig. 7 instead of traversing the entire grid while explicitly omitting all cells of different levels.

5 RESULTS

To test our integration we performed some tests on a shared memory machine with 72 physical cores (Intel® Xeon® CPU E7-8880 @2.30GHz) and 512 GB RAM. In 5.1, we compare our new `p4est` grid using a regular discretisation to the existing regular grid. To mimic spa-

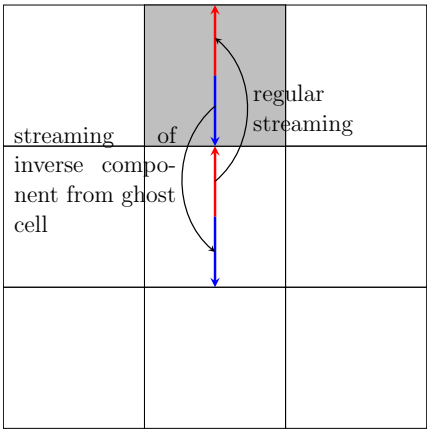


Figure 11: Illustration of the streaming step at neighbouring cells from the ghost layer (grey). Instead of streaming from the local cell into the ghost cell (red) we perform the inverse streaming step from the ghost cell to the current cell (blue).

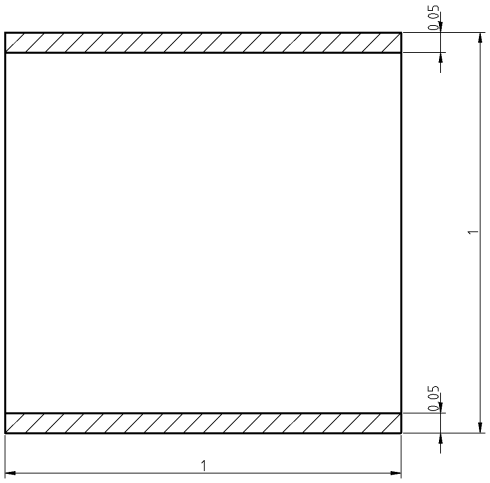


Figure 12: Geometry of the considered Poiseuille flow scenario. Note that the channel is not cylindrical but extruded to better match with the unit hypercube.

tial adaptivity in terms of realistic run-time, we do not use the information that the actually used grid is regularly refined and constant over time. Instead, we reconstruct the entries of `p4est_ghost` and `p4est_mesh` in each time step. In 5.2, we analyse the additional costs of this choice, i.e., the time it takes to create both structures. The section is concluded by comparing the run-time for traversing the grid using `p4est_iterate` and `p4est_mesh` in 5.3.

5.1 Run-time of LBM on Different Grids

To test our integration, we performed some scalability tests for a simple Poiseuille flow scenario as it is shown in Fig. 12. Note that the channel was not rotated to a cylindrical shape but that it has been extruded to the size of the unit cube, thus containing more fluid and fewer obstacle cells compared to a cylindrical channel.

We compared our `p4est` based implementation using a regularly refined octree and a collision optimised data layout to the existing regular grid with lexicographic cell-ordering using a propagation optimised data layout. For simulating the run-time effects of dynamic grid adaptivity, we did not use the information that we have both a constant layer of ghost cells

Proc	NUPS p4est					NUPS regular grid				
	Level 4	Level 5	Level 6	Level 7	Level 8	Level 4	Level 5	Level 6	Level 7	Level 8
16	1,323.85	2,824.1	3,621.2	4,044.3	4,300.2	2,185.7	9,019.6	11,083.8	13,934.5	13,867.5
32	1,411.93	4,567.6	6,450.7	7,513.4	8,294.2	2,200.96	11,583	21,357.7	23,125	27,104
64	1,422.23	6,261.7	9,461.7	13,600.2	15,498.8	1,895.42	11,279.9	36,880	39,871	46,580

Table 2: Comparison of lattice-Boltzmann run-times using the p4est grid with existing regular grids on various discretisation levels for an increasing number of MPI processes in NUPS (node updates per second).

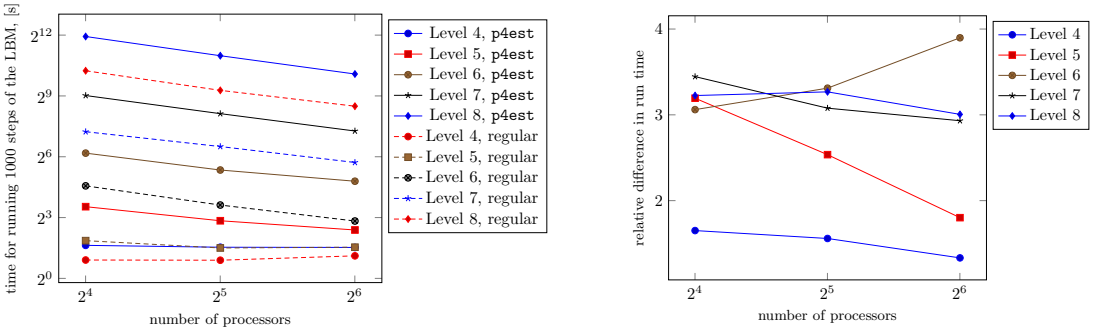


Figure 13: Comparing the run-time for $1,000$ time-steps of the LBM on various discretisation levels using the existing regular grid and the newly integrated p4est grid for an increasing number of processes (left). On the right we show the relative factor that our implementation is slower.

and a constant neighbourhood. Instead, we discarded and re-created the information about the ghost layer in p4est_ghost as well as the lookup tables for the neighbouring information in p4est_mesh in each time step. The overhead of dynamical grid adaptivity was bound by a factor of 4 as shown in Table 2 and Fig. 13. We therefore could prove, that even in a worst case scenario, where the grid would be allowed to change in every fine time step, the performance of the adaptive grid does not degrade compared to an optimised regular grid. Note that this is true in spite of using a data layout that is known to reduce the performance of a regular discretisation and despite performing additional grid traversals for rebuilding p4est_ghost and p4est_mesh. We use the collision optimised data layout, because it will reduce the amount of data that has to be transferred in case of dynamic spatial adaptivity, cf. 4.3. The impact on performance for the different data layouts in case of a regular grid is described in [54].

Further optimisations reusing the overhead are to be done in the future. E.g., by restricting the grid adaptivity to the end of the coarsest time steps, we can omit rebuilding the information of p4est_ghost and p4est_mesh during the smaller time steps of the smaller cells.

5.2 Creating p4est_ghost and p4est_mesh

Building the information of neighbourhood and ghost cells in case of a regular grid can be done in constant time using simple index calculations. In case of a dynamically-adaptive grid, this process is more challenging. A comparison between creating the ghost layer and accessing neighbours in case of regular and octree-based grids is schematically shown in Fig. 14. To investigate the overhead of the more complicated ghost layer as well as the creation of lookup tables for the neighbourhood information, we performed run-time and scalability experiments using our regularly refined grids within p4est for the Poiseuille flow scenario.

We measured the run-time for creating both, p4est_mesh and p4est_ghost using a single octree with periodic boundaries for different discretisation levels and a varying number of processes. Both, ghost information and lookup tables have only been created for faces and

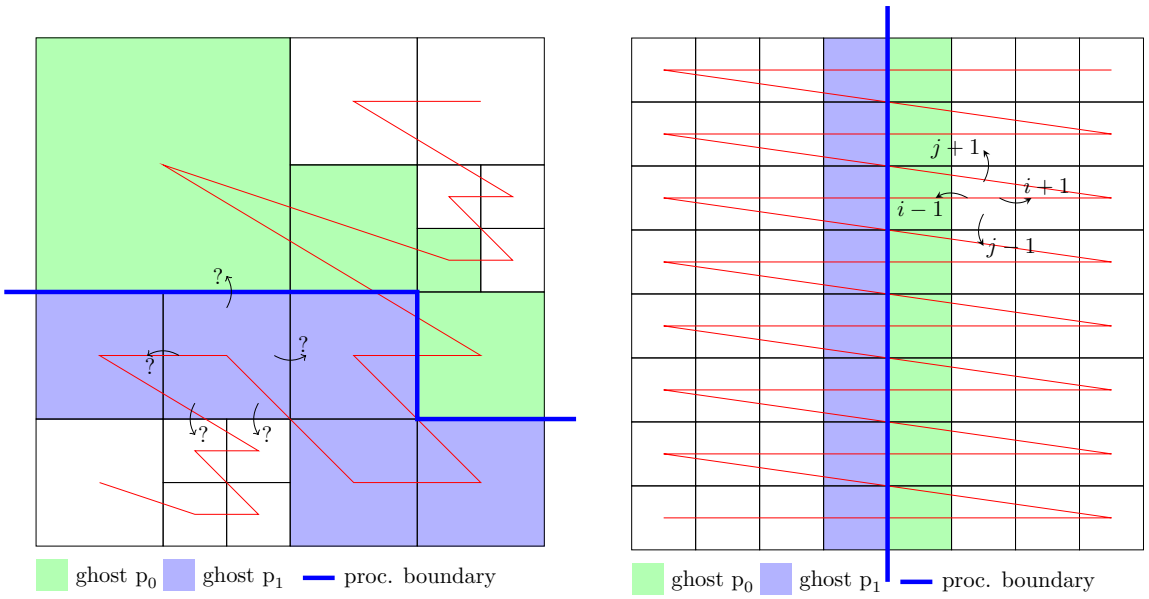


Figure 14: Differences in the structure and calculation of ghost layers and neighbour in formation in a regular grid versus a spatially adaptive grid.

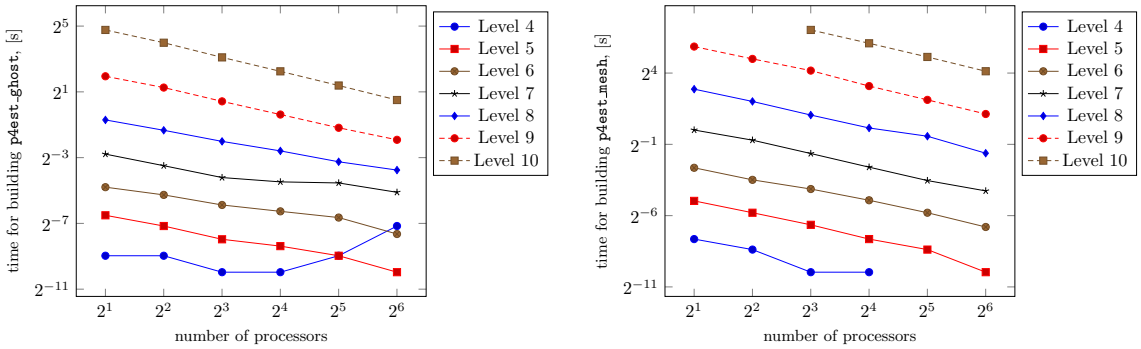


Figure 15: Left: run-times for creating `p4est_ghost`; right: run-times for creating `p4est_mesh` for a regularly refined grid on various discretisation levels for an increasing number of processes.

edges, i.e., no information about corner neighbours was gathered as this is not required by the D3Q19 lattice-Boltzmann method. The results are shown in Fig. 15. Note that the volume to surface ratio is low in the grid of level 4 which limits the scalability for creating the ghost information.

5.3 Run-Times of Grid Traversals

To compare the computational costs of traversing the grid in the streamlined and optimised way using `p4est_iterate` and the more flexible `p4est_mesh` we applied some stencil operations similar to the one used by the D3Q19 LBM. Again, we investigated the run-time for traversing a regularly discretised `p4est` based grid with periodic boundary conditions once while accessing all nearest and next-nearest neighbours. We varied both, discretisation level and number of processes. The results are shown in Fig. 16. We observe that traversing the grid using `p4est_mesh` is on average a factor of 2 slower than using `p4est_iterate`. This represents in a sense the price we pay for the minimal-invasiveness, in particular for being able to preserve the existing grid traversal logic and the direct neighbour access in ESPResSo. Note that on discretisation level four there is only a very limited number of quads associated to each

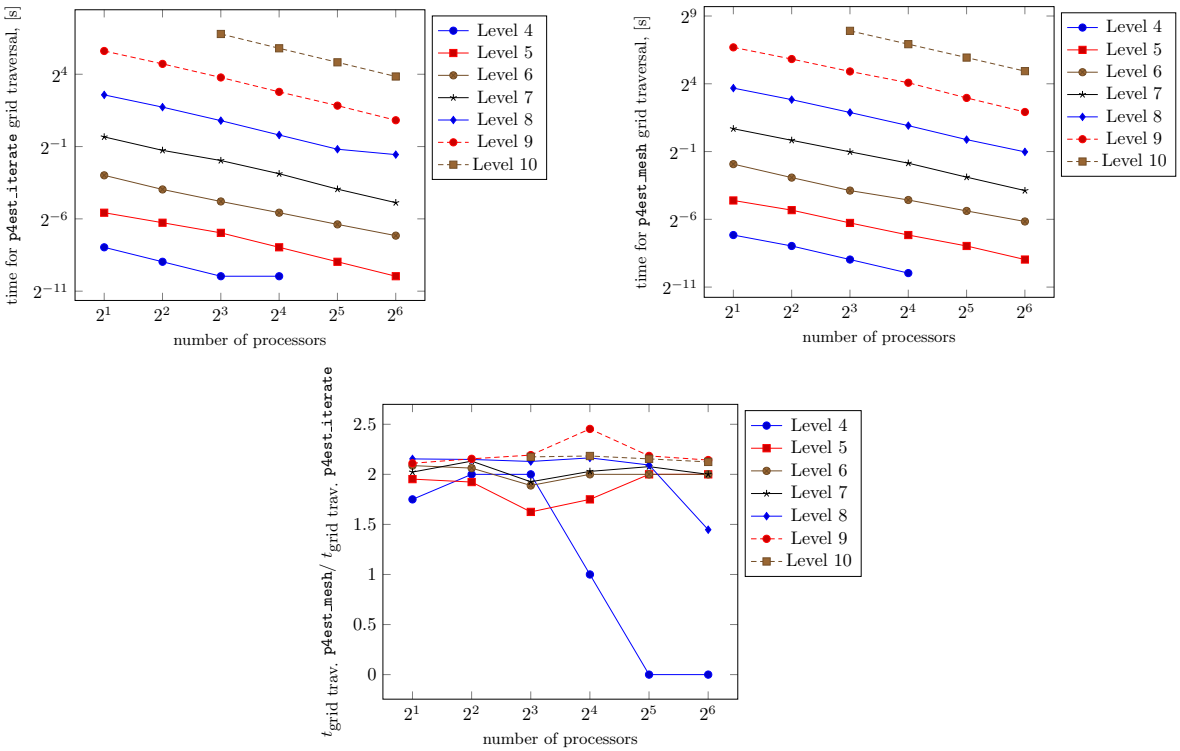


Figure 16: Comparison of the run-times of a grid traversal. Top left: using the highly optimised `p4est_iterate`; top right: using the much more flexible `p4est_mesh`; bottom: relative overhead of `p4est_mesh` compared to `p4est_iterate`. In both cases, we traverse a regularly refined grid using an increasing number of processes.

processing node such that in both cases the run-time was less than a microsecond.

6 CONCLUSION AND OUTLOOK

With the combination of ESPResSo and `p4est`, we have prepared a powerful implementation of a lattice-Boltzmann solver on adaptively refined octree grids. We are not the first group setting up such a solver. However, there are a few major differences to other approaches: [55] and [56] both implement the lattice-Boltzmann method on top of a highly optimised octree grid implementation. I.e., the solver itself had to be developed more or less from scratch in a way that is adapted to the grid structure and its specific memory-optimal traversal in the Peano framework [57] or waLBerla [58]. Also, [56] provides a solver on statically adaptive meshes, which allows for much more sophisticated solutions in terms of run-time optimisation per core and on massively parallel hardware. In contrast, we focus on full dynamical adaptivity as required by the physical systems simulated with ESPResSo, where regions with high refinement requirements move with particles or molecules immersed in the flow. In addition, the task was to port the full lattice-Boltzmann functionality in ESPResSo to the new grid type with only minimal changes to the code.

Our results show that `p4est` is well-suited for minimally-invasive integration of an octree-based grid into existing applications. Some changes in the code of the existing application are necessary, as at least the additional numerical features at boundaries between different refinement levels have to be developed from scratch and i, j, k-loops have to be exchanged by more general variants. However, most of the core code can be preserved. Even in worst-case tests where we assumed full dynamical adaptivity in every fine grid time step, we observe good performance and scalability. The flexibility and minimal invasiveness of the `p4est_mesh` comes

with an acceptable overhead when compared to specialised versions such as [58, 57].

In the future, we intend to tackle complex real world applications using the new solver. Scenarios that we expect to clearly benefit from spatial and time adaptivity are, e.g., the translocation of DNA through a nanopore [59] or the filtering of dust particles in charged cabin air filters [60]. To actually be able to handle these scenarios efficiently, the most important steps to be taken are: 1) the full implementation of the virtual cell functionalities at refinement boundaries; 2) the adaption of the iterations to level-wise iterations accounting for the natural time-adaptivity of the LBM; 3) the development of a suitable weighting concept for octree cells on different tree levels, accounting for the different time-step sizes and possible overhead at partition boundaries in the dynamical load-balancing of `p4est`; 4) the coupling of electrostatics, electrokinetics, and the molecular dynamics modules of ESPResSo to the lattice-Boltzmann solver; 5) the further optimisation of the `p4est_mesh` and its generation.

ACKNOWLEDGEMENTS

B. gratefully acknowledges travel and guest support provided by the Hausdorff Center for Mathematics funded by the DFG (German Research Foundation).

All other authors gratefully acknowledge funding provided by the collaborative research center 716 (Dynamic Simulation of Systems with Large Particle Numbers), funded by the DFG (German Research Foundation).

REFERENCES

[1] H. J. Limbach, A. Arnold, B. A. Mann, and C. Holm. ESPResSo – an extensible simulation package for research on soft matter systems. *Computer Physics Communications*, 174(9):704–727, May 2006.

[2] A. Arnold, O. Lenz, S. Kesselheim, R. Weeber, F. Fahrenberger, D. Röhm, P. Košovan, and C. Holm. ESPResSo 3.1 — molecular dynamics software for coarse-grained models. In M. Griebel and M. A. Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VI*, volume 89 of *Lecture Notes in Computational Science and Engineering*, pages 1–23. Springer Berlin Heidelberg, 2013.

[3] C. Burstedde, L. C. Wilcox, and O. Ghattas. p4est: Scalable algorithms for parallel adaptive mesh refinement on forests of octrees. *SIAM Journal on Scientific Computing*, 33(3):1103–1133, January 2011.

[4] C. Burstedde. `p4est`: Parallel AMR on forests of octrees. `http://www.p4est.org/`, last accessed February 25, 2016.

[5] T. Weinzierl and M. Mehl. Peano—a traversal and storage scheme for octree-like adaptive Cartesian multiscale grids. *SIAM Journal on Scientific Computing*, 33(5):2732–2760, October 2011.

[6] H. Sundar, R. S. Sampath, and G. Biros. Bottom-up construction and 2:1 balance refinement of linear octrees in parallel. *SIAM Journal on Scientific Computing*, 30(5):2675–2708, 2008.

[7] H. G. Klimach, M. Hasert, J. Zudrop, and S. P. Roller. Distributed octree mesh infrastructure for flow simulations. In *European Congress on Computational Methods in Applied Sciences and Engineering*, pages 1–15, 2012.

[8] V. I. Kolobov, R. R. Arslanbekov, V. V. Aristov, A. A. Frolova, and S. A. Zabelok. Unified solver for rarefied and continuum flows with adaptive mesh and algorithm refinement. *Journal of Computational Physics*, 223(2):589–608, 2007.

[9] M. Geier, A. Greiner, and J. G. Korvink. Bubble functions for the lattice Boltzmann method and their application to grid refinement. *The European Physical Journal Special Topics*, 171(1):173–179, 2009.

[10] H. Chen. Volumetric formulation of the lattice boltzmann method for fluid dynamics: Basic concept. *Physical Review E*, 58(3):3955–3963, 1998.

[11] A. Arnold, K. Breitsprecher, F. Fahrenberger, S. Kesselheim, O. Lenz, and C. Holm. Efficient algorithms for electrostatic interactions including dielectric contrasts. *Entropy*, 15(11):4569–4588, 2013.

[12] M. Deserno and C. Holm. How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines. *Journal of Chemical Physics*, 109:7678, 1998.

[13] A. Arnold and C. Holm. A novel method for calculating electrostatic interactions in 2D periodic slab geometries. *Chemical Physics Letters*, 354:324–330, 2002.

[14] A. Arnold, J. de Joannis, and C. Holm. Electrostatics in periodic slab geometries I. *Journal of Chemical Physics*, 117:2496–2502, 2002.

[15] A. Arnold, J. de Joannis, and C. Holm. Electrostatics in periodic slab geometries II. *Journal of Chemical Physics*, 117:2503–2512, 2002.

[16] A. Arnold and C. Holm. MMM1D: A method for calculating electrostatic interactions in 1D periodic geometries. *Journal of Chemical Physics*, 123(12):144103, September 2005.

[17] C. Tyagi, M. Süzen, M. Sega, M. Barbosa, S. S. Kantorovich, and C. Holm. An iterative, fast, linear-scaling method for computing induced charges on arbitrary dielectric boundaries. *The Journal of Chemical Physics*, 132:154112, 2010.

[18] S. Tyagi, A. Arnold, and C. Holm. ICMMM2D: An accurate method to include planar dielectric interfaces via image charge summation. *Journal of Chemical Physics*, 127:154723, 2007.

[19] S. Tyagi, A. Arnold, and C. Holm. Electrostatic layer correction with image charges: A linear scaling method to treat slab 2D + h systems with dielectric interfaces. *Journal of Chemical Physics*, 129(20):204102, 2008.

[20] J. J. Cerdà, V. Ballenegger, O. Lenz, and C. Holm. P3M algorithm for dipolar interactions. *Journal of Chemical Physics*, 129:234104, 2008.

[21] D. Roehm and A. Arnold. Lattice Boltzmann simulations on GPUs with ESPResSo. *The European Physical Journal Special Topics*, 210:89–100, 2012.

[22] G. Inci, A. Arnold, A. Kronenburg, and R. Weeber. Modeling nanoparticle agglomeration using local interactions. *Aerosol Science and Technology*, 48(8):842–852, 2014.

[23] J. Höpfner, T. Richter, P. Košovan, C. Holm, and M. Wilhelm. Seawater desalination via hydrogels: Coarse grained simulations and practical realisation. *Progr. Colloid. Polym. Sci.*, 140(140), 2013.

[24] B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Muller, K. Kremer, and M. Deserno. Aggregation and vesiculation of membrane proteins by curvature-mediated interactions. *Nature*, 447(7143):461–464, May 2007.

[25] M. Kuron and A. Arnold. Role of geometrical shape in like-charge attraction of dna. *European Physical Journal E: Soft Matter*, 38:20, 2015.

[26] S. Kesselheim, M. Sega, and C. Holm. Effects of dielectric mismatch and chain flexibility on the translocation barriers of charged macromolecules through solid state nanopores. *Soft Matter*, 8(36):9480–9486, 2012.

[27] K. Breitsprecher, P. Košovan, and C. Holm. Coarse grained simulations of an ionic liquid-based capacitor i: density, ion size, and valency effects. *Journal of Physics: Condensed Matter*, 26:284108, 2014.

[28] S. Succi. *The lattice Boltzmann equation*. Oxford university press New York, 2001.

[29] B. Dünweg and A. J. C. Ladd. Lattice boltzmann simulations of soft matter systems. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, volume 221 of *Advances in Polymer Science*, pages 89–166. Springer-Verlag Berlin, Berlin, Germany, 2009.

[30] S. Succi, M. Sbragaglia, and S. Ubertini. Lattice boltzmann method. *Scholarpedia*, 5(5):9507, 2010.

[31] K. Grass. *Towards realistic modelling of free-solution electrophoresis: a case study on charged macromolecules*. PhD thesis, Goethe-Universität Frankfurt, 2009.

[32] L.-S. Luo, W. Liao, X. Chen, Y. Peng, and W. Zhang. Numerics of the lattice boltzmann method: Effects of collision models on the lattice boltzmann simulations. *Physical Review E*, 83(5):056710, 2011.

[33] U. D. Schiller. *Thermal fluctuations and boundary conditions in the lattice Boltzmann method*. PhD thesis, Johannes Gutenberg-Universität, Mainz, 2008.

[34] B. Dünweg, U. Schiller, and A. J. C. Ladd. Statistical mechanics of the fluctuating lattice-boltzmann equation. *Phys. Rev. E*, 76:36704, 2007.

[35] W. C. Rheinboldt and C. K. Mesztenyi. On a data structure for adaptive finite element mesh refinements. *ACM Transactions on Mathematical Software*, 6(2):166–187, 1980.

[36] D. A. Kopriva, S. L. Woodruff, and M. Y. Hussaini. Computation of electromagnetic scattering with a non-conforming discontinuous spectral element method. *International Journal for Numerical Methods in Engineering*, 53(1):105–122, 2002.

[37] D. Rosenberg, A. Fournier, P. Fischer, and A. Pouquet. Geophysical-astrophysical spectral-element adaptive refinement (GASpAR): Object-oriented h-adaptive fluid dynamics simulation. *Journal of Computational Physics*, 215(1):59–80, 2006.

[38] M. Bader. *Space-Filling Curves: An Introduction with Applications in Scientific Computing*. Texts in Computational Science and Engineering. Springer, 2012.

[39] D. Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*, 38:459–460, 1891.

[40] G. M. Morton. A computer oriented geodetic data base; and a new technique in file sequencing. Technical report, IBM Ltd., 1966.

[41] M. Griebel and G. W. Zumbusch. Parallel multigrid in an adaptive PDE solver based on hashing and space-filling curves. *Parallel Computing*, 25:827–843, 1999.

[42] H.-J. Bungartz, M. Mehl, and T. Weinzierl. A parallel adaptive Cartesian PDE solver using space–filling curves. *Euro-Par 2006 Parallel Processing*, pages 1064–1074, 2006.

[43] T. Tu, D. R. O'Hallaron, and O. Ghattas. Scalable parallel octree meshing for terascale applications. In *SC '05: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2005.

[44] C. Burstedde, O. Ghattas, M. Gurnis, E. Tan, T. Tu, G. Stadler, L. C. Wilcox, and S. Zhong. Scalable adaptive mantle convection simulation on petascale supercomputers. In *SC08: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM/IEEE, 2008.

[45] R. S. Sampath and G. Biros. A parallel geometric multigrid method for finite elements on octree meshes. *SIAM Journal on Scientific Computing*, 32(3):1361–1392, 2010.

[46] J. R. Stewart and H. C. Edwards. A framework approach for developing parallel adaptive multiphysics applications. *Finite Elements in Analysis and Design*, 40(12):1599–1617, 2004.

[47] W. Bangerth, R. Hartmann, and G. Kanschat. deal.II – a general-purpose object-oriented finite element library. *ACM Transactions on Mathematical Software*, 33(4):24, 2007.

[48] T. Isaac, C. Burstedde, and O. Ghattas. Low-cost parallel algorithms for 2:1 octree balance. In *Proceedings of the 26th IEEE International Parallel & Distributed Processing Symposium*. IEEE, 2012. `http://dx.doi.org/10.1109/IPDPS.2012.47`.

[49] M. Mirzadeh, A. Guittet, C. Burstedde, and F. Gibou. Parallel level-set methods on adaptive tree-based grids. Submitted, 2015.

[50] T. Isaac, C. Burstedde, L. C. Wilcox, and O. Ghattas. Recursive algorithms for distributed forests of octrees. *SIAM Journal on Scientific Computing*, 37(5):C497–C531, 2015.

[51] J. Rudi, A. C. I. Malossi, T. Isaac, G. Stadler, M. Gurnis, P. W. J. Staar, Y. Ineichen, C. Bekas, A. Curioni, and O. Ghattas. An extreme-scale implicit solver for complex pdes: highly heterogeneous flow in earth's mantle. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, page 5. ACM, 2015.

[52] A. Müller, M. A. Kopera, S. Marras, L. C. Wilcox, T. Isaac, and F. X. Giraldo. Strong scaling for numerical weather prediction at petascale with the atmospheric model NUMA. http://arxiv.org/abs/1511.01561, 2015.

[53] M. Rohde, D. Kandhai, J. J. Derksen, and H. E. A. van den Akker. A generic, mass conservative local grid refinement technique for lattice-boltzmann schemes. *International Journal for Numerical Methods in Fluids*, 51(4):439–468, 2006.

[54] G. Wellein, T. Zeiser, G. Hager, and S. Donath. On the single processor performance of simple lattice Boltzmann kernels. *Computers and Fluids*, 35(8-9):910–919, September 2006.

[55] P. Neumann and T. Neckel. A dynamic mesh refinement technique for lattice boltzmann simulations on octree-like grids. *Computational Mechanics*, 51(2):237–253, January 2013.

[56] C. Feichtinger, J. Habich, H. Köstler, G. Hager, U. Rüde, and G. Wellein. A flexible patch-based lattice boltzmann parallelization approach for heterogeneous gpu-cpu clusters. *Parallel Computing*, 37:536–549, 2011.

[57] T. Weinzierl and M. Mehl. Peano — a traversal and storage scheme for octree-like adaptive cartesian multiscale grids. *SIAM Journal on Scientific Computing*, 33(5):2732–2760, January 2011.

[58] C. Feichtinger, S. Donath, H. Köstler, J. Götz, and U. Rüde. WaLBerla: HPC software design for computational engineering simulations. *Journal of Computational Science*, 2:105–112, 2011.

[59] S. Kesselheim, M. Sega, and C. Holm. Applying ICC$\star$ to dna translocation: Effect of dielectric boundaries. *Computer Physics Communications*, 182(1):33–35, 2011.

[60] M. J. Lehmann, J. Weber, A. Kilian, M. Heim, and MANN+HUMMEL GmbH, Germany. Micro scale simulation as part of fibrous filter media development processes - from real to virtual media. Feb 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

AUTOMATIC CONFORMAL DECOMPOSITION OF ELEMENTS CUT BY NURBS

Jakob W. Steidl[1] **and Thomas-Peter Fries**[1]

[1]Institute of Structural Analysis, Graz University of Technology
Lessingstrae 25/II, 8010 Graz, Austria
e-mail: {steidl, fries}@tugraz.at

Keywords: conformal decomposition, fictitious domain method, NURBS, signed distance, level set method

Abstract. *In fictitious domain and extended finite element methods, elements cut by boundaries or interfaces are frequently present. They pose challenges with respect to the application of boundary and interface conditions, numerical integration, and the conditioning of the system of equations. A different path is the automatic decomposition of cut elements into smaller, conforming finite elements. Herein, we propose a concept for the automatic decomposition of elements cut by (curved) NURBS which automatically generates higher-order Lagrange elements on the two sides of the interfaces. The NURBS data, defining boundaries and interfaces in two dimensions, is converted into implicit level-set data by evaluating signed distance computations at all nodes in the background mesh. Based on the findings in [1] this data is then used for the decomposition into sub-elements. First numerical results are achieved as a proof-of-concept.*

1 INTRODUCTION

Within the scientific community of numerical engineering, recently efforts are made to incooperate exact geometry definitions delivered from designers. The reasons for including the exact geometry definition are due to the tighter integration of simulation tools within design software. However, also problems sensitive to the geometry, such as buckling of shells or in fluid mechanics can benefit.

An effort almost as old as the finite element method itself is the automatic generation of meshes suitable for simulations. While proven meshing approaches are available for some time now (see [2] for an historic overview), in practice, generated meshes are often manually reworked by the user. This intervention, however, is often a problem because the link to the original geometry description is lost, which makes automatic mesh re-generation as in refinement or in moving interface problems difficult.

Also classic mesh triangulation is time consuming and yields a high number of nodes if complex boundaries or interfaces have to be discretized with high accuracy. These two problems are tackled by a number of methods that do not capture boundaries explicitly in terms of element edges. This allows the use of a structured *background mesh* with a wealth of potential speedups in computation time (e.g. constant Jacobians). The trade-off is that elements from the background mesh intersected by a boundary or interface require special treatments.

The kind of treatment of cut elements is what the methods differ in. For example, in the extended Finite Element Method (XFEM) [3], the approximation space is enriched by additional shape functions tailored to capture discontinuities across interfaces. Thereby cut elements yield additional degrees of freedom and customized integration schemes are required.

Another class of methods which naturally involve elements cut by interfaces are immersed boundary methods [4, 5]. The ‘interface’ is then rather the boundary of the domain and, herein, we do no longer distinguish between interfaces and boundaries. These methods embed the original domain into a much simpler, usually rectangular, domain spanned by the background mesh. The methods of this class differ in the treatment of boundary conditions, numerical integration and conditioning issues in the resulting system of equations.

A different approach to handle these intersected elements is to decompose them into subelements that *conform* to the boundary or interface and is referred to as the Conformal Decomposition FEM (CDFEM) in [6]. The advantage of their approach is that new nodes are being introduced to the mesh alongside the interface, which makes the application of boundary conditions trivial. However it also becomes necessary to modify the mesh and to (at least partially) give up the advantages of the structured background mesh. So far conformal decomposition has only been reported using linear elements for reconstruction.

Herein, we are concerned with the extension to higher-order accuracy. The interfaces are defined by NURBS in the beginning. This information is converted into implicit level-set data by performing signed-distance computations. Then, based on the approach in [1] elements cut by the zero-level set are automatically decomposed into sub-elements.

2 TOWARDS A HIGHER-ORDER CONFORMAL DECOMPOSITION METHOD

The aim is to develop a higher-order accurate CDFEM, which uses NURBS for the definition of interfaces. NURBS are able to represent all surface types common in engineering exactly.

The main part of the paper deals with converting the NURBS into level-set data by evaluating distances of nodes in the background mesh to the NURBS. Then, the approach in [1] is used to generate the sub-elements in cut background elements.

2.1 Level Set Method

For implicit interface descriptions, the level set method [7] is frequently used. In this framework interfaces Γ can be described as the zero-isoline of a scalar function $\phi(\mathbf{x})$, i.e.

$$\Gamma = \{\mathbf{x} \mid \phi(\mathbf{x}) = 0\}, \tag{1}$$

where $\mathbf{x}$ describes arbitrary points in space and ϕ typically is the shortest distance from $\mathbf{x}$ to Γ multiplied by a sign to denote the side of Γ where $\mathbf{x}$ lies. In its discrete form, used for numerical analysis, values for ϕ are usually stored at the mesh nodes, that is, for every node in the mesh the distance to Γ needs to be computed. In between the nodes, level set values are interpolated based on the shape function of the background mesh.

2.2 Non-Uniform Rational B-Splines

Non-Uniform Rational B-Splines (NURBS) can be seen as B-Splines projected onto a projective plane. This is what allows them to represent arcs and conic sections exactly. B-Splines are a generalization of Beziér curves, which again are a linear combination of Bernstein polynomials. A NURBS curve, often also denoted as univariate NURBS is defined in [8] as

$$\mathbf{C}(u) = \frac{\sum_{i=0}^{n} N_{i,p}(u) w_i \mathbf{P}_i}{\sum_{i=0}^{n} N_{i,p}(u) w_i} \qquad u_0 < u < u_m \qquad u \in \mathbb{R}$$

where P_i are the control points, w_i the associated weights and $N_{i,p}(u)$ B-Spline basis functions, which can be defined in their recursive form:

$$N_{i,p}(u) = \frac{u - u_i}{u_{i+p} - u_i} N_{i,p-1}(u) + \frac{u_{i+p+1} - u}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(u)$$

$$N_{i,0}(u) = \begin{cases} 1 & u_i < u < u_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

Associated with this is the non-decreasing sequence of $U = \{u_0, \ldots, u_m\}$ called the *knot vector* with $m = n + p + 1$. It is common to repeat u_0 and u_m with a multiplicity of $p + 1$ to ensure the endpoints of the curve are interpolatory with the corresponding control points. The linear connection of the ordered $\mathbf{P}_i$ is referred to as the *control polygon*. Properties of NURBS and Beziér curves include:

- endpoints are interpolatory, ie $\mathbf{C}(u_0) = \mathbf{P}_0$ and $\mathbf{C}(u_m) = \mathbf{P}_n$
- strong convex hull property: for $u \in [u_i, u_{i+p})$ will be inside the convex hull $\mathcal{H}$ spanned by the control points $\mathbf{P}_{i-1} \ldots \mathbf{P}_i$. For Beziér curves this holds for the whole curve.

3 SIGNED MINIMAL DISTANCE TO NURBS CURVES

The signed, minimal distance function defines the shortest Euclidean distance from a *test point* $\mathbf{p}_t$ to a given curve $\mathbf{C}(u)$ and can be written as

$$\phi(\mathbf{p}) = \min(\|\mathbf{C}(u) - \mathbf{p}\|) \cdot s \tag{2}$$

with

$$s = \text{sign}(\langle \mathbf{C}(u) - \mathbf{p}_t, \mathbf{n}_{\mathbf{C}(u)} \rangle. \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\langle\cdot,\cdot\rangle$ the dot product, $\mathbf{n}_{\mathbf{C}(u)}$ the normal vector of $\mathbf{C}(u)$ and

$$\min(\|\mathbf{C}(u) - \mathbf{p}_t\|) = \|\mathbf{C}(u^*) - \mathbf{p}_t\| \tag{4}$$

with the criterion for the *projection point* $\mathbf{p}_p = \mathbf{C}(u^*)$ to $\mathbf{p}_t$

$$\langle \mathbf{C}(u^*) - \mathbf{p}_t, \mathbf{C}_{,u}(u^*)\rangle = 0 \quad \Longleftrightarrow \quad \|\mathbf{C}(u) - \mathbf{p}_t\|_{,u} = 0. \tag{5}$$

Note that for the criteria in (5) a unique solution is not guaranteed. If no solution exists, $\mathbf{p}_p$ must be one of the curve's endpoints ($\mathbf{C}(u_0)$ or $\mathbf{C}(u_m)$). This case can be handled quite easily. In order to find all u^* that can lead to the solution of (4), tailored root finding algorithms have been developed by the authors and others.

3.1 Root finding algorithms

3.1.1 Literature review

For arbitrary polynomial orders p, there is no closed form solution to (5). A common approach is to solve the equation iteratively via a Newton-Raphson iteration scheme as in

$$u^*_{k+1} = u^*_k \cdot \frac{\|\mathbf{C}(u^*_k) - \mathbf{p}_t\|_{,u}}{\|\mathbf{C}(u^*_k) - \mathbf{p}_t\|_{,uu}}. \tag{6}$$

It is well known, that the iteration will only converge for start values close to the actual solution of (5). Therefore a major task is to find start values $U_k = \{u^*_0\}$ such that ultimately the shortest distance to the NURBS is found. It is noted that depending on the NURBS, several local minimal of (5) may exist. Hence, different start values may lead to different converged solutions of (5). It is, therefore, not a trivial task to define suitable start values. In the CADG community the following procedure has emerged: Decompose the NURBS into a set of (projective) Beziér segments, drop segments that can not contain the projection point, refine remaining segments and repeat with the refinement and dropping procedure until a certain exit criterion is reached.

Refinement: Apart from splitting the Beziér curves, in [9] the Beziér curves are further divided into *simple* (i.e. without a self crossing control polygon) and convex curves.

Culling: In order to reduce the number of Beziér curves, in [10] and [11] simplifications of the Beziér curves' convex hull property and interpolation property are employed: Segments with a distance $\min(\|\mathbf{p}_t - \mathcal{H}\|) > \min(\|\mathbf{p}_t - \mathbf{C}(\{u_0, u_m\})\|$ can be dropped because they can not contain the solution.

Stopping criterion: The stopping criterion proposed in [9] checks if the remaining segments are 'flat' enough to be approximated as a straight line. If so, the test point will be projected on that line to get a start value u^*_k. This, however, is not a sufficient criterion to guarantee only one root of (5) and so far only one sufficient criterion has been reported in [12]. However, it is not trivial to implement and only covers curves in $\mathbb{R}^2$.

3.1.2 Implementation

Herein an algorithm is proposed that is tailored towards the reliable computation of $\phi(\mathbf{p}_t)$ for a larger number of test points and which can be extended easily to surfaces in $\mathbb{R}^3$. The algorithm is outlined as follows:

1. Evaluate NURBS 'often enough' at parameters u_j, so that there is not more than one solution of (5) in between each interval.

2. Determine $d_j = d(u_j) = \|\mathbf{C}(u_j) - \mathbf{p}_t\|$ for each u_j
3. Using piecewise linear interpolation of the discrete distance function d_j find its roots. This yields starting points u_j^* for the NR-iteration.
4. Using the NR-Iteration scheme (6), improve all u_j^* to get all u^* that fulfill (5).
5. Evaluate distance for all $\mathbf{C}(u^*)$ and the endpoints of $\mathbf{C}(u)$ and finally,
6. return $\phi(\mathbf{p}_t)$.

The question of 'how often is often enough' to capture all roots for (5) extensively using randomly generated curves and test points has been studied. The solutions were compared against a brute force algorithm, that iteratively subdivided the NURBS into a polygon until the projection point was determined up to a defined precision.

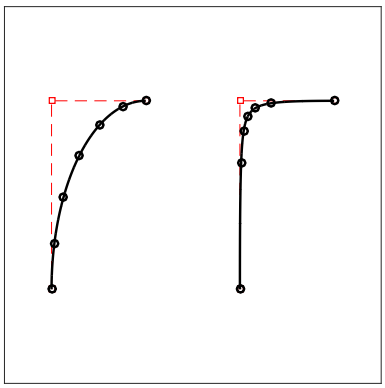


Figure 1: On the right hand curve the middle weight ($w_1/w_0 = 5$) is decreasing the uniformity of the otherwise equally (alongside u) spaced dots.

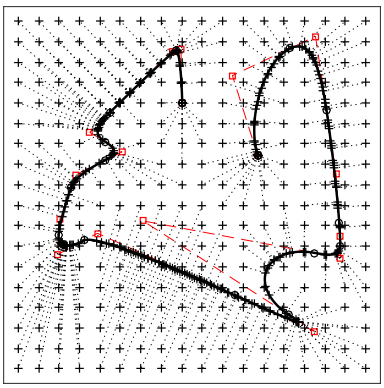


Figure 2: A curve used to validate our algorithm and showing shortest distance from gridded test points.

To arrive at the start values u_k for the bisection we subdivide the interval spanned by the knot vector by a factor $\psi = \psi_p \cdot \psi_{w,i}$ that is different for each interval of the knot vector U. It is based on a base subdivision

$$\psi_p = 2p + 1 \tag{7}$$

and a factor

$$\psi_{w,i} = \frac{w_i}{\min(\{w\})} \tag{8}$$

to account for the 'distortion effect' (see Fig. 1) caused by the weights w_i. Note that $\psi_{w,i}$ only affects the knot spans in $[i, \ldots, i + p - 1]$.

Although this approach leads to a high number of evaluation points on the NURBS, it pays off when ϕ needs to be computed for a high number of test points (i.e. nodes in the mesh). Because in that case the computationally expensive NURBS evaluations is independent from $\mathbf{p}_t$ and hence needs to be done only once for each curve. There is even a further potential of avoiding NURBS evaluations. If one is only interested in ϕ (and not for example $\mathbf{p}_p$ or derivatives of ϕ) then the result of the bisection part can be evaluated directly without the need of a Newton-Raphson iteration. That is because the point found by the bisection is already very close to the exact solution and the curve's tangent is (almost) normal to the vector between $\mathbf{p}_t$ and $\mathbf{p}_P$.

3.2 Extension to Multiple Curves

Interface descriptions may consist of multiple NURBS. It is therefore mandatory to handle (2) for more than one NURBS. Let Φ_i denote the set of $\{\phi(\mathbf{C}_i(u), \mathbf{p}_t)\}$ to a set of multiple curves, then the the sought solution of the minimal signed distance extends to

$$\phi(\mathbf{C}_i(u), \mathbf{p}_t) = \min(|\Phi_i|). \tag{9}$$

It is important to note that there might exist multiple minima for (9). From a practical point of view this is not a problem as long as all minima are of the same sign. In sets of curves which define domains, only for test points that project to endpoints of curves (i.e. in corners made of multiple curves) the found minima might be of different signs.

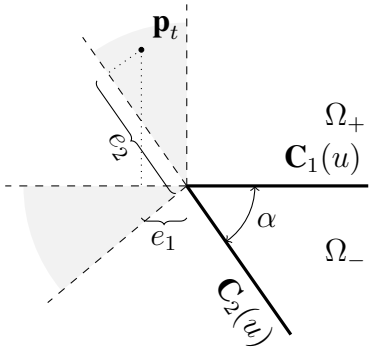


Figure 3: Situation for corners with $0 < \alpha \leqslant 90°$

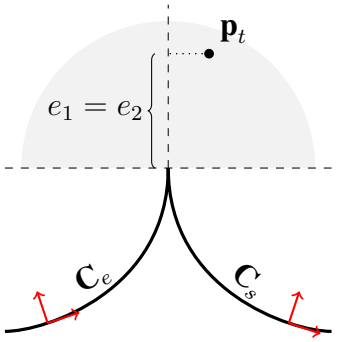


Figure 4: Situation for corners with $\alpha = 0$. The ↳ denote the tangent and normal vectors of each curve.

As seen in Fig. 3 in corners there are two zones where the sign computation (3) would return ambiguous signs for each curve. It is also seen that this zone vanishes for $\alpha > 90°$.

The curve to deliver the correct sign can be determined by

$$\begin{aligned} e_i &= \langle \mathbf{p}_t - \mathbf{p}_{corner}, \mathbf{C}_{i,u}(u) \rangle \qquad i = 1, 2 \\ s &= \text{sign}(\min(e_1, e_2)). \end{aligned} \tag{10}$$

In Fig. 3 it can also be seen, that for $\alpha = 0 \iff e_1 = e_2$. In such cases, it is useful to compare the curvatures of both curves:

$$s = \begin{cases} 1 & \text{for } \kappa_{C,e} > -\kappa_{C,s} \\ -1 & \text{otherwise} \end{cases}, \tag{11}$$

where $\kappa_{C,e}$ and $\kappa_{C,s}$ are the curvatures of the curves whose end-point (respectively, start-point) define the corner. The curvature is defined so that $\kappa > 0$ if the curve bends towards its normal vector. As long as interface lines do not overlap there exists a unique solution to (11).

4 DECOMPOSITION OF CUT ELEMENTS

Once, the signed distance to the interface has been computed for all nodes, this information is used to determine which elements are cut and and then to automatically decompose these elements into conforming sub-elements. The procedure follows ideas presented in the context of integration in [1].

The decomposition algorithm can be outlined as follows: Based on a sample grid in a reference background element, it is determined whether an element is cut or not. Using the element shape functions, the zero-level set in the reference element is detected and meshed by higher-order interface elements (blue line in Fig. 4). Based on how these elements cut the reference background element, topologically different situations result, leading to different sub-cells, see e.g. the gray and yellow elements in Fig. 4. Higher-order elements are then mapped to these sub-cells with one higher-order side so that they conform with the interface.

For a successful reconstruction, a few things need to be considered: To achieve a valid cutting pattern, recursive refinement of the cut elements might be required. To reconstruct the zero-isoline inside the element, a Newton-Raphson scheme with prescribed search paths is used. The suitability of the sub-elements for numerical analysis heavily depends on the search paths used and also on the mappings to the sub-cells.

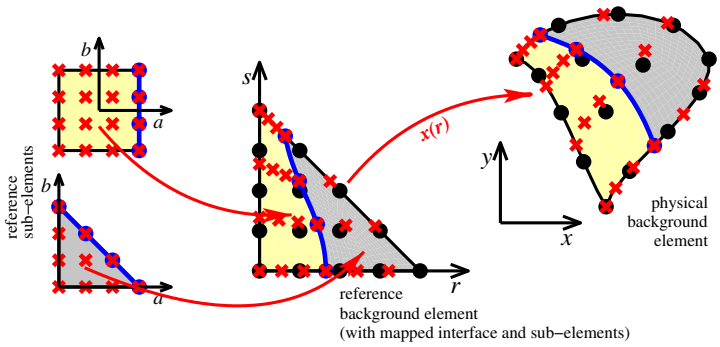


Figure 5: An example of a cut triangular element decomposed into a triangluar and a quadrilateral sub-element. Note the interface element (blue) and the additional mapping $\mathbf{r}(\mathbf{a})$.

5 NUMERICAL STUDY

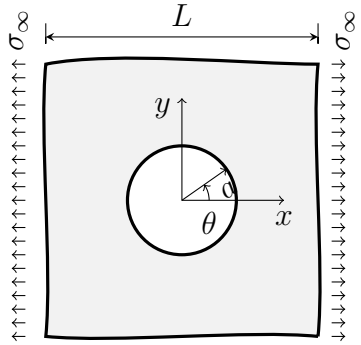


Figure 6: Infinite plate with hole under uniaxial load

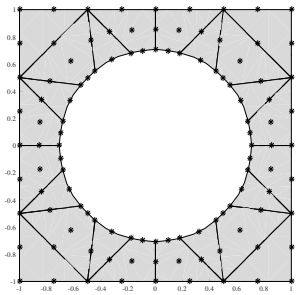


Figure 7: Coarsest mesh used in study with triangular elements in background mesh.

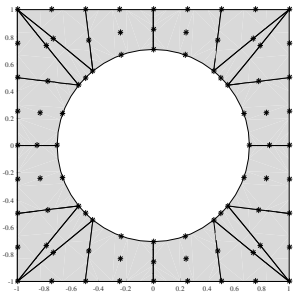


Figure 8: Coarsest mesh used in study with quadrilateral elements in background mesh.

The numerical study validates the conversion from NURBS to level-set data and demonstrates the applicability of higher order conformal decompositions. The study is carried out on an infinite plate under uniaxial load (plane strain) as shown in Fig. 6. In order to reduce the size of the computational domain, exact tractions were prescribed at the boundary. A closed

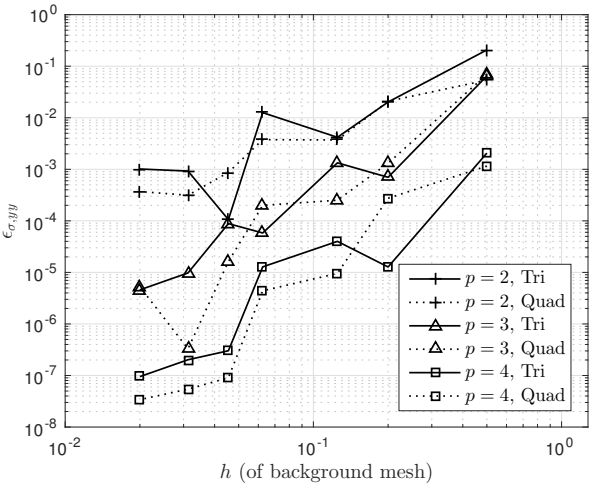


Figure 9: Error in stresses σ_{yy} on left side of hole

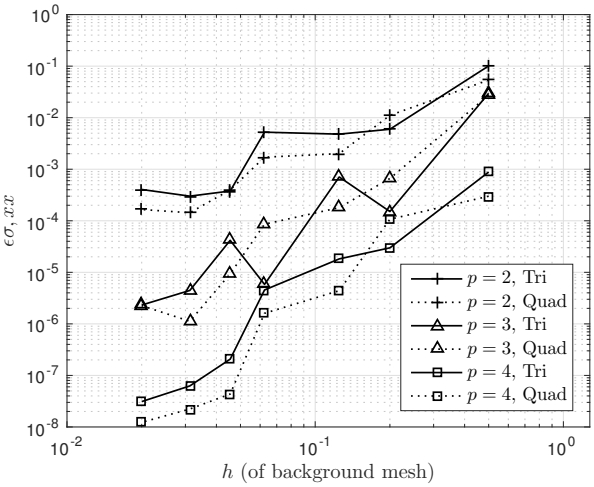


Figure 10: Error in stresses σ_{xx} on top side of hole

form solution for the stresses and displacements within the domain is found in [13]. As material parameters, a Young's modulus of $E = 10000$ and a Poisson ratio of $\nu = 0.3$ are chosen. For the hole a ratio of $L/a = 2\sqrt{2}$ is used. The rate of convergence is studied with triangular and quadrilateral background elements and element orders of $p = \{2, 3, 4\}$ and with different element sizes ($L/h = n = \{4, 10, 16, 32, 44, 64, 100\}$).

In order to avoid an ill-conditioning of the system matrix, the ratio of element areas was limited by moving nodes from the background mesh away from the interface if they were too close.

One measure is the relative error of the stresses on the left and on the top side of the hole

$$\epsilon_{\sigma_{ii}} = \left| \frac{\sigma_{ii}^{h} - \sigma_{ii}^{ex}}{\sigma_{ii}^{ex}} \right| \tag{12}$$

where σ_{xx} is evaluated at $(r = a, \theta = 90°)$ and σ_{yy} at $(r = a, \theta = 180°)$. To get an impression of the overall approximation quality, the total error

$$\epsilon_{\Omega} = \|\mathbf{u}^{ex} - \mathbf{u}^{h}\|_{L2} \tag{13}$$

is also studied.

It can be seen in Figures 9, 10, 11 that the achieved convergence rates are very good, they are optimal in the L_2-norm (13). It is interesting but not surprising that the error for the stresses at the selected points is less straight. This can most likely be attributed to the different element shapes around the nodes in question.

6 CONCLUSIONS

In this paper a 'proof of concept' for a higher-order accurate CDFEM was outlined. It promises to combine fully automated efficient meshing with trivial application of boundary conditions, optimal convergence rates and NURBS-based interface and boundary descriptions. A foundation for this method is the integration scheme proposed in [1].

Herein, the focus was on an algorithm to compute the signed distance to NURBS. Therefore a substantial effort is needed to define start values for the Newton-Raphson iteration required to obtain the minimum distance. In a numerical example, optimal convergence rates suggest that higher order convergence and conformal decomposition do work together.

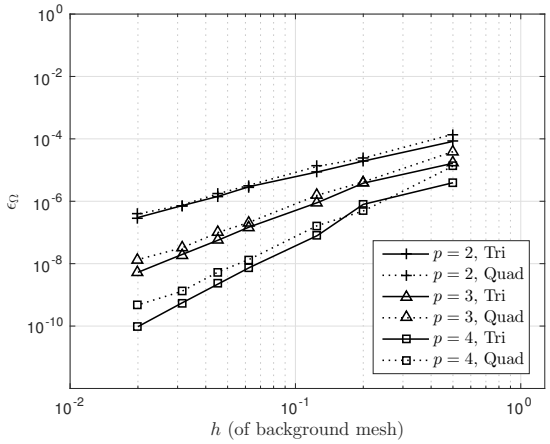


Figure 11: Overall approximation error ϵ_Ω

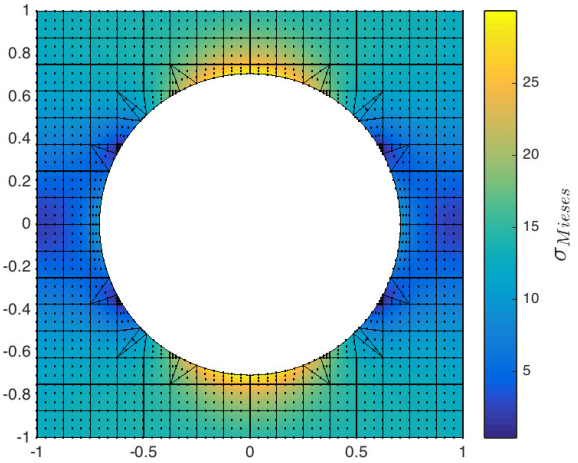


Figure 12: Distribution of von Mieses Stress based on the mesh plotted in the foreground.

The new method is, however, still in an early stage. Further steps are planned towards adaptivity and the extension to three dimensions.

REFERENCES

[1] T. Fries, S. Omerović, Higher-order accurate integration of implicit geometries. *International Journal for Numerical Methods in Engineering*, 2015. early view.

[2] K. Ho-Le, Finite element mesh generation methods: a review and classification. *Computer-aided design*, **20**(1), 27–38, 1988.

[3] T. Fries, T. Belytschko, The extended/generalized finite element method: An overview of the method and its applications. *International Journal for Numerical Methods in Engineering*, **84**, 253–304, 2010.

[4] R. Mittal, G. Iaccarino, Immersed boundary methods. *Annual Review of Fluid Mechanics*, **37**, 239 – 261, 2005.

[5] J. Parvizian, A. Düster, E. Rank, Finite cell method. *Computational Mechanics*, **41**(1), 121–133, 2007.

[6] D. R. Noble, E. P. Newren, J. B. Lechman, A conformal decomposition finite element method for modeling stationary fluid interface problems. *International Journal for Numerical Methods in Fluids*, **63**(6), 725–742, 2010.

[7] S. Osher, R. P. Fedkiw, Level set methods: An overview and some recent results. *Journal of Computational Physics*, **169**(2), 463–502, 2001.

[8] L. Piegl, W. Tiller, *The NURBS book*. Springer, 1997.

[9] Y. L. Ma, W. Hewitt. Point inversion and projection for NURBS curve: control polygon approach. In *Proceedings of Theory and Practice of Computer Graphics, 2003.*, 113–120. Institute of Electrical & Electronics Engineers (IEEE), 2003.

[10] I. Selimovic, Improved algorithms for the projection of points on NURBS curves and surfaces. *Computer Aided Geometric Design*, **23**(5), 439–445, 2006.

[11] Y.-T. Oh, Y.-J. Kim, J. Lee, M.-S. Kim, G. Elber, Efficient point-projection to freeform curves and surfaces. *Computer Aided Geometric Design*, **29**(5), 242–254, 2012.

[12] X.-D. Chen, J.-H. Yong, G. Wang, J.-C. Paul, G. Xu, Computing the minimum distance between a point and a NURBS curve. *Computer-Aided Design*, **40**(10), 1051–1054, 2008.

[13] B. Szabó, I. Babuška, *Introduction to finite element analysis: Formulation, verification and validation*. John Wiley & Sons, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A DESIGN-THROUGH-ANALYSIS APPROACH USING THE FINITE CELL METHOD (ECCOMAS CONGRESS 2016)

Benjamin Wassermann[1], **Tino Bog**[1], **Stefan Kollmannsberger**[1], **and Ernst Rank**[1,2]

[1] Technical University of Munich
Arcisstr. 21, 80333 Munich, Germany
e-mail: {benjamin.wassermann, tino.bog, kollmannsberger}@tum.de

[2] Technical University of Munich
Institute for Advanced Study
Lichtenbergstrae 2a, 85748 Garching, Germany
e-mail: ernst.rank@tum.de

Keywords: Constructive Solid Geometry, CSG, Finite Cell Method, FCM, Point-in-membership test

Abstract. *Modern 3D Computer-Aided-Design (CAD) systems use mainly two types of geometric models. Classically, objects are defined by a Boundary Representation (B-Rep), where only the objects surfaces with their corresponding edges and nodes are stored. One disadvantage concerning a numerical simulation is that B-Rep models are not necessarily water-tight. These 'dirty geometries' cause major difficulties in computational analysis because even basic geometric operations such as point-in-membership tests fail, not to mention meshing as required by classical boundary conforming finite element methods. Alternatively, objects may be represented by Constructive Solid Geometry (CSG), which is strongly related to Procedural Modeling (PM). In this context, the model is created using Boolean operations on primitives. The modeling process is then either stored as a sequence (PM), or as a construction tree (CSG). In contrast to B-Rep models, CSG models are intrinsically water-tight. To run a finite element simulation on a water-tight CSG model, two alternatives are possible: (i) it can either be converted to a B-Rep-model to obtain a finite element mesh or (ii) its implicit description can be used directly by applying an embedded domain approach, like the Finite Cell Method (FCM). In this contribution, we present a design-through analysis methodology using CSG and FCM. A crucial point in FCM is a fast and reliable point-in-membership test which can be directly derived from the CSG model. We present the outline of the modeling approach, the realization of the point-in-membership test as a sequence of CSG-operations, and discuss advantages and limitations on complex models of relevance in mechanical engineering.*

1 INTRODUCTION

Computer aided engineering often requires an an iterative process to find an optimal design. This iterative process consists of a modeling phase followed by a numerical simulation and an analysis phase. For the second phase, a common choice is the classical finite element method (FEM) in which the finite elements are conforming with the physical boundaries of the model. An estimation of the relative time required in a representative design process at Sandia National Laboratories [1] has shown, that the transition from the geometric model to the simulation model causes more than 80 % of the engineering effort.

Various methodologies have been developed to overcome the difficulties involved in this transition process. The most prominent method is Isogeometric Analysis (IGA) as proposed by Hughes et al. [2]. IGA aims at bridging the gap between the CAD model and computational analysis by using the same shape functions in CAD and FEA. To this end, B-Splines and NURBS are used. These functions offer several desirable properties such as the possibility of straightforward refinements in grid size and polynomial degree and the possibility to control the continuity within a patch. Furthermore, as B-Splines and NURBS are functions of higher order, they offer the potential to deliver high convergence rates in case of smooth solutions of the underlying problem.

Geometric models in CAD systems are often described using a boundary representation (B-Rep) [3]. Here, IGA was first applied to surface bodies, which where made up of several conforming two-dimensional B-Spline or NURBS patches. More complicated topologies are usually generated by trimming, which may lead to non-water tight geometric models. Remedies for this problem range from classic re-parametrization [4] to the use of T-Splines [5]. B-Rep solids still pose challenges for IGA, since they are defined by a collection of their bounding surfaces. Hence, the B-Rep does not provide three-dimensional patch to directly discretize the volume.

However, B-Rep is not the only possible way to represent geometries. Constructive Solid Geometry (CSG) [6] expresses the underlying construction process by combining simple solid primitives with Boolean operations. Many modern CAD systems use a hybrid representation combining B-Rep and constructive solid geometry (CSG) [7]. In this context the B-Rep model provides the additional information necessary e.g. for visualization purposes. From a design point of view, CSG offers a more intuitive approach of geometric modeling. Additionally, CSG can efficiently be used for parametric and feature based design [8] for which a description of the construction history, dependencies and constraints is mandatory. At first sight, IGA seems to be more closely related to B-Rep models. However, even before the IGA idea became popular, Natekar et al. [9] proposed a method to combine spline-based element formulations with two-dimensional CSG model descriptions. In contrast to the approach presented in the contribution at hand, heavy use is made of an explicit representation of boundaries and a decomposition into sub-domains. Recently, Zuo et al. [10] proposed an approach in which each CSG primitive is treated separately. The resulting sub-domains are then coupled with the Mortar method. This poses the difficulty that an explicit boundary representation needs to be set up also for inter-subdomain boundaries to span the Mortar boundary.

In parametric modeling a change of parameters, or constraints has hardly any impact on the CSG model, but may require a complete reconstruction of the entire corresponding B-Rep model. Hence, a simulation technique is desirable, which uses the explicit description of volumes by CSG as much and its B-Rep representation as little as possible. To this end we propose a combination of CSG and the Finite Cell Method. We denote our approach as 'design-through-analysis'

as it allows, like IGA, a very close interaction of the (geometric) design process and the (numerical) analysis, where an engineer can immediately investigate consequences of a variation of the geometric design on the mechanical behavior of a structural object.
The Finite Cell Method (FCM) was first proposed by Parzivan et al. [11]. The FCM is a high-order fictitious domain method, which embeds an arbitrary complex geometry into an extended domain which can easily be meshed by a Cartesian grid. The complexity of the geometry is handled only on the integration level. This renders the method very flexible as the only information FCM needs from the CAD model is a reliable and robust point-in-membership test, i.e. whether an integration point lies inside or outside of the physical model. This point-in-membership test is directly provided by the CSG model description. The interplay between CSG and FCM was already investigated for simple primitives and proved to be a "very accurate and efficient method for analyzing trimmed NURBS patch structures" [12]. The goal of the present paper is to extend the combination of FCM and CSG to more complex geometric models and to solid construction processes of industrial relevance.
This paper is organized as follows: In Section 2 a short overview on geometric representations and the Finite Cell Method is given. In Section 3 the relevant methods for the combination of CSG and FCM are presented. Section 4 provides an example showing the relevance and potential for practical applications.

2 OVERVIEW

This section describes concepts of geometric modeling and the Finite Cell Method in more detail.

2.1 Geometric Modeling

Modern three-dimensional CAD systems use mainly two different types of geometry descriptions. One of them is the Boundary Representation (B-Rep), where only the objects' surfaces with their corresponding edges and nodes are stored (see figure 1) [13]. Although B-Rep has several advantages, such as the direct access to the surfaces, it has also some disadvantages especially with respect to a subsequent numerical simulation. B-Rep models are not necessarily water-tight. Therefore, for these invalid solids, even basic topological operations such as a point-in-membership test fail.

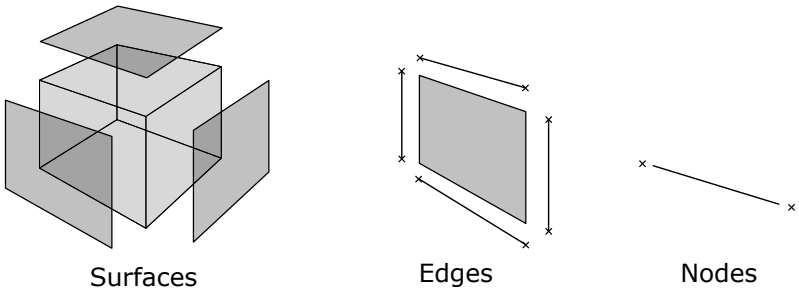


Figure 1: Boundary Representation

Alternatively a 3D object is often described as a procedural model. Procedural modeling is strongly related to Constructive Solid Geometry (CSG). In CSG a 3D object is created out of a set of primitives, such as cubes, cylinders, cones, spheres, etc. These primitives are combined

with the three Boolean operations: union, intersection and difference. The resulting CSG object is stored implicitly as a CSG tree (see figure 2).

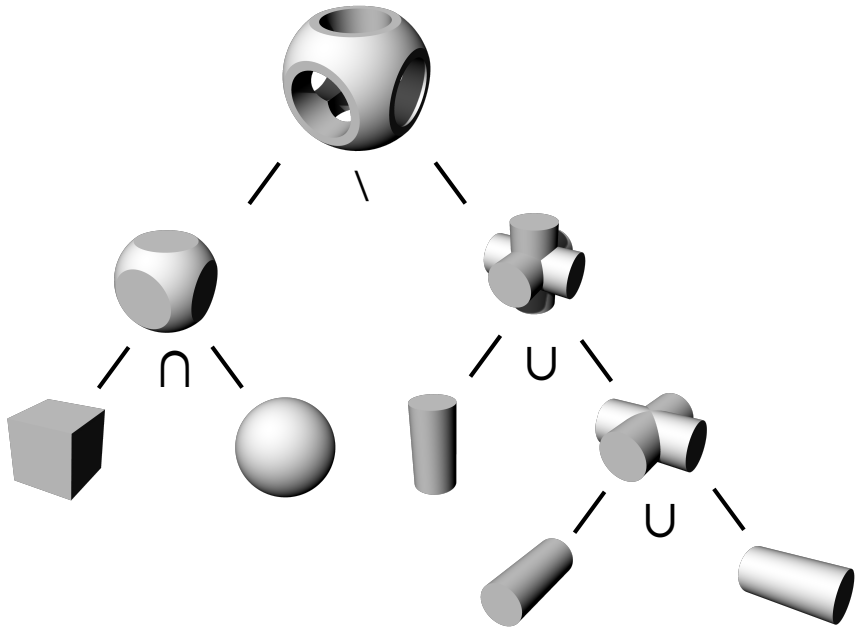


Figure 2: CSG Tree with the three Boolean operations: union $\cup$, intersection $\cap$, difference $\setminus$ on primitives

Procedural modeling extents CSG modeling with additional operations and primitives. Extra operations, such as chamfer, fillet, drilling a hole, draft, etc. are in fact just a sequence of the original three Boolean operations, which are summarized for convenience (see figure 3). A further extension allows the use of additional primitives such as sweeps, lofts or revolved objects (see figure 4).

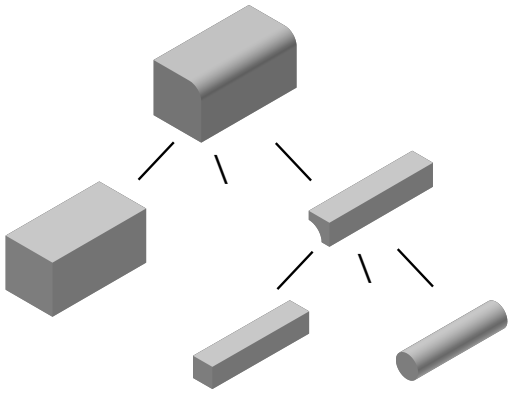


Figure 3: Extended operations can be expressed by the classical Boolean operations: union, intersection, difference. The example shows filleting an edge.

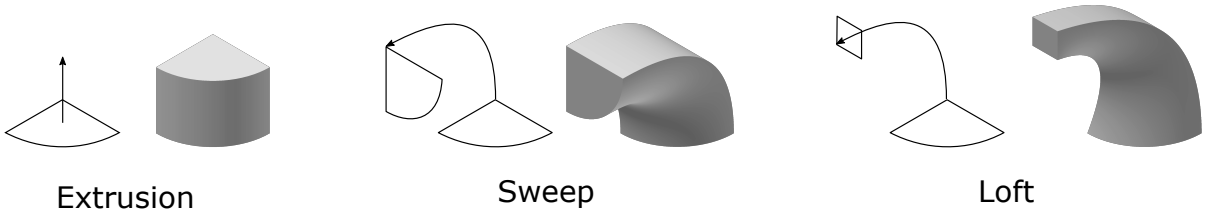


Figure 4: Extended Primitives: Extrusion, Sweep, Loft

One big advantage of the procedural CSG model over the B-Rep model regarding a numerical simulation is the inherent water-tightness of the implicit model.

2.2 Finite Cell Method

With a valid geometric description, as provided by the CSG-procedural model, it is possible to obtain a mesh on which a numerical simulation can be performed. However, mesh generators rely on a explicit (watertight) surface description of the entire boundary. This conversion may be error prone and must be conducted after each change of model. Alternatively, we aim at directly using the implicit description of the volume during the simulation process.
The Finite Cell Method is perfectly suited for this purpose, since it does not need to deal with complex geometries on the mesh level. The relevant geometric information is requested on the integration level, where for each integration point a point-in-membership test is performed. Points lying outside of the physical domain are penalized with a small factor [14].

2.2.1 Weak form

Consider a linear-elastic problem on a physical domain Ω_{phy} with the boundary $\mathrm{d}\Omega$ divided into Dirichlet and Neumann parts Γ_D and Γ_N. By applying the principle of virtual work the weak form of the underlying partial differential equation reads

$$\mathcal{B}(\mathbf{u}, \mathbf{v}) = \int_{\Omega_{phy}} \nabla \mathbf{v} : \mathbb{C} : \nabla \mathbf{u} \, \mathrm{d}\Omega \tag{1}$$

for the inner work and

$$\mathcal{F}(\mathbf{v}) = \int_{\Omega_{phy}} \mathbf{b} \cdot \mathbf{v} \, \mathrm{d}\Omega + \int_{\Gamma_N} \hat{\mathbf{t}} \cdot \mathbf{v} \, \mathrm{d}\Gamma \tag{2}$$

for the external work, where $\mathbf{u}$ is the displacement, $\mathbf{v}$ the test function and $\mathbb{C}$ the elasticity tensor. $\mathbf{b}$ and $\hat{\mathbf{t}}$ denote the body load and the prescribed boundary traction applied on the Neumann boundary, respectively.

2.2.2 Concept of FCM

In FCM the physical domain Ω_{phy} is extended by a fictitious domain Ω_{fict} in such way that the resulting domain $\Omega_{\cup}$ has a simple shape and can thus be meshed easily. (see figure 5).

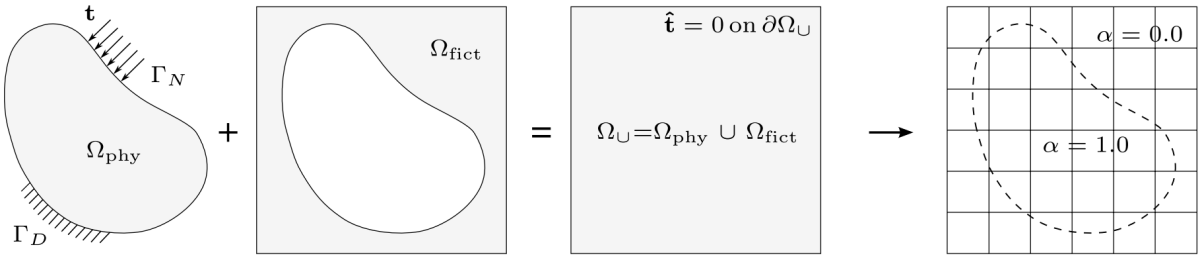


Figure 5: Concept of Finite Cell Method, taken from [15]

The weak formulation is modified by defining it over the extended domain $\Omega_\cup$. Additionally, the virtual work terms are multiplied by a scalar field $\alpha(x)$: [14]

$$\mathcal{B}_e(\mathbf{u}, \mathbf{v}) = \int_{\Omega_\cup} \nabla\mathbf{v} : \alpha\mathbb{C} : \nabla\mathbf{u} \, \mathrm{d}\Omega \tag{3}$$

$$\mathcal{F}_e(\mathbf{v}) = \int_{\Omega_\cup} \alpha\mathbf{b} \cdot \mathbf{v} \, \mathrm{d}\Omega + \int_{\Gamma_N} \hat{\mathbf{t}} \cdot \mathbf{v} \, \mathrm{d}\Gamma \tag{4}$$

with α defined as:

$$\alpha = \begin{cases} 1 & \forall \mathbf{x} \in \Omega_{phy} \\ 10^{-q} & \forall \mathbf{x} \in \Omega_{fict} \end{cases} \tag{5}$$

To minimize the influence of the fictitious domain, while not obtaining a singular system, q is typically set in the range of 5 to 10. The extended computational domain $\Omega_\cup$ is discretized by high-order finite elements, called cells. Current implementations use Lagrange polynomials, B-Splines or integrated Legendre polynomials [16]. Further, for an accurate integration of the bi-linear form, adaptive schemes are employed as presented e.g. in [17]. Recently, the discretizational framework of the finite cell method has also been extended to hierarchic refinements [18].

2.2.3 Boundary conditions

In FCM the boundaries of the physical domain Ω_{phy} typically do not coincide with the boundaries of the cells in the extended domain $\Omega_\cup$. This requires an enforcement of Neumann and Dirichlet boundary conditions in a weak sense.
Inhomogeneous Neumann conditions can be applied by integrating the prescribed traction forces $\hat{\mathbf{t}}$ along the boundary Γ_N (see equation (2)). Here, an explicit description of the boundary must be available. However, as the continuity requirements for this integration are much lower than for finite element meshes, the surface description may even be non-water-tight.
Dirichlet boundary conditions can be enforced weakly using methods like the penalty method [14], Nitsche's method [19], Discontinuous Galerkin methods [20] or Lagrange multipliers [21]. An explicit surface description must be available. The continuity requirements of that surface depend on the chosen method.

3 METHODS

3.1 Point-in-membership test

The FCM performs a point-in-membership test on the integration level i.e. at each Gaussian point. The only information which is needed from the geometric model is a reliable and fast

statement, if a point lies inside the physical domain Ω_{phy} or inside the fictitious domain Ω_{fict}. Classically this test is done on a B-Rep model casting a ray from the concerned point into an arbitrary direction. All intersections with the boundary are then counted. The point lies inside the domain, if the number of intersections is odd, and outside otherwise. This classic test has the drawback that it fails for non-water tight B-Rep models and its computational effort rises with the complexity of the surface description, i.e. a model consisting of many non-planar surfaces or highly resolved surface triangularizations.
By contrast, a point-in-membership test can be performed much faster on a CSG tree. Here, the root element is queried, which forwards the request to (a selection of) its children. Since the tree is build from primitives the individual tests are very cheap. Therefore, the complexity is at its worst proportional to the number of bodies involved which is in general orders of magnitude lower than the number of surfaces for all practical applications.

3.2 Point-in-membership test on primitives

For classical primitives a simple analytical function is available. Consider a primitive $\mathbf{B}_i$ which is created axis-aligned on the $x-y$ plane and assume that we define each primitive as a closed body, i.e. the boundary is included in the body. The test whether a point $\mathbf{P} = \{x, y, z\}$ is inside a primitive reads as follows for a:

- **Sphere** with center point $\mathbf{C}_{Shpere}$ and radius r_0

$$\mathbf{P} \in \mathbf{B}_{Sphere} \quad iff \qquad ||\overline{\mathbf{PC}}_{Sphere}||_2 \leq r_0, \tag{6}$$

- for a **Cuboid** defined by two corner points lying on its diagonal $P_{start} = [x_s, y_s, z_s]$ and $P_{end} = [x_e, y_e, z_e]$

$$\mathbf{P} \in \mathbf{B}_{Cuboid} \quad iff \qquad x \in [x_s, x_e] \wedge y \in [y_s, y_e] \wedge z \in [z_s, z_e], \tag{7}$$

- and for a **Cylinder** defined by its center point $\mathbf{C}_{Cylinder} = \{x_c, y_c, z_c \equiv 0\}$, radius r_0, and height h_0

$$\mathbf{P} \in \mathbf{B}_{Cylinder} \quad iff \qquad ||\overline{\tilde{\mathbf{P}}\mathbf{C}}_{Cylinder}||_2 \leq r_0 \wedge z \in [0, h_0] \tag{8}$$

  where point $\tilde{\mathbf{P}} = \{x, y, 0\}$ is the projection of point $\mathbf{P}$ onto the $x-y$ plane.

There are also fast analytical solutions for other primitives like cones, pyramids, tori and frustums. However, it is not likely that these primitives are only constructed axis-aligned on the $x-y$ plane. Therefore, at a suitable position a local orthonormal coordinate system A is constructed. It is spanning a work plane on which the respective primitive can be constructed. To perform a point-in-membership test, the point of interest P needs to be mapped from the Cartesian space E to the local base A

$$\mathbf{P_A} = Q_{EA} \cdot \mathbf{P_E} + \mathbf{v} \tag{9}$$

with v the transposition vector between the center points of the Cartesian and local basis system

$$\mathbf{v} = \mathbf{C_A} - \mathbf{C_E} = \mathbf{C_A} \tag{10}$$

and

$$Q_{EA} = \begin{bmatrix} A_{1x} & A_{2x} & A_{3x} \\ A_{1y} & A_{2y} & A_{3y} \\ A_{1z} & A_{2z} & A_{3z} \end{bmatrix} \tag{11}$$

with $\mathbf{A_i}$ being the base vectors of the local basis system A.

3.3 Point-in-membership test on sweeps

Point-in-membership tests are more involved for ”primitives” generated by sweeps or lofts. Typically, no analytic solution is available for those primitives. Nevertheless, it is possible to perform a fast, reliable test also on these bodies. The basic idea is to reduce the dimension of the problem. For this consider the set-up of a sweep. A sweep consists of a 2D sketch, which is moved along a sweep path. (see figure 6)

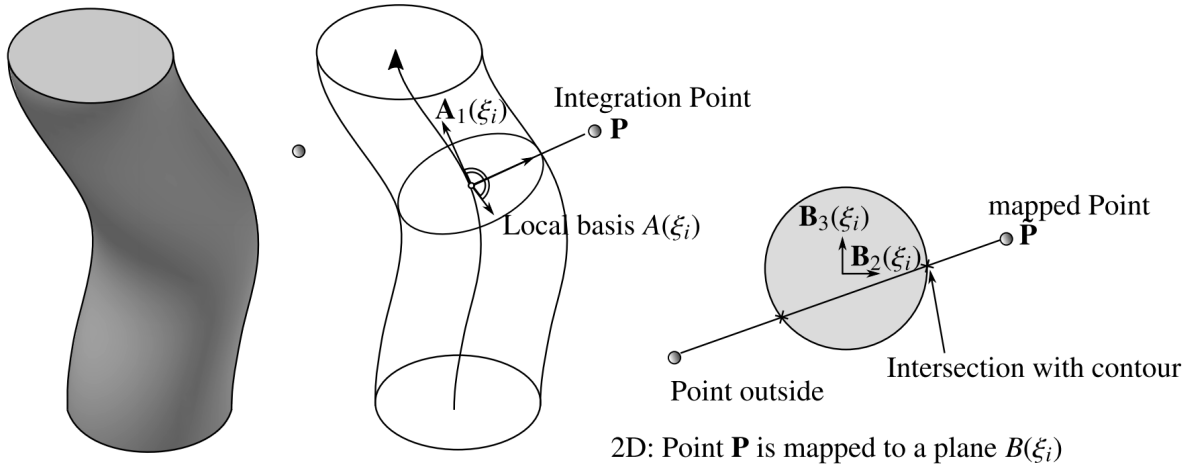


Figure 6: Point-in-membership test on intermediate sketch of a loft

For the simple case that (i) the sweep path is orthogonal to the sketch plane of the starting sketch and (ii) the local basis system follows the tangent of the path, a point-in-membership carries out the following steps:

- The closest point $\mathbf{C}(\xi_{cp})$ on the sweep path is computed either analytically or, if not possible, using Newton's method:

$$f(\xi) = \dot{\mathbf{C}}(\xi) \cdot (\mathbf{P} - \mathbf{C}(\xi)) \equiv 0 \tag{12}$$

$$\xi_{i+1} = \xi_i - \frac{f(\xi_i)}{f'(\xi_i)} = \xi_i - \frac{(\dot{\mathbf{C}}(\xi_i) \cdot (\mathbf{P} - \mathbf{C}(\xi_i))}{\ddot{\mathbf{C}}(\xi_i) \cdot (\mathbf{P} - \mathbf{C}(\xi_i)) + |\dot{\mathbf{C}}(\xi_i)|^2} \tag{13}$$

  where $\mathbf{C}$ denotes an arbitrary curve description with its first and second derivative ($\dot{\mathbf{C}}$, $\ddot{\mathbf{C}}$) and $\mathbf{P}$ being the point of interest.

- On the closest point an auxiliary plane $WP(\xi_{cp})$ is created. For this purpose, the tangent vector at $\mathbf{C}(\xi_{cp})$ is evaluated and a local base system is created using e.g. the Frenet base [22], or a base system, where one base vector is always in a plane parallel to an arbitrary plane.

- The point of interest $\mathbf{P}$ is mapped to the local coordinate system of the plane to obtain $\tilde{\mathbf{P}}$.

- A point-in-membership test using a ray-test is performed in 2D with the contour line. In case of a sweep a ray cast on the initial swept object, i.e. the two-dimensional curve is carried out.

The fall back to a point-in-membership test in two dimensions poses a draw-back at first sight. However, a two-dimensional ray test is much simpler and more robust to implement than a general three-dimensional one. Clearly, a necessary pre-requisite is that the curve is closed to produce a closed object during the sweep operation. Again, this proves to be much simpler than assuring closed surfaces in three dimensions.

4 EXAMPLE

The following example was constructed as a procedural model and then transformed to a CSG tree (see figure 7). It combines several simple primitives and two sweeps along a B-Spline ($p = 2$) curve. The (round) base plate was fixed and a predefined deflection $\hat{\mathbf{u}} = 1$ was applied onto the left (quadratic) base plate. Homogeneous Neumann boundary conditions were applied to all other surfaces.

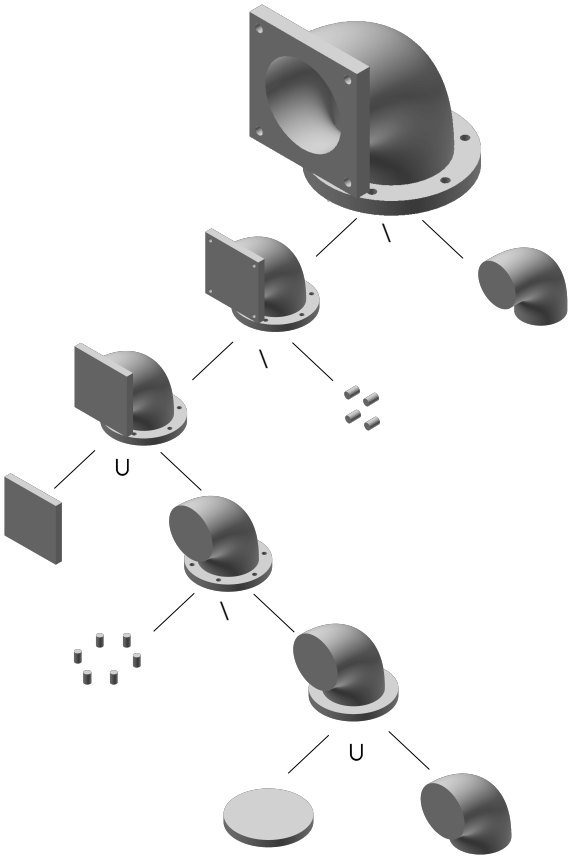


Figure 7: CSG tree of example

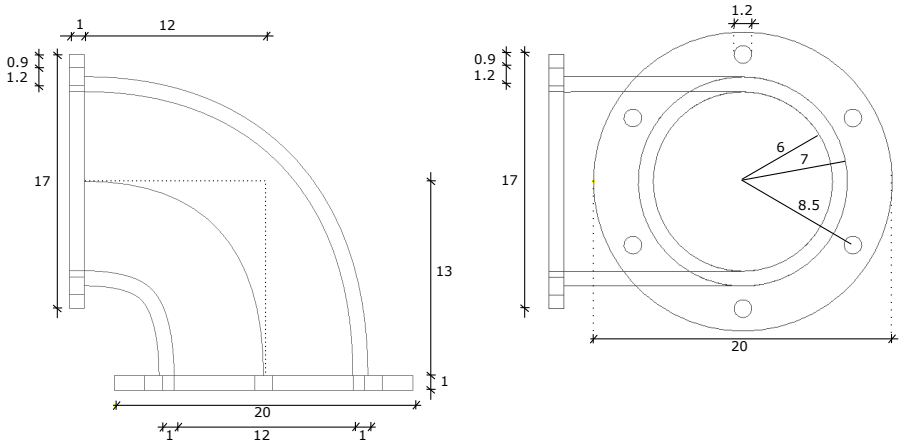


Figure 8: Dimensions of the example

For the simulation in each direction 15 cells with (integrated) Legendre shape functions and a polynomial degree of $p = 4$ were chosen. For a precise integration of the stiffness matrix the cells were partitioned with an octree to a maximum depth of four subdivisions. Only cells containing parts of the physical domain were considered. This reduces the number of degrees of freedom by 72 % to 55,296 dofs.

Figure 9 shows the finite cells embedding the structural model and the computed displacements for the example. The von Mises stresses are depicted in figure 10. These stresses provide a good overall insight into the structural load carrying behavior. Local stresses are not always fully resolved for example at re-entrant corners. A locally very accurate resolution of these singularities is possible by application of hierarchical refinements as recently developed in [23]. Further extensions include the development of point-in-membership tests for more complex CSG models such as lofts.

It is noteworthy that only the CSG model was used in all involved steps, i.e. from the setup of the model until the computation itself. The only point at which a conversion from the CSG-model to an explicit B-rep was carried out was for the post-processing step. Here, the marching cubes algorithm was used to derive a triangulated surface on which the results were post-processed [24]. However, even this conversion is not mandatory as volumetric post-processing is a possible option as well.

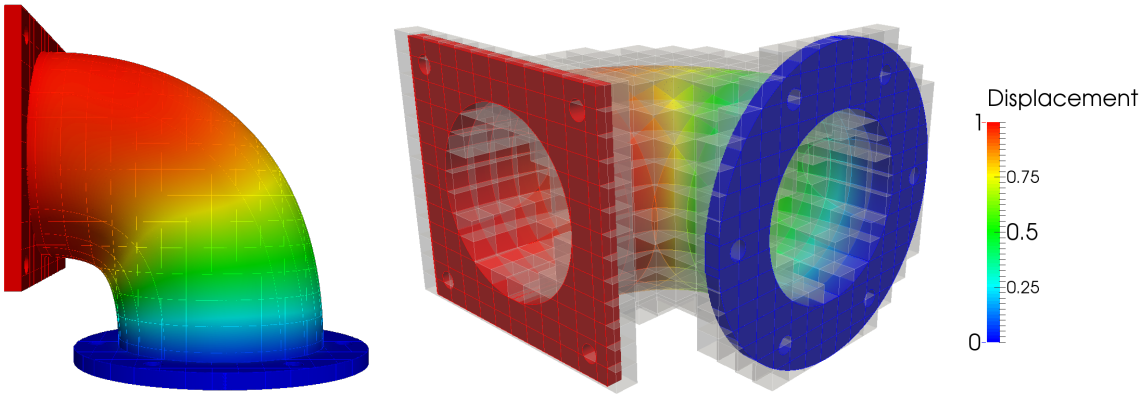


Figure 9: Displacement

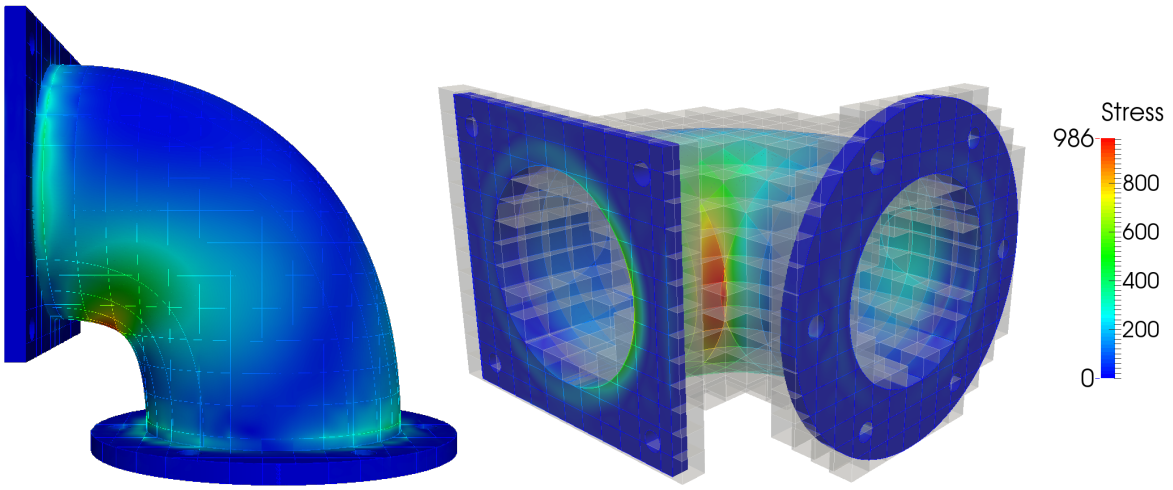


Figure 10: Stresses

5 CONCLUSION

An integration of the design process and numerical analysis without a complex transition like meshing is of high relevance for the industry and has been in the focus of several research groups for the last years. While Isogeometric Analysis provides an excellent method for the numerical simulation of boundary representation models and shell structures, this paper has focused on models created with Constructive Solid Geometry. A design-through analysis approach combining CSG modelling and the Finite Cell Method (FCM) has been presented. FCM is able to use the implicit model description provided by the CSG model and hence greatly simplifies the meshing process. It was shown that point-in-membership tests can be carried out efficiently for complex geometries like sweeps. Further steps include the development of point-in-membership tests for other primitives like lofts and sweeps with rotated sketches.

ACKNOWLEDGMENTS

This work has been founded by the German Research Foundation (Project RA 624/22-1). This support is gratefully acknowledged.

REFERENCES

[1] J. Cottrell, T. J. Hughes, and Y. Bazilevs, *Isogeometric Analysis: Toward Integration of CAD and FEA*. New York: Wiley and Sons, 2009.

[2] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs, "Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement," *Computer Methods in Applied Mechanics and Engineering*, vol. 194, no. 39–41, pp. 4135–4195, Oct. 2005.

[3] J. D. Foley, A. V. Dam, S. K. Feiner, J. F. Hughes, and R. L. Phillips, *Introduction to Computer Graphics*. Addison-Wesley, 1997.

[4] L. Piegl and W. Tiller, *The NURBS Book*, ser. Monographs in Visual Communication. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997.

[5] Y. Bazilevs, V. M. Calo, J. A. Cottrell, J. A. Evans, T. J. R. Hughes, S. Lipton, M. A. Scott, and T. W. Sederberg, "Isogeometric analysis using T-splines," *Computer Methods in Applied Mechanics and Engineering*, vol. 199, no. 5–8, pp. 229–263, Jan. 2010.

[6] A. A. G. Requicha and H. B. Voelker, *Constructive Solid Geometry, TM-25*. Production Automation Project, University of Rochester, 1977.

[7] A. J. P. Gomes and J. G. Teixeira, "Form feature modelling in a hybrid CSG/BRep scheme," *Computers & Graphics*, vol. 15, no. 2, pp. 217–229, 1991.

[8] J. J. Shah and M. Mäntylä, *Parametric and Feature-Based CAD/CAM: Concepts, Techniques, and Applications*. John Wiley & Sons, 1995.

[9] D. Natekar, X. Zhang, and G. Subbarayan, "Constructive solid analysis: a hierarchical, geometry-based meshless analysis procedure for integrated design and analysis," *Computer-Aided Design*, vol. 36, no. 5, pp. 473–486, Apr. 2004.

[10] B.-Q. Zuo, Z.-D. Huang, Y.-W. Wang, and Z.-J. Wu, "Isogeometric analysis for CSG models," *Computer Methods in Applied Mechanics and Engineering*, vol. 285, pp. 102–124, Mar. 2015, 00000.

[11] J. Parvizian, A. Düster, and E. Rank, "Topology optimization using the finite cell method," *Optimization and Engineering*, vol. 13, no. 1, pp. 57–78, Jul. 2011.

[12] E. Rank, M. Ruess, S. Kollmannsberger, D. Schillinger, and A. Düster, "Geometric modeling, isogeometric analysis and the finite cell method," *Computer Methods in Applied Mechanics and Engineering*, vol. 249-252, pp. 104–115, Dec. 2012.

[13] H.-J. Bungartz, M. Griebel, and C. Zenger, *Introduction to Computer Graphics*. Charles River Media, 2004.

[14] A. Düster, J. Parvizian, Z. Yang, and E. Rank, "The finite cell method for three-dimensional problems of solid mechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 197, no. 45–48, pp. 3768–3782, Aug. 2008.

[15] D. Schillinger, M. Ruess, N. Zander, Y. Bazilevs, A. Düster, and E. Rank, "Small and large deformation analysis with the p- and B-spline versions of the Finite Cell Method," *Computational Mechanics*, vol. 50, no. 4, pp. 445–478, Feb. 2012.

[16] E. Stein, *Encyclopedia of computational mechanics. 2, 2,.* Chichester: Wiley, 2004.

[17] D. Schillinger and E. Rank, "An unfitted hp-adaptive finite element method based on hierarchical B-splines for interface problems of complex geometry," *Computer Methods in Applied Mechanics and Engineering*, vol. 200, no. 47-48, pp. 3358–3380, Nov. 2011.

[18] D. Schillinger, A. Düster, and E. Rank, "The hp-d-adaptive finite cell method for geometrically nonlinear problems of solid mechanics," *International Journal for Numerical Methods in Engineering*, vol. 89, no. 9, pp. 1171–1202, 2012.

[19] M. Ruess, D. Schillinger, Y. Bazilevs, V. Varduhn, and E. Rank, "Weakly enforced essential boundary conditions for NURBS-embedded and trimmed NURBS geometries on the basis of the finite cell method," *International Journal for Numerical Methods in Engineering*, vol. 95, no. 10, pp. 811–846, Sep. 2013.

[20] S. Kollmannsberger, A. Özcan, J. Baiges, M. Ruess, E. Rank, and A. Reali, "Parameter-free, weak imposition of Dirichlet boundary conditions and coupling of trimmed and non-conforming patches," *International Journal for Numerical Methods in Engineering*, vol. 101, no. 9, pp. 670–699, Mar. 2015, 00001.

[21] E. Burman and P. Hansbo, "Fictitious domain finite element methods using cut elements: I. A stabilized Lagrange multiplier method," *Computer Methods in Applied Mechanics and Engineering*, vol. 199, no. 41-44, pp. 2680–2686, Oct. 2010.

[22] S. Sternberg, *Lectures on differential geometry*. Prentice-Hall, 1964.

[23] N. Zander, T. Bog, S. Kollmannsberger, D. Schillinger, and E. Rank, "Multi-level hp-adaptivity: high-order mesh adaptivity without the difficulties of constraining hanging nodes," *Computational Mechanics*, vol. 55, no. 3, pp. 499–517, Feb. 2015.

[24] William E. Lorensen and Harvey E. Cline, "Marching Cubes: A High Resolution 3d Surface Construction Algorithm," *Computer Graphics,*, vol. 21, no. 4, p. 7, Jul. 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DISCONTINUOUS GALERKIN COMPUTATION OF GASEOUS MIXTURE COAXIAL JETS

N. Franchina[1], **M. Savini**[1], **and F. Bassi**[1]

[1] Università degli Studi di Bergamo, Dipartimento di Ingegneria e Scienze Applicate
viale Marconi, 5 Dalmine BG, Italy
e-mail: {nicoletta.franchina,marco.savini,francesco.bassi}@unibg.it

Keywords: Discontinuous Galerkin, high-order accurate discretization, multicomponent flow, pressure L^2 projection, preferential diffusion, coaxial jets

Abstract. *A novel approach based upon Discontinuous Galerkin (DG) discretization, applied to the divergence form of the multicomponent Navier-Stokes equations, is here presented and used to compute non reactive turbulent axisymmetric gaseous jets. The original key feature is the use of L^2-projection form of the (perfect gas) equation of state. This choice mitigates problems typically encountered by the front-capturing schemes in computing multicomponent flow fields, i.e. spurious oscillations across material and contact surfaces where the mixture composition is changing. The solver makes also use of a shock-capturing technique based on artificial dissipation selectively added into the equations and tuned in connection with the magnitude of inviscid residuals of the equations and on suitable coefficients accounting for the variation of the unknown variables within and across grid elements. A simple limiting procedure is introduced in order to avoid the occurrence of unphysical gas properties due to negative and/or greater that one mass fractions values within the domain. The DG code based on the proposed novel technique for multicomponent flow computation is here employed to study the mixing mode and the preferential diffusion mechanism of a mixture jet of helium and carbon dioxide in a surrounding flow of air, both in laminar and turbulent flow regimes. Mass diffusion is modelled by means of Fick's first law and use is made of constant Prandtl and Schmidt numbers in Wilcox's (2008) $k-\omega$ model. Third-order accurate results are presented, discussed and compared with the available experimental data. They confirm the possible existence in coaxial jets of different periodic flow structures, greatly affecting mixing rates, and different species diffusive mass fluxes. The relative importance of both phenomena depends on the flow regime and its characteristics. The tests carried out give at the same time indications about the accuracy of the proposed method and its effectiveness in computing complex unsteady flow fields.*

1 INTRODUCTION

The increased power and enhanced performance of the available computational resources achieved during these last years allowed a more widespread use of high-order accurate discretizations for the numerical prediction of complex flow fields of academic and industrial interest. The discontinuous Galerkin finite element method is one of the most promising numerical method due to the great geometrical flexibility, good scalability for high-performance computing and inherent high-order accuracy.

In the current work, the development of the DG code MIGALE [1, 2, 3, 4] for the computation of multicomponent flows, is presented. The solver is considered as one of our preliminary steps to deal with homogeneous (gas phase) combustion problems by means of a high-order CFD-DG tool and at present it neglects the modeling of chemical reactions.

The increasing importance of an accurate prediction of combustion processes, especially for the control of pollutant emissions, is nowadays indeed demanding further development of high-fidelity solvers, accurate enough to trace the evolution of chemical species down to ppm scale and, at the same time, robust and reliable enough to be applied to problems of industrial interest. The DG approach marries the ability to allow discontinuities in a natural way with its great accuracy granted even on general meshes, both important properties to compute the wrinkled flame fronts encountered in turbulent combustion, and indeed reactive DG codes have recently appeared (see *e.g.* [5, 6]).

The study of coaxial jets, with and without swirl, is of great importance in the assessment of the burner efficiency and it is an essential requirement to subsequently investigate the burner behaviour with partially premixed and non-premixed flames. In addition to this, the impact of coaxial jets fluid dynamics is crucial not only on the performances of rockets and combustion devices, but also in non-reactive components such as injectors and exhausts of turbofan engines. Studies on the behaviour of coaxial jets date back to the seventies with the pioneering experimental investigations presented in [7, 8, 9]; recent works [10, 11] confirm the existence of different flow configurations depending on the velocity and density ratios between inner and outer jets, as well as on geometrical details of the apparatus. The two main distinctive features are the "wake-like" and "jet-like" turbulent shear structures that, roughly speaking, develop respectively when the jets velocity ratio is near unity or far away from this value. Typically the flow field is turbulent and unsteady, with large scale fluctuations generated by Kelvin-Helmoltz instabilities at the shear interface between the jets, superimposed on the broad-band small scale turbulent background; thus the accurate simulation of gaseous coaxial jets represents indeed a challenging task for any multicomponent flow solver. Moreover, the growing interest in the use of syngas derived both from biomasses and coal, possibly added by methane in order to raise its enthalpy content (LHV), and/or in methane enrichment by hydrogen (to increase its flame speed and its lean stability limit) in turbogas burners, draws the attention toward the study of the effects induced by the faster mixing of one species (*e.g.* H_2), as compared with other species, on the whole process of combustion. As shown in [12], these effects are quite significant at low Reynolds numbers.

Numerical investigation on the mixing and diffusion phenomena in a coaxial jet of helium and carbon dioxide in air, at two different velocity ratios, is herein carried out using the multicomponent DG MIGALE code. The adopted DG solver is based on the "exact" Riemann solver [13] and on the BR2 scheme [1] for the computation of the interface convective and diffusive flux vector, respectively. It is well-known that the discretization of multicomponent flow fields has to cope with numerical oscillations that are not encountered with single component

flows. Unphysical fluctuations may appear near contact and material interfaces, where the mixture composition changes, even with low-order accurate discretizations, depending on the sort of approximation adopted, see *e.g.* [14, 15]. Following the approach presented in [16] to mitigate spurious pressure oscillations at contact/material interface, the pressure is computed from the weak form of the equation of state for ideal gases. This can be viewed as a "regularization" of pressure that reduces its polynomial order of representation. Numerical oscillations that may also occur, due to flow field discontinuities, are treated by means of a shock-capturing scheme that explicitly adds artificial dissipation based on the inviscid residuals of the equations.

Two unsteady test cases have been computed, with jet-to-coflow velocity ratio V ranging from 5 to 1, and their results compared with experiments carried out at the Politecnico di Milano [17]. The computations allow a twofold check on the accuracy of the developed code, capable to investigate both the different mixing modes in coaxial jets and the concurrent mechanism of preferential diffusion of mass species.

In Section 2 we first describe the governing system of equations adopted for multicomponent gas mixture and the formulas employed to compute the mixture properties, then the DG approximation, the artificial dissipation and limiting procedure are presented. Section 3 reports the computed solutions along with a discussion on the jet structures and the mixing of chemical species. Final remarks are summarized in Section 4.

2 NUMERICAL METHOD

The numerical discretization presented in this section applies to the conservative form of the RANS equations. Their system for turbulent non reactive multicomponent flows is obtained through Favre averaging, neglecting the contributions due to the product of temperature and species mass fractions fluctuations. The system is written for both planar and axisymmetric 2D flows as follows

$$\frac{\partial}{\partial t}\left(\rho r\right)+\frac{\partial}{\partial x}\left(\rho r\mathrm{v}_x\right)+\frac{\partial}{\partial r}\left(\rho r\mathrm{v}_r\right)=0, \tag{1}$$

$$\frac{\partial}{\partial t}\left(\rho r\mathrm{v}_x\right)+\frac{\partial}{\partial x}\left[r\left(p+\rho\mathrm{v}_x^2\right)\right]+\frac{\partial}{\partial r}\left(\rho r\mathrm{v}_x\mathrm{v}_r\right)=\frac{\partial}{\partial x}\left(r\tau_{xx}\right)+\frac{\partial}{\partial r}\left(r\tau_{xr}\right), \tag{2}$$

$$\frac{\partial}{\partial t}\left(\rho r\mathrm{v}_r\right)+\frac{\partial}{\partial x}\left(\rho r\mathrm{v}_x\mathrm{v}_r\right)+\frac{\partial}{\partial r}\left[r\left(p+\rho\mathrm{v}_r^2\right)\right]-\phi p=\frac{\partial}{\partial x}\left(r\tau_{xr}\right)+\frac{\partial}{\partial r}\left(r\tau_{rr}\right), \tag{3}$$

$$\frac{\partial}{\partial t}\left(\rho re_t\right)+\frac{\partial}{\partial x}\left(\rho rh_t\mathrm{v}_x\right)+\frac{\partial}{\partial r}\left(\rho rh_t\mathrm{v}_r\right)=A-rP+\beta^\star\rho r\bar{k}\mathrm{e}^{\widetilde{\omega}_r}+\frac{\partial}{\partial x}\left(rh_iJ_{i,x}\right)+\frac{\partial}{\partial r}\left(rh_iJ_{i,r}\right) \tag{4}$$

$$\frac{\partial}{\partial t}\left(\rho rk\right)+\frac{\partial}{\partial x}\left(\rho r\mathrm{v}_xk\right)+\frac{\partial}{\partial r}\left(\rho r\mathrm{v}_rk\right)=\frac{\partial}{\partial x}\left(\bar{\mu}_kr\frac{\partial k}{\partial x}\right)+\frac{\partial}{\partial r}\left(\bar{\mu}_kr\frac{\partial k}{\partial r}\right)+rP-\beta^\star\rho r\bar{k}\mathrm{e}^{\widetilde{\omega}_r}, \tag{5}$$

$$\frac{\partial}{\partial t}\left(\rho r\widetilde{\omega}\right)+\frac{\partial}{\partial x}\left(\rho r\mathrm{v}_x\widetilde{\omega}\right)+\frac{\partial}{\partial r}\left(\rho r\mathrm{v}_r\widetilde{\omega}\right)=\frac{\partial}{\partial x}\left(\bar{\mu}_\omega r\frac{\partial\widetilde{\omega}}{\partial x}\right)+\frac{\partial}{\partial r}\left(\bar{\mu}_\omega r\frac{\partial\widetilde{\omega}}{\partial r}\right)+\frac{\alpha r\widetilde{\omega}}{\bar{k}}P-\beta\rho r\mathrm{e}^{\widetilde{\omega}_r}, \tag{6}$$

$$\frac{\partial}{\partial t}\left(\rho ry_i\right)+\frac{\partial}{\partial x}\left(\rho r\mathrm{v}_xy_i\right)+\frac{\partial}{\partial r}\left(\rho r\mathrm{v}_ry_i\right)=\frac{\partial}{\partial x}\left(rJ_{i,x}\right)+\frac{\partial}{\partial r}\left(rJ_{i,r}\right),\qquad i=1,\ldots,\mathrm{N}-1, \tag{7}$$

where ρ, p, $\mathbf{v}=(\mathrm{v}_x,\mathrm{v}_r)$, e_t, $h_t=e_t+p/\rho$, are the density, pressure, velocity, total internal energy and total enthalpy of the mean motion respectively, while y_i is the mass fraction of the

i-th species. The inclusion of the geometrical factor r allows to use a unique implementation; r is equal to the radius in the axisymmetric case and equal to one in the planar case, whilst the coefficient ϕ in Eq. (3) is equal to one and zero in the former and latter case, respectively.

Equations (1) and (7) express the conservation of the whole mass mixture and of $\mathrm{N}-1$ mass species, while Equations (2)–(4) express the momentum and energy conservation. As the total mass of the mixture is the sum of the mass of all species, the last (N-th) species conservation equation could be added to (7) in place of Eq. (1). Our choice, which is widely used in academic and commercial codes, ensures the global mass conservation, unaffected by chemical reactions and diffusive transport, and implicitly assumes that

$$y_{\mathrm{N}} = 1 - \sum_{i=1}^{\mathrm{N}-1} y_i, \quad y_i \geq 0 \quad \forall i = 1, \ldots, \mathrm{N} \tag{8}$$

instead of computing the mixture density as $\rho = \sum_{i=1}^{\mathrm{N}} \rho_i$.

The term A in the energy equation (4) summarizes the divergence of the work of laminar stresses and the divergence of the "effective" heat flux vector $\mathbf{q}_h^e$, defined as

$$\mathbf{q}_h^e = -\left(k + \frac{\mu_t c_p}{Pr_t}\right) \nabla T. \tag{9}$$

P is the production term that models the energy exchange between mean and turbulent field (see [18] for their complete expressions).

The diffusive mass flux of the i-th species J_i of Eqs. (4) and (7) is computed according to the Fick's first law

$$J_i = -\rho D_{i,m}^e \nabla y_i, \tag{10}$$

where $D_{i,m}^e$ is the "effective" mass diffusion coefficient of the species i in the mixture obtained as the sum of molecular and turbulent contributions,

$$D_{i,m}^e = D_{i,m} + \mu_t/(\rho Sc_t),$$

and Sc_t is the turbulent Schmidt number introduced to model the product of species and velocity fluctuations through the classical gradient diffusion hypothesis. The $D_{i,m}$ coefficient is computed using a modified form of the Stefan's formula reported in [19], where y_i has been replaced by X_i (molar fraction of the i-th species) for self-consistency, *i.e.* at least when all $D_{ij} = D$, the formula guarantees that $D_{im} = D$, and hence $\sum_{i=1}^{N} J_i = 0$,

$$D_{i,m} = \frac{1 - X_i}{\displaystyle\sum_{j=1,\, j\neq i}^{\mathrm{N}} \frac{X_j}{D_{ij}}}. \tag{11}$$

The binary diffusion coefficients D_{ij} are approximated as

$$D_{i,j} = D_{(i,j)_0} \frac{(T/T_0)^n}{p/p_0}, \tag{12}$$

where $D_{(i,j)_0}$ is a reference diffusion coefficient at a reference state given by the pressure and temperature values p_0 and T_0 ([20]). Due to the constraint on the sum of mass diffusive fluxes,

$J_{\mathrm{N}} = -\sum_{i=1}^{\mathrm{N}-1} J_i$, the mass diffusivity coefficient of the last species in the mixture $D_{\mathrm{N},m}$ is implicitly defined as

$$D_{\mathrm{N},m} = \frac{\sum_{i=1}^{\mathrm{N}-1} D_{i,m} \nabla y_i}{\sum_{i=1}^{\mathrm{N}-1} \nabla y_i}. \tag{13}$$

Note that, being working with non-reactive flows, the mass diffusion due to temperature gradients (Soret effect), as well as terms related to pressure gradients and body forces, have been neglected.

The diffusion flux J_i gives rise to an energy flux $h_i J_i$ which, summed over all the species present in the mixture, provides the last terms in the RHS of Eq. (4), where the summation convention over repeated indices is adopted.

The turbulence model employed is the "high-Reynolds" version of Wilcox's $k-\omega$ turbulence model [21, 22]. Notice that the model adopted is not in its standard form and employs the modifications introduced in [2], *i.e.*, in Equations (4)-(6) the turbulent kinetic energy k is replaced by a limited value $\bar{k}$ to avoid the occurrence of negative values; besides the model uses $\widetilde{\omega} = \log \omega$ instead of ω. Furthermore, "realizable" $\widetilde{\omega}_r$ (*i.e.* properly limited $\widetilde{\omega}$) values are used in the source term of Equation (6) and in the eddy viscosity computation to guarantee physical turbulent stress values. The standard RANS model employed makes use of the simple assumption of constant turbulent Prandtl number and of unique constant turbulent Schmidt number for each species. In this work they are fixed at 0.9 and 0.75, respectively.

The dynamic viscosity μ and thermal conductivity λ of the mixture are computed using the Wilke's formula [23] and the corrected expression given in [24] respectively, (see [25])

$$\mu = \sum_{i=1}^{N} \frac{X_i \mu_i}{\sum_{j=1}^{N} X_j \phi_{ij}}, \qquad \lambda = \sum_{i=1}^{N} \frac{X_i \lambda_i}{\sum_{j=1}^{N} X_j \phi_{ij}} \tag{14}$$

with

$$\phi_{ij} = \frac{\left\{1 + \left(\frac{\mu_i}{\mu_j}\right)^{1/2} \left(\frac{R_i}{R_j}\right)^{1/4}\right\}^2}{\left\{8\left(1 + \frac{R_j}{R_i}\right)\right\}^{1/2}}. \tag{15}$$

The specific heat at constant pressure $c_{p,i}$, the dynamic viscosity μ_i and the thermal conductivity λ_i of each species i are defined as a function of the temperature through polynomial expansion up to the desired order of approximation. As in the present application pressure and temperature are nearly constant, all species thermofluidynamic properties are assumed constant.

In the hypothesis of perfect gas, the constitutive equation of state for the mixture reads $p = \rho R T$, where R is the mixture gas constant, $\sum_{i=1}^{\mathrm{N}} y_i R_i$. Herein the weak form of the EOS is considered and p is defined in a weak sense via an integral relation:

$$\int v p \, \mathrm{d}V = \int v \rho R T \, \mathrm{d}V,$$

where v is every function in the chosen functional space, belonging to the Hilbert space of square integrable functions. The pressure degrees of freedom p_j are computed in a straightforward way as

$$p_j(t) = \int v_j \rho R T dV, \tag{16}$$

due to the selected basis functions v, orthonormal in the physical reference frame.

2.1 Discontinuous Galerkin Discretization

The governing equations (1)–(7) can be written in compact form as

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{F}\left(\mathbf{u}, \nabla \mathbf{u}\right) + \mathbf{s}_t\left(\mathbf{u}, \nabla \mathbf{u}\right) = \mathbf{0}, \tag{17}$$

where $\mathbf{u} \in \mathbb{R}^m$ is the unknown solution vector of the m conservative variables, $\mathbf{F} \in \mathbb{R}^m \otimes \mathbb{R}^d$ is the sum of inviscid and viscous flux functions (with d indicating the number of dimensions) and $\mathbf{s}_t \in \mathbb{R}^m$ is the turbulent source term vector.

The discretization is defined over a computational mesh made of a set of non-overlapping elements $\mathcal{T}_h = \{K\}$. We thus consider the system (17) in its weak form over $\mathcal{T}_h$ and use a discontinuous Galerkin scheme based on the following space consisting of polynomial functions,

$$V_h \stackrel{\text{def}}{=} [\mathbb{P}_d^k(\mathcal{T}_h)]^m, \tag{18}$$

where

$$\mathbb{P}_d^k(\mathcal{T}_h) \stackrel{\text{def}}{=} \left\{ v_h \in L^2(\Omega) \,|\, v_{h|K} \in \mathbb{P}_d^k(K),\, \forall K \in \mathcal{T}_h \right\}$$

being $\mathbb{P}_d^k(K)$ the space of polynomials of degree at most k defined on the element K. The DG discretization requires to compute $\mathbf{u}_h \in V_h$, such that:

$$\begin{aligned}
&\sum_{K \in \mathcal{T}_h} \int_K v_i v_j \frac{\mathrm{d}U_j}{\mathrm{d}t}\, r\, \mathrm{d}\mathbf{x} - \sum_{K \in \mathcal{T}_h} \int_K \frac{\partial v_i}{\partial x_n} F_{j,n}\left(\mathbf{u}_h, \nabla_h \mathbf{u}_h + \mathbf{R}(\llbracket \mathbf{u}_h \rrbracket)\right)\, r\, \mathrm{d}\mathbf{x} \\
&\quad + \sum_{e \in \mathcal{F}} \int_e \llbracket v_i \rrbracket_n \widehat{F}_{j,n}\left(\mathbf{u}_h^{\pm}, (\nabla_h \mathbf{u}_h + \eta_e \mathbf{r}_e \llbracket \mathbf{u}_h \rrbracket)^{\pm}\right)\, r\, \mathrm{d}\sigma \\
&\quad + \sum_{K \in \mathcal{T}_h} \int_K v_i s_j\left(\mathbf{u}_h, \nabla_h \mathbf{u}_h + \mathbf{R}\left(\llbracket \mathbf{u}_h \rrbracket\right)\right)\, r\, \mathrm{d}\mathbf{x} = 0,
\end{aligned} \tag{19}$$

for each function $v_i \in V_h$, where $s_j = s_{t,j} + s_{ax,j}$ is the sum of turbulent and axisymmetric source terms. In Eq. (19) $\mathcal{F}$ defines the collection of all edges e and the functions $\mathbf{R}$ and $\mathbf{r}_e : [L^2(e)]^d \to [\mathbb{P}_d^k(\mathcal{T}_h)]^d$ are the global and local lifting operators

$$\int_\Omega \mathbf{r}_e\left(\mathbf{v}\right) \cdot \boldsymbol{\tau}_h \, \mathrm{d}\mathbf{x} = -\int_e \{\boldsymbol{\tau}\} \cdot \mathbf{v} \, \mathrm{d}\sigma, \qquad \forall \boldsymbol{\tau}_h \in [\mathbb{P}_d^k(\mathcal{T}_h)]^d, \tag{20}$$

$$\mathbf{R}\left(\mathbf{v}\right) = \sum_{e \in \mathcal{F}} \mathbf{r}_e\left(\mathbf{v}\right), \tag{21}$$

used by the BR2 scheme [1] to define a DG discretization of the viscous part of the governing equations. The jump $\llbracket \cdot \rrbracket$ and average $\{\cdot\}$ trace operators are defined as usual in the DG context to conveniently deal with discontinuities at elements interface as

$$\llbracket \mathbf{u}_h \rrbracket \stackrel{\text{def}}{=} \mathbf{u}_h^+ \mathbf{n}^+ + \mathbf{u}_h^- \mathbf{n}^-, \quad \{v_h\} \stackrel{\text{def}}{=} \frac{v_h^+ + v_h^-}{2}, \tag{22}$$

where the $\pm$ superscripts indicate the trace of the numerical solution over a generic edge computed from the two adjacent elements K^+ and K^-. The coefficient η_e is a stability parameter depending on the number of elements faces; see [26] for a detailed analysis of the BR2 method.

Suitable numerical flux functions ($\widehat{\mathbf{F}}_j$ in Eq. (19)) are used in place of the physical fluxes (convective and diffusive) to properly deal with the non-uniqueness of these vectors on the elements interfaces. The inviscid part of the flux is computed via "exact" Riemann solver [13] while the diffusive part as the average of the fluxes computed from neighboring elements.

The conservative formulation of the governing equations is used only for the purpose of fluxes computation, whilst the unknowns are given by the primitive variables vector

$$\mathbf{q} = [\rho, T, v_x, v_r, k, \widetilde{\omega}, y_1, y_2, ..., y_{\mathrm{N}-1}]^T.$$

The use of the primitive variables is computationally more efficient in the evaluation of the diffusive fluxes, where the gradient of $\mathbf{q}$ is required, since it is already available. Differently from monocomponent flow simulations, in multi-fluid flow computations we employ the density ρ in place of the pressure p. In most reactive gas dynamic environments the pressure is in fact slowly varying or nearly homogeneous, whereas strong density gradients exist near flame fronts and thermal boundary layers.

The new multicomponent DG solver here presented closely follows the implementation of the monocomponent DG MIGALE code [1, 2, 3, 4]. A fully implicit discretization is employed with Jacobian matrices analytically computed to account in exact manner for the dependence of residuals on the values of variables and gradients, also including the treatment of lifting operators and boundary conditions.

2.2 Time Integration

The DG space discretization of the governing equations (19) is here written in the form of the following non-linear ODE system

$$\mathbf{M}\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} + \mathbf{R}(\mathbf{U}) = \mathbf{0}, \tag{23}$$

where $\mathbf{U}$ is the vector of global unknown degrees of freedom, $\mathbf{M}$ the global block diagonal mass matrix and $\mathbf{R}$ the residuals vector. The time discretization of Eq. (23) is efficiently performed by means of a linearly implicit Rosenbrock-type Runge-Kutta schemes, following the implementation presented in [27], that can be written as:

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \sum_{j=1}^{s} m_j \mathbf{W}_j, \tag{24}$$

$$\left(\frac{\mathbf{I}}{\gamma\Delta t} + \widetilde{\mathbf{J}}\right)^n \mathbf{W}_i = -\widetilde{\mathbf{R}}\left(\mathbf{U}^n + \sum_{j=1}^{i-1} a_{ij}\mathbf{W}_j\right) + \sum_{j=1}^{i-1} \frac{c_{ij}}{\Delta t}\mathbf{W}_j, \quad i = 1, \ldots, s, \tag{25}$$

where

$$\mathbf{J} = \frac{\partial \mathbf{R}}{\partial \mathbf{U}}, \quad \widetilde{\mathbf{R}} = \mathbf{M}^{-1}\mathbf{R}, \quad \widetilde{\mathbf{J}} = \frac{\partial \widetilde{\mathbf{R}}}{\partial \mathbf{U}} = \mathbf{M}^{-1}\left(\mathbf{J} - \frac{\partial \mathbf{M}}{\partial \mathbf{U}}\widetilde{\mathbf{R}}\right), \tag{26}$$

and c_{ij}, a_{ij}, m_j, γ are the scheme coefficients. The Eq. (25) is then replaced by the following efficient implementation that avoids the computation of $\mathbf{M}^{-1}\widetilde{\mathbf{J}}$,

$$\left(\frac{\mathbf{M}}{\gamma\Delta t} + \mathbf{J} - \frac{\partial \mathbf{M}}{\partial \mathbf{U}}\widetilde{\mathbf{R}}\right)^n \mathbf{W}_i = -\mathbf{M}^n\left[\widetilde{\mathbf{R}}\left(\mathbf{W}^n + \sum_{j=1}^{i-1} a_{ij}\mathbf{W}_j\right) - \sum_{j=1}^{i-1} \frac{c_{ij}}{\Delta t}\mathbf{W}_j\right], \tag{27}$$

for $i = 1, \ldots, s$. For steady problems the time discretization is performed using the backward Euler scheme with the Jacobian matrix analytically computed to achieve quadratic convergence. For unsteady computations, accurate Runge-Kutta schemes up to fourth-order accuracy are adopted.

The linear system is finally solved using the restarted GMRES algorithm with 120 Krylov vectors, 240 iterations and a relative tolerance of 10^{-6}. The code is parallel and employs the METIS package to handle grid partitioning and the PETSc library [28] for linear algebra.

2.3 Shock-capturing technique

High-order DG methods, like any linear (*i.e.* independent of the local solution) high-order method, suffer from numerical oscillations in the vicinity of flow field discontinuities. In multicomponent flow computations, spurious fluctuations appear near material fronts already with a first-order approximate conservative scheme. Thus, to obtain a crisp representation of discontinuities and preserve the accuracy within smooth flow field regions, it is necessary to adopt *ad-hoc* numerical techniques, such as shock capturing schemes suited to control oscillations.

In this work we employ the shock-capturing technique presented in [16], originally inspired by [29] which is based on the explicit introduction of artificial viscosity into the governing equations. The following element-wise artificial dissipation contribution is thereby added to the system (19):

$$\int_K \varepsilon \left(\mathbf{u}_h^{\pm}, \mathbf{u}_h\right) \frac{\partial v_i}{\partial x_n} \frac{\partial u_j}{\partial x_n} \, \mathrm{d}\mathbf{x}. \tag{28}$$

The artificial viscosity coefficient ε depends on two residual-based quantities: i) the difference between the numerical and physical inviscid fluxes and ii) the divergence of the inner convective fluxes,

$$\varepsilon = C h_K^2 \sum_{n=1}^{3} \left(\frac{|S_n| + |D_n|}{\widetilde{q}_n^*} \right) F_s, \tag{29}$$

where C is a user-defined value, h_K the characteristic dimension of the element K defined in 2D as

$$h_K = \frac{1}{\sqrt{\dfrac{1}{(\Delta x)^2} + \dfrac{1}{(\Delta r)^2}}}, \tag{30}$$

where Δx and Δr are the dimensions of the quadrilateral enclosing K, scaled in such a way that their product matches the area of K.

The viscosity is selectively introduced based on the behavior of the variables $\widetilde{\mathbf{q}} = (\rho, T, y_M)$, *i.e.* the density, the temperature and the mass fraction y_M of the species showing the local greatest variation, chosen to tune the dissipation. Notice that in Eq. (29), the vector $\widetilde{\mathbf{q}}^* = (\rho, T, 1)$ is used instead of $\widetilde{\mathbf{q}}$ to avoid break-up caused by too small values of y_M. The S_n terms account for changes occurring at elements interface and based on the difference between the "numerical" and inner convective flux, herein denominated s_i. The D_n terms account for any steep gradient in the flow field occurring within the elements and are based on the divergence of the convective inner flux, denoted by d_i,

$$S_n\left(\mathbf{u}_h^{\pm}, \mathbf{u}_h\right) = \sum_{i=1}^{M} \frac{\partial \widetilde{q}_{h_n}(\mathbf{u}_h)}{\partial \mathbf{u}_{h_i}} s_i(\mathbf{u}_h^{\pm}), \qquad D_n\left(\mathbf{u}_h\right) = \sum_{i=1}^{M} \frac{\partial \widetilde{q}_{h_n}(\mathbf{u}_h)}{\partial \mathbf{u}_{h_i}} d_i(\mathbf{u}_h). \tag{31}$$

The artificial viscosity is then multiplied by a factor F_s based on the maximum absolute value of the temperature, density and mass fraction scaled gradients, depending on polynomial order k and mesh coarseness parameter h_K, given by

$$F_s = \left(|\mathbf{d\tilde{q}_j}| \, \frac{h_K}{k} \right)^k , \tag{32}$$

with

$$|\mathbf{d\widetilde{q}_j}| = \max_j \left(\frac{|\nabla_h \widetilde{q}_j|}{\widetilde{q}_j^*} \right), \qquad j = 1, 2, 3.$$

The user-defined coefficient C has been determined through numerical experiments and in the present work it typically varies between 0.1 and 1.

2.4 Limiting on mixture properties

During computations, out of bounds mass fraction values may appear in the solution flow field and this may lead to the attainment of unphysical gas properties. Different strategies have been considered by various authors to deal with oscillations arising in non monotonic approximation scheme, *e.g.* [30]. In this work, we do not constraint species degrees of freedom and thus negative or greater than one mass fraction values may appear. To keep this potential problem under control, we employed a simple limiting procedure that locally uses modified mass fraction values, satisfying Eq. (8), for the sole purpose of gas properties computation. The technique is presented in [16]; it first corrects (if necessary) each of the N species mass fraction to a minimum value and then divides each one by the sum of the new N mass fraction values.

3 He-CO_2 JET IN A CO-FLOW STREAM OF AIR

In this section third-order accurate ($\mathbb{P}^2$) computations are discussed and compared with available experimental data [17].

Experiments were performed at ambient conditions in a configuration typical of non-premixed coaxial burners, consisting of an inner long pipe ($x/D = 37.5$) delivering a mixture of helium (11.8% on a mass basis) and carbon dioxide, surrounded by much a shorter air annulus ($x/D_m = 3.9$). The closed chamber diameter is 5.6 times the annulus outer diameter, therefore confinement effects should not be relevant. The post thickness is 3.5 mm. Other geometrical parameters of the apparatus are listed in Table 1.

	jet	air inner	air outer	chamber
Diameter	8	15	36	200
Length	300	80	80	300

Table 1: Test rig dimensions [mm]

The considered test cases are summarized in Table 2, where the Reynolds numbers, based on bulk velocity, (hydraulic) diameter and kinematic viscosity, are reported alongside with density (D), velocity (V) and momentum flux (M) inner-to-outer ratios.

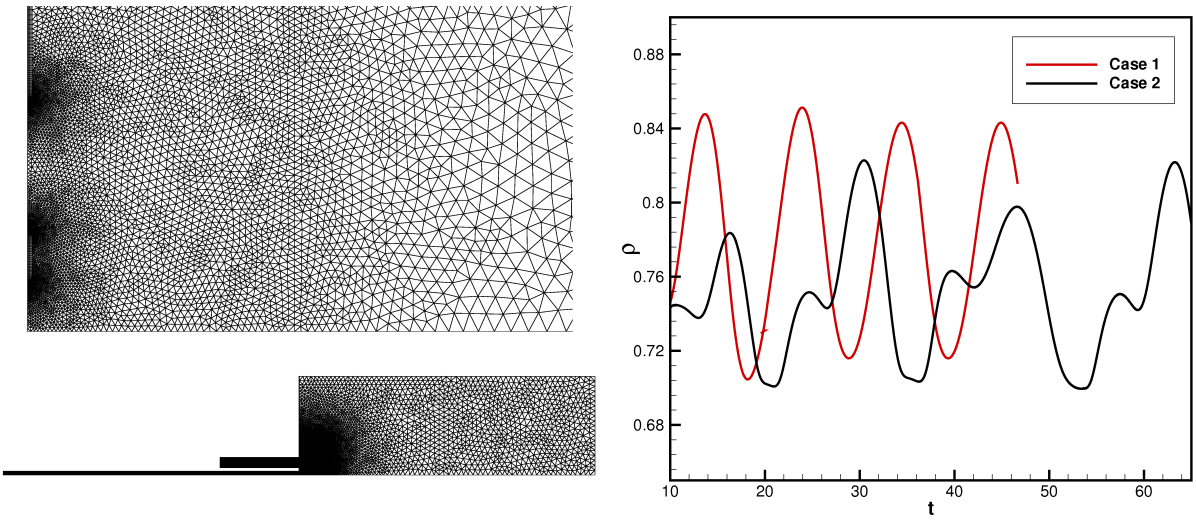
Notice that in [17], due to a typing error, what is referred to as bulk velocity is actually the centerline velocity at the exit. The flow regime is incompressible, being the jet freestream Mach number $Ma_{jet} \simeq 0.025$.

An unstructured grid consisting of 42722 elements of mixed triangular and quadrilateral type is employed to discretize half geometry. Figure 1(a) shows the computational domain (bottom

Case	Re_{jet}	Re_{air}	V	D	M
1	3140	2350	4.82	0.695	16.13
2	3140	9480	1.19	0.695	1.01

Table 2: Test case Reynolds number, velocity, density and momentum flux ratios

part of figure) and a close up of the mixing region (top part) near the merging of jet and co-flow. Note that the whole length of the tube has been discretized since it is not long enough to obtain a fully developed pipe flow.



(a) grid (b) time history

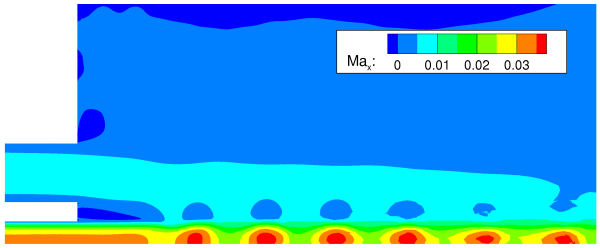
Figure 1: Computational mesh (left) and time histories of Case 1 – red and Case 2 –black (right)

The flow field is unsteady in both test cases, so computations were run in time-accurate mode until a periodic solution was finally set. To check the attainment of periodicity, fluidynamic quantities were continuously sampled at three points in the initial region of jets interaction ($x/D = 1$, $0.5 \leq r/D \leq 0.95$); an example of density time histories is shown in Figure 1(b).

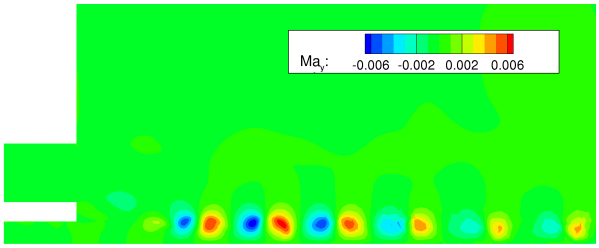
In Case 1 both the inner and outer channel boundary layer Reynolds numbers are in the transitional range; even if they are well above the threshold for the development of turbulence in the subsequent jet free shear layer, flow field features may heavily depend on displacement and momentum thicknesses at the pipe exit. Therefore, for test Case 1 both laminar and turbulent computations were performed.

Figure 2 displays the Mach numbers and species mass fractions of the laminar solution at a generic instant of time, while Figure 3 shows the time-averaged periodic solution. At the jet exit, the shear layer (Kelvin-Helmholtz) instability generates a continuous detachment of vortices moving toward the axis, all counter-clockwise rotating, *i.e.* with positive vorticity. When they encounter their mirror image at the axis of symmetry, a local reinforcement is clearly visible prior to their slow downstream decay under the influence of viscosity. One single dominant frequency is present in the time spectrum, corresponding to a Strouhal number

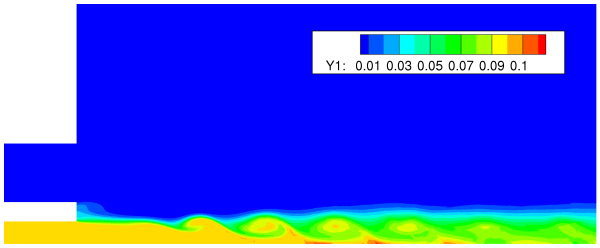
$$St = \frac{\omega D}{V_{bulk}} \simeq 0.49.$$



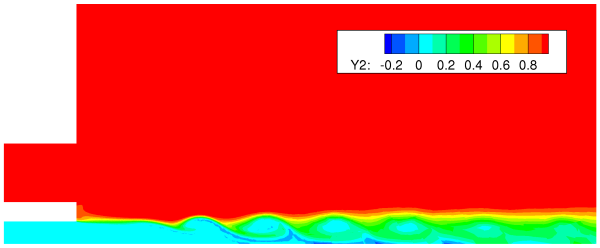
(a) Ma_x



(b) Ma_y

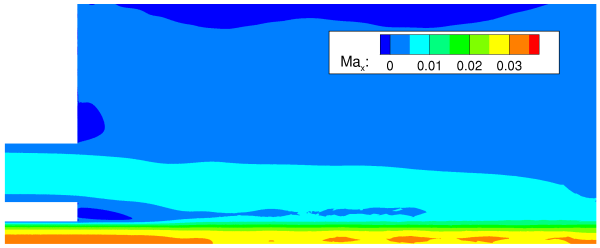


(c) y_{He}

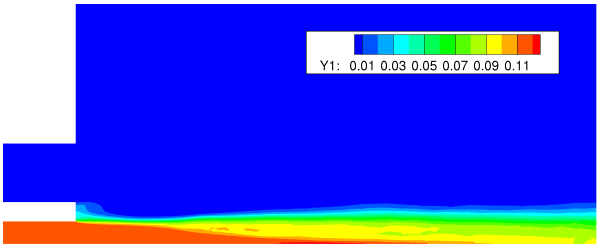


(d) y_{air}

Figure 2: Case 1, laminar: instantaneous contours



(a) Ma_x



(b) y_{He}

Figure 3: Case 1, laminar: time-averaged contours

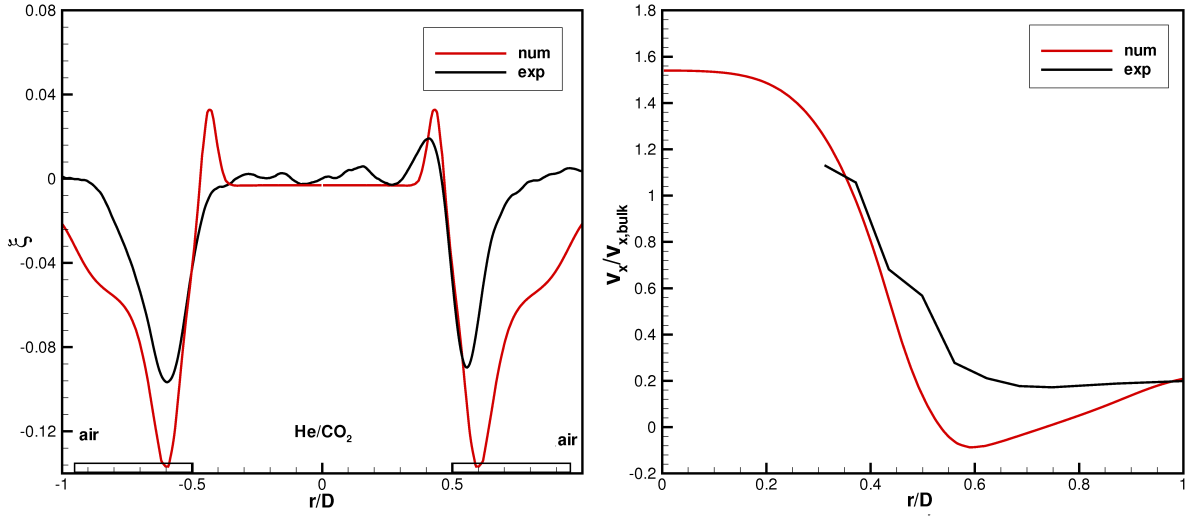
In case it were a fully developed (Poiseuille) pipe flow, this would correspond to a Strouhal number based on momentum thickness of 0.021, at the upper edge of the classical laminar range $(0.01 - 0.02)$.

In order to evaluate the mechanism of preferential diffusion, we compute a local diffusion parameter ξ defined as follows,

$$\xi = \frac{X_{\mathrm{CO_2}}}{X^0_{\mathrm{CO_2}}} - \frac{X_{\mathrm{He}}}{X^0_{\mathrm{He}}},$$

where X^0_i is the initial mole fraction of the species i. Positive values of this parameter can be interpreted as a local excess of carbon dioxide as compared with helium; the opposite is true for negative values.

Figure 4 shows the time-averaged numerical and experimental distributions of ξ and axial non-dimensional velocity in the coaxial jets near field. As one can see, the laminar solution overestimates maxima and minima in both ξ and velocity profiles; in particular, trace of the recirculation bubble due to the finite post thickness is still present at $x/D = 0.75$, causing a zone of local backflow. We then turned to turbulent computations, using Wilcox's 2008 (high-Re) turbulence model.



(a) ξ (b) velocity

Figure 4: Case 1, laminar: time-averaged ξ coefficient and axial velocity at $x/D = 0.75$

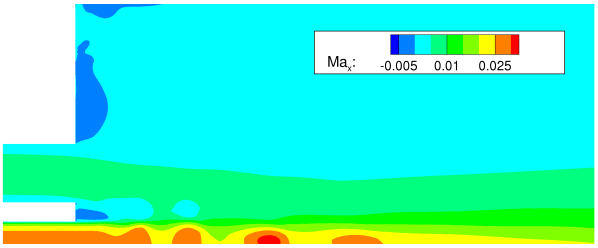
Figure 5 displays the Mach numbers, species mass fractions, vorticity and turbulence intensity of the turbulent solution at a generic instant of time, whilst Figure 6 shows the time-averaged periodic solution. The vorticity is scaled with the pipe radius and the air ambient sound speed; in the turbulence intensity the turbulent kinetic energy is scaled with the square inlet air velocity. Unsteady coherent structures are still well evident, even if they are less pronounced and more smeared, owing to the diffusive action of turbulence. In the reported figures it can be noticed the different jet spreading rate, greater than in the laminar case, as should be expected. The axial distance travelled by perturbations before reaching the axis of symmetry is therefore shorter.

Figure 7 is, for the turbulent case, the companion of Figure 4 . The results are now in a much better agreement with experimental data, confirming how the turbulence plays a key role in the mass and momentum mixing process. Maybe different results could have been obtained turning on turbulence in the simulation starting from the jet discharge axial point, but this is left as a future exercise. Test Case 2 was therefore only run under turbulent assumption.

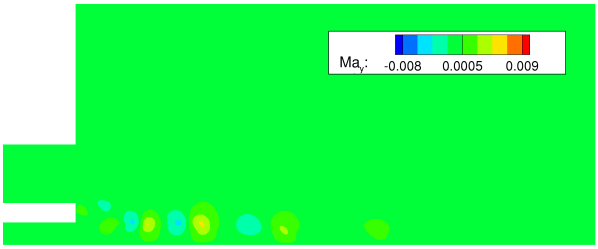
Figure 8 shows a snapshot of the Mach numbers, species mass fractions, vorticity and turbulence intensity. This pictures are not at all alike their analogous obtained at greater velocity ratio. Vortices are leaving the air co-flow at the outer diameter as well and, moreover, the region enclosed between the two coaxial jets is now of the "wake-like" type with vortices of alternating sign visible close to the exit.

The coherent structures generated at the interface persist much longer downstream and the eventual setting of a self-similar solution is expected to occur farther away. The most important parameter governing what kind of configuration is likely to appear in the near field is indeed the momentum flux ratio M and it is well established that a value of M around unity leads to a deficit velocity profile, and possibly to an unsteady vortex street, typical of wake flows.

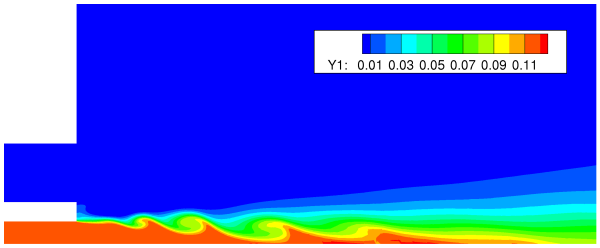
Furthermore, a much more complex flow field is apparent as a natural outcome of the jets strong mutual interaction. This is also verified in Figure 9, where relevant steady spatial oscillations are noticeable. It is sensible to infer that removing the symmetry constraint and simulating the whole coaxial jet may result in a smoother average flow field, since perturbations are not forced to develop themselves simultaneously.



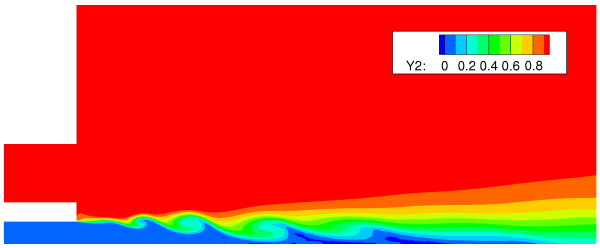
(a) Ma_x



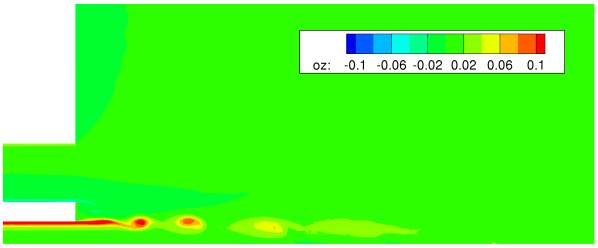
(b) Ma_y



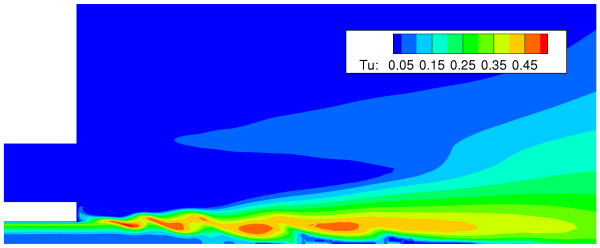
(c) y_{He}



(d) y_{air}

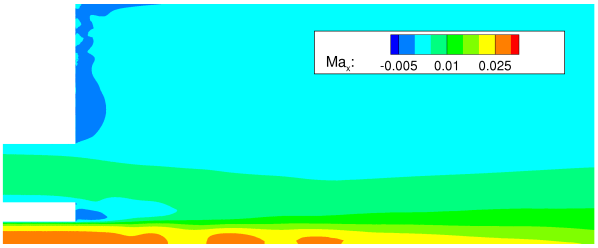


(e) vorticity

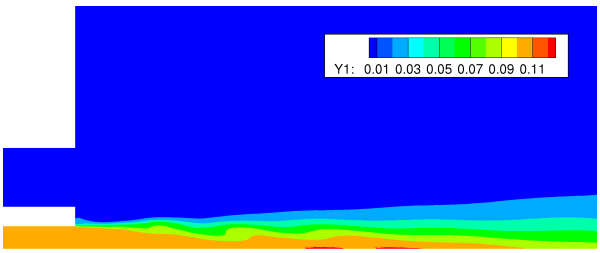


(f) turbulence intensity

Figure 5: Case 1, turbulent: instantaneous contours

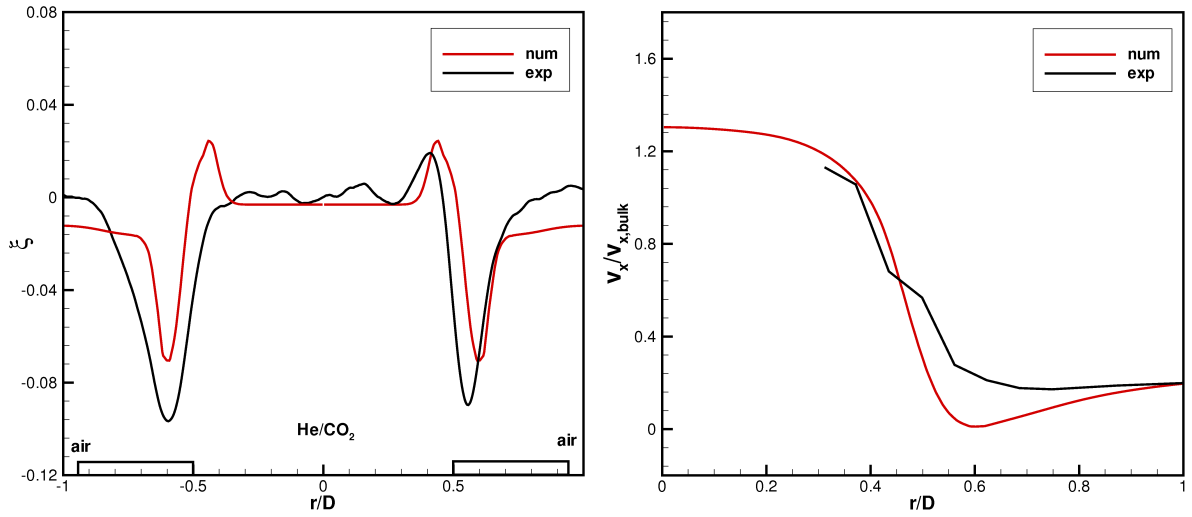


(a) Ma_x



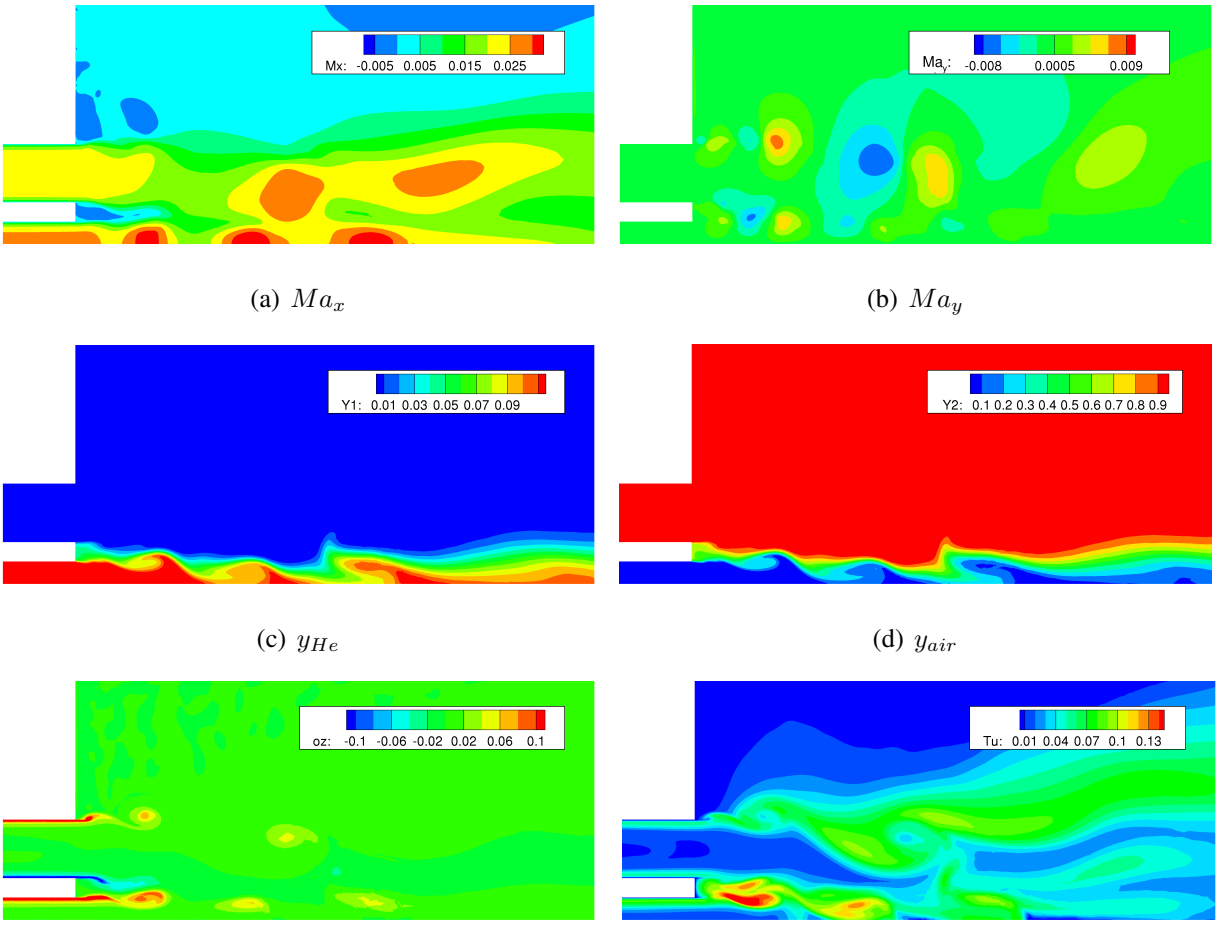
(b) y_{He}

Figure 6: Case 1, turbulent: time-averaged contours



(a) ξ (b) velocity

Figure 7: Case 1, turbulent: time-averaged ξ coefficient and axial velocity at $x/D = 0.75$

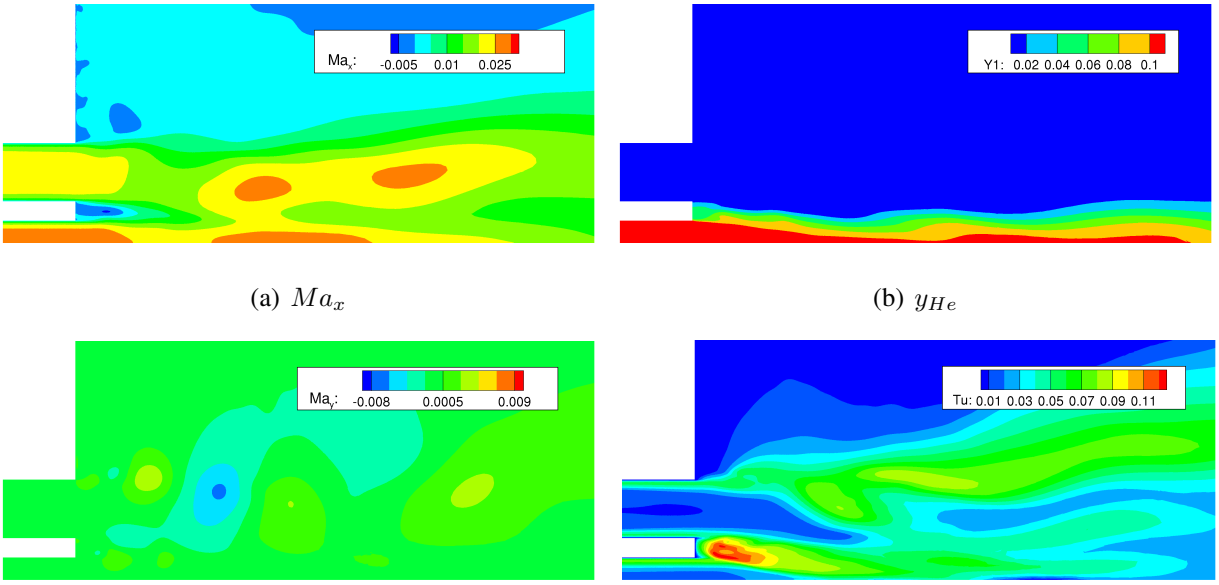


(a) Ma_x (b) Ma_y

(c) y_{He} (d) y_{air}

(e) vorticity (f) turbulence intensity

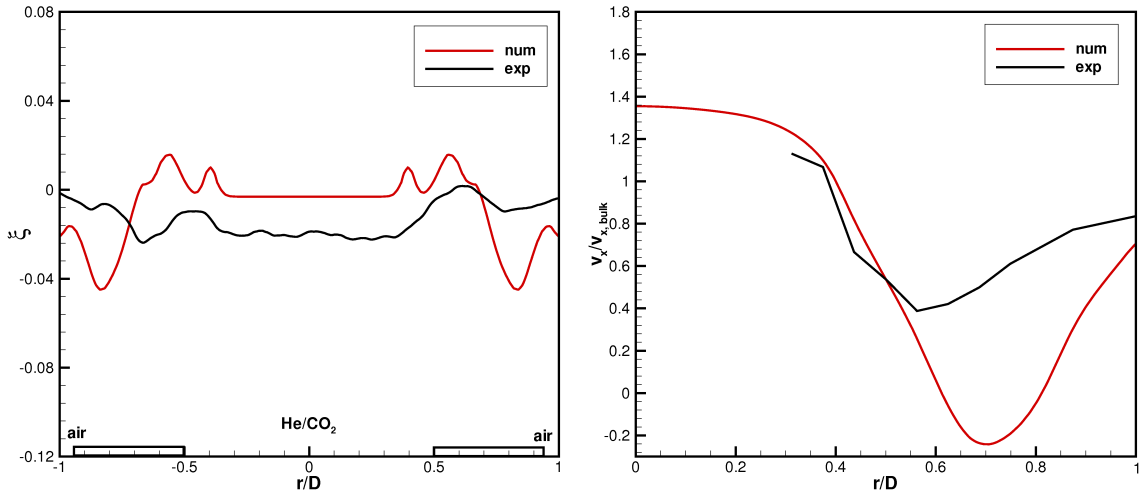
Figure 8: Case 2: instantaneous contours



(a) Ma_x (b) y_{He}

(c) Ma_y (d) turbulence intensity

Figure 9: Case 2: time averaged contours



(a) ξ (b) velocity

Figure 10: Case 2: time-averaged ξ coefficient and axial velocity at $x/D = 0.75$

Time-averaged ξ and non-dimensional v_x graphs are reported in Figure 10. The computed radial profile of helium and carbon dioxide mass fraction is satisfactory, keeping in mind the uncertainties inherent in the measurements (note the lack of symmetry and the nearly absence of positive values in the experimental data) and the fact that the local ξ rms is of the same order of its mean value; proof of this last statement can be found in Figure 11. On the other hand, there is an evident mismatch in the radial velocity profile. It seems due to the overestimation of the post recirculation zone in the numerical solution, similarly to what already manifest, even if to a lesser extent, in the Case 1 turbulent computation. The change of flow configuration from the "jet-like" to the "wake-like" type enlarged the rear separation bubble and shifted radially outward the point of minimum average velocity at $x/D = 0.75$.

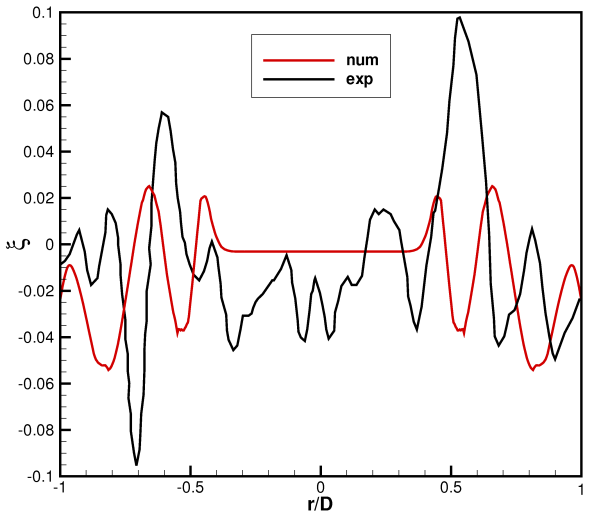


Figure 11: Case 2: instantaneous ξ coefficient at $x/D = 0.75$

4 CONCLUSIONS

A high-order accurate DG scheme suited to multicomponent flow computation has been presented and applied to investigate the complex mixing and diffusion mechanism occurring in coaxial jets. Oscillations typically encountered by numerical schemes applied to the conservative form of the multicomponent gas equations are herein kept under control by means of: i) the use of the weak form of the equation of state that allows to compute a more regular pressure field, ii) the addition of a shock-capturing term based upon local inviscid residuals and interface flux jumps, and iii) a limiting technique on mass fraction values, in order to avoid mixture properties out of admissible bounds.

The DG solver has been used to compute the flow of a mixture of helium and carbon dioxide discharging in a surrounding co-flow of air at two different Reynolds numbers and inner-to-outer momentum flux ratios. Laminar and turbulent unsteady computations have been carried out to assess the capability of the code and to investigate the jet and co-flow mixing process taking place. The results obtained indicate in all cases the presence of a strongly unsteady flow field with the development of noteworthy coherent structures. Moreover, even at Re as small as about 10^3, the assumption of turbulent flow is necessary to obtain a reasonable agreement with measurements.

The preferential diffusion of helium with respect to carbon dioxide is evident (and well captured) only at the smaller Re and higher velocity ratio, when also the computed time-averaged velocity field in the near region is close to the experimental one. In this case the mixing process resembles the one, well known, experienced by a jet emanating in a quiescent ambient, whereas a complete different picture exists when jet and co-flow possess similar velocity (more correctly, similar momentum flux). In this latter case one observes a “wake-like” mixing mode, characterized by a more complex interaction and a much long-lasting presence of vortices.

Finally, it is worth to mention that in the former case (jet) the unsteadiness is focused on a single time scale and the unsteady vorticity field is positive definite, while in the latter case (wake) more frequencies are present altogether with an unsteady vorticity field showing both signs. These and other interesting topics are currently under investigation and will be the subject of a forthcoming paper.

REFERENCES

[1] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997.

[2] F. Bassi, A. Crivellini, S. Rebay, and M. Savini. Discontinuous Galerkin solution of the Reynolds-averaged Navier-Stokes and k-ω turbulence model equations. *Comput. & Fluids*, 34:507–540, 2005.

[3] F. Bassi, A. Colombo, N. Franchina, A. Ghidoni, and S. Rebay. Robust and efficient implementation of very high-order discontinuous Galerkin methods in CFD. *Notes on Fluid Mechanics*, 113:287–299, 2010.

[4] F. Bassi, L. Botti, A. Colombo, A. Crivellini, C. De Bartolo, N. Franchina, A. Ghidoni, and S. Rebay. Time Integration in the Discontinuous Galerkin code MIGALE - Steady Problems. In Norbert Kroll, Charles Hirsch, Francesco Bassi, Craig Johnston, and Koen Hillewaert, editors, *IDIHOM: Industrialization of High-Order Methods - A Top-Down Approach*, volume 128 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 179–204. Springer International Publishing, 2015.

[5] G. Billet, J. Ryan, and M. Borrel. A Runge Kutta Discontinuous Galerkin approach to solve reactive flows on conforming hybrid grids: the parabolic and source operators. *Computers & Fluids*, 95:98–115, 2014.

[6] Y. Lv and M. Ihme. Discontinuous Galerkin method for multicomponent chemically reacting flows and combustion. *J. Comput. Phys.*, 270:105–137, 2014.

[7] F. H. Champagne and I.J. Wygnanski. An experimental investigation of coaxial turbulent jets. *J. Heat and mass transfer*, 14:1445–1464, 1971.

[8] A. Antonia and R. W. Bilger. An experimental investigation of an axisymmetric jet in a co-flowing air stream. *J. Fluid Mech.*, 61:805–822, 1973.

[9] N.W.M. Ko and A.S.H. Kwan. The initial region of coaxial jets. *J. Fluid Mech.*, 73:305–332, 1976.

[10] G. Buresti, P. Petagna, and A. Talamelli. Experimental investigation on the turbulent near-field of coaxial jets. *Exp. Thermal and Fluid Science*, 17:18–36, 1998.

[11] S. A. Schumaker. *An experimental investigation of reacting and nonreacting coaxial jet mixing in a laboratory rocket engine*. PhD thesis, University of Michigan, 2009.

[12] F. Cozzi and A. Coghe. Behavior of hydrogen-enriched non-premixed swirled natural gas flames. *Int. Journal of Hydrogen Energy*, 31:669–677, 2006.

[13] J. J. Gottlieb and C. P. T. Groth. Assessment of Riemann solvers for unsteady one-dimensional inviscid flows of perfect gases. *J. Comput. Phys.*, 78:437–458, 1988.

[14] R. Abgrall. How to prevent pressure oscillations in multicomponent flow calculations: a quasi conservative approach. *Rapporte de recherche, INRIA*, 2372, Octobre 1994.

[15] E. Johnsen and F. Ham. Preventing numerical errors generated by interface-capturing schemes in compressible multi-material flows. *J. Comput. Phys.*, pages 5705–5717, 2012.

[16] N. Franchina, M. Savini, and F. Bassi. Multicomponent gas flow computations by a discontinuous Galerkin scheme using L^2-projection of perfect gas EOS. *J. Comput. Phys.*, 2015, 10.1016/j.jcp.2016.03.059.

[17] F. Cozzi, A. Olivani, L. Caratti, and A. Coghe. Investigation of fine structure of turbulent and molecular diffusion in coaxial jets of He/CO_2 in air by LDA and Rayleigh scattering. *Exp. Thermal and Fluid Science*, 34:316–322, 2010.

[18] F. Bassi, F. Cecchi, N. Franchina, S. Rebay, and M. Savini. High-order discontinuous Galerkin computation of axisymmetric transonic flows in safety valves. *Comput. & Fluids*, 49:203–213, 2011.

[19] J. Warnatz, U. Maas, and R.W. Robert. *Combustion: Physical and Chemical Fundamentals, Modeling and Simulation*. Springer, 2006.

[20] N. B. Vargaftik and Y.K. Vinogradov. *Handbook of physical properties of liquid and gases: pure substances and mixtures*. Begel House, 1996.

[21] D. C. Wilcox. *Turbulence Modelling for CFD*. DCW industries Inc., La Cañada, CA 91011, USA, 2006.

[22] D. C. Wilcox. Formulation of the k-ω turbulence model revisited. *AIAA Journal*, 44:2823–2838, 2008.

[23] C.R. Wilke. A viscosity equation for gas mixture. *J. Chem. Phys.*, 18:517–519, 1950.

[24] E.A. Mason and S.C. Saxena. Physics of fluids. *J. Fluids Eng.*, pages 361–369, 1958.

[25] B. R. Byron, W.E. Stewart, and E. N. Lightfoot. *Transport phenomena*. John Wiley & Sons, Inc., 2002.

[26] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[27] F. Bassi, L. Botti, A. Colombo, A. Crivellini, A. Ghidoni, and F. Massa. Linearly implicit Rosenbrock-type Runge–Kutta schemes applied to the Discontinuous Galerkin solution of compressible and incompressible unsteady flows. *Comput. & Fluids*, 118:305–320, 2015.

[28] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. Curfman McInnes, K. Rupp, B.F. Smith, S. Zampini, and H. Zhang. PETSc Web page. `http://www.mcs.anl.gov/petsc`, 2014.

[29] F. Bassi, L. Botti, A. Colombo, A. Crivellini, N. Franchina, A. Ghidoni, and S. Rebay. Very high-order discontinuous Galerkin computations of transonic turbulent flows on aeronautical configurations. *Notes on Fluid Mechanics*, 113:25–38, 2010.

[30] L. Krivodonova. Limiters for high-order discontinuous Galerkin methods. *J. Comput. Phys.*, 226:879–896, 2007.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DISCONTINUOUS GALERKIN SOLUTION OF THE REYNOLDS-AVERAGED NAVIER–STOKES AND K_L-K_T-$\tilde{\omega}$ TRANSITION MODEL EQUATIONS

A. Bassi[1], **A. Colombo**[1], **A. Ghidoni**[2], **M. Lorini**[2], **and G. Noventa**[2]

[1] Dipartimento di Ingegneria Industriale, Università di Bergamo
Viale Marconi 5, 24044 Dalmine (BG), Italy
e-mail: {francesco.bassi,alessandro.colombo}@unibg.it

[2] Dipartimento di Ingegneria Meccanica e Inustriale, Università di Brescia
via Branze 38, 25123 Brescia, Italy
e-mail: {antonio.ghidoni,m.lorini006,g.noventa002}@unibs.it

Keywords: RANS Equations, DG Method, Transition Modeling, local model, laminar kinetic energy.

Abstract. *Transition modelling represents a key ingredient to improve the performances of modern turbomachinery, affecting losses and the heat-transfer phenomenon. In this paper the transition model proposed by Walters [3] was considered. It is based on the k-ω formulation with the addition of a third transport equation (k_L equation), which allows predicting the magnitude of low-frequency velocity fluctuations in the pre-transitional boundary layer. The model was implemented into a parallel high-order accurate Discontinuous Galerkin code, named MIGALE, which allows to solve the Reynolds averaged Navier-Stokes (RANS) equations coupled with the k-$\widetilde{\omega}$ ($\tilde{\omega} = \log(\omega)$) turbulence model. The model was applied to two test cases, namely the flat plate (T3A and T3B configuration) and the T106A turbine cascade. Results obtained with and without transition model were compared with experimental data.*

1 INTRODUCTION

In the last decade, Computational Fluid Dynamic (CFD) has become a tool commonly adopted for industrial analysis and design. This wide spread can be partially explained by the development of reliable and robust turbulence models. However, these turbulence models are very well suited for high Reynolds number flows, whereas for low Reynolds numbers flows, where a large part of the boundary layer is laminar or transitional, they can provide wrong results. Therefore, the transition modeling represents a key ingredient to improve the prediction capabilities of standard RANS solvers. Different methods have been proposed in the finite volume context to predict the laminar-turbulent transition [1, 2, 3], but, according to the literature, there was only an attempt to couple a non-local transition model [17] with a high-order method. In this paper the local transition model proposed by Walters [3] was considered. The model is based on the k-ω formulation with the addition of a third transport equation (k_L equation), which allows predicting the magnitude of low-frequency velocity fluctuations in the pre-transitional boundary layer. The closure of the model is based on a phenomenological (*i.e.* physics-based) rather than a purely empirical approach. The model was implemented into a parallel high-order accurate Discontinuous Galerkin (DG) code, named MIGALE [4, 5], which allows solving the RANS equations coupled with the k-$\widetilde{\omega}$ turbulence model. The objective of this work is to assess the capability of high-order DG methods in accurately computing transitional flows on two benchmark testcases, i.e. the transitional flow around a flat plate (T3A and T3B configurations) and through the T106A turbine cascade, comparing high-order results with available numerical and experimental data.

2 GOVERNING EQUATIONS

The complete set of RANS and k_L-k_T-$\widetilde{\omega}$ model equations can be written as

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j) = 0 \ , \tag{1}$$

$$\frac{\partial}{\partial t}\left(\rho u_i\right) + \frac{\partial}{\partial x_j}(\rho u_j u_i) = -\frac{\partial p}{\partial x_i} + \frac{\partial \widehat{\tau}_{ji}}{\partial x_j} \ , \tag{2}$$

$$\frac{\partial}{\partial t}(\rho E) + \frac{\partial}{\partial x_j}(\rho u_j H) = \frac{\partial}{\partial x_j}\left[u_i \widehat{\tau}_{ij} - \widehat{q}_j\right] - \rho(P_{k_T} - \overline{k}_T \widetilde{\omega}) \ , \tag{3}$$

$$\begin{aligned}\frac{\partial}{\partial t}(\rho k_T) + \frac{\partial}{\partial x_j}(\rho u_j k_T) =& \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\rho \alpha_T}{\sigma_k}\right)\frac{\partial k_T}{\partial x_j}\right] \\ &+ \rho\left(P_{k_T} + R_{BP} + R_{NAT} - \overline{k}_T\widetilde{\omega} - D_T\right) \ ,\end{aligned} \tag{4}$$

$$\frac{\partial}{\partial t}(\rho k_L) + \frac{\partial}{\partial x_j}(\rho u_j k_L) = \frac{\partial}{\partial x_j}\left[\mu \frac{\partial k_L}{\partial x_j}\right] + \rho\left(P_{k_L} - R_{BP} - R_{NAT} - D_L\right) \ , \tag{5}$$

$$\begin{aligned}\frac{\partial}{\partial t}(\rho\widetilde{\omega}) + \frac{\partial}{\partial x_j}(\rho u_j \widetilde{\omega}) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\rho\alpha_T}{\sigma_\omega}\right)\frac{\partial \widetilde{\omega}}{\partial x_j}\right] + \rho\left[\left(\frac{C_{\omega R}}{f_W} - 1\right)\frac{\widetilde{\omega}}{\overline{k}_T}\left(R_{BP} + R_{NAT}\right)\right. \\ \left. + C_{\omega 1}\frac{\widetilde{\omega}}{\overline{k}_T}P_{k_T} - f_2\widetilde{\omega}^2 + P_{BL}\right] + \left(\mu + \frac{\rho\alpha_T}{\sigma_\omega}\right)\frac{\partial\widetilde{\omega}}{\partial x_k}\frac{\partial\widetilde{\omega}}{\partial x_k} \ ,\end{aligned} \tag{6}$$

where u_i is the absolute velocity.

The total energy, the total enthalpy, the pressure, the turbulent and total stress tensors, the heat flux vector, the eddy viscosity and the limited value of laminar and turbulent kinetic energy are given by

$$E = \hat{e} + u_k u_k/2, \tag{7}$$
$$H = h + u_k u_k/2, \tag{8}$$
$$p = (\gamma - 1)\rho \left(E - u_k u_k/2\right), \tag{9}$$
$$\tau_{ij} = 2\mu_t \left[S_{ij} - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right] - \frac{2}{3}\rho \overline{k}_T \delta_{ij}, \tag{10}$$
$$\widehat{\tau}_{ij} = 2\mu \left[S_{ij} - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right] + \tau_{ij}, \tag{11}$$
$$\widehat{q}_j = -\left(\frac{\mu}{\mathrm{Pr}} + \rho\alpha_\theta\right)\frac{\partial h}{\partial x_j}, \tag{12}$$
$$\mu_t = \rho\left(\nu_{T,s} + \nu_{T,l}\right), \tag{13}$$
$$\overline{k}_T = \max\left(0, k_T\right), \quad \overline{k}_L = \max\left(0, k_L\right), \tag{14}$$

where $\hat{e}$ is the internal energy, h the enthalpy, γ the ratio of gas specific heats, Pr the molecular Prandtl number and

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

is the mean strain-rate tensor.

The turbulence model was implemented using $\widetilde{\omega} = \log(\omega)$ instead of ω, limiting the values of k_T and k_L. At solid walls the homogeneous Neumann condition for the specific dissipation rate $\partial\widetilde{\omega}/\partial n = 0$ is prescribed.

Furthermore, due to the no-slip condition, the velocity is set to zero. Since the velocity is equal to zero, the kinetic energy at the wall must also reduce to zero. Therefore, turbulent and laminar kinetic energy are set to zero at the wall.

2.1 TRANSITION MODEL

The transition model considered in this work (see Eqs. 4, 5 and 6) is based on the model proposed by Walters [3]. This model includes several damping functions for the different physical phenomena observed (such as shear sheltering and wall reflection), which require empirically determined constants. However, the amount of empiricism necessary is still significantly less if compared to correlation-based intermittency models.

The various terms in the model equations represent advection, production, destruction and diffusion. The generation of turbulence due to small-scale turbulent fluctuations and the generation of laminar kinetic energy due to large-scale turbulent fluctuations are modelled as

$$P_{k_T} = \nu_{T,s} S^2 \,, \tag{15}$$
$$P_{k_L} = \nu_{T,l} S^2 \,, \tag{16}$$

where $S = \sqrt{2S_{ij}S_{ij}}$ is the mean strain-rate magnitude, $\nu_{T,s}$ and $\nu_{T,l}$ are the small-scale and large-scale components of the eddy-viscosity, respectively.

The small-scale eddy-viscosity is defined as

$$\nu_{T,s} = f_\nu f_{INT} C_\mu \sqrt{k_{T,s}} \lambda_{eff} \ , \tag{17}$$

where $k_{T,s}$ is the effective small-scale turbulence

$$k_{T,s} = f_{SS} f_W k_T \ . \tag{18}$$

The kinematic wall effect is included through an effective wall-limited turbulence length scale

$$\lambda_{eff} = \min\left(C_\lambda d, \lambda_T\right),$$

and damping function

$$f_W = \left(\frac{\lambda_{eff}}{\lambda_T}\right)^{\frac{2}{3}},$$

where

$$\lambda_T = \frac{\sqrt{k_T}}{\omega},$$

and d denotes the wall distance.

The viscous wall effect is incorporated through a viscous damping function

$$f_\nu = 1 - \exp\left(-\frac{\sqrt{Re_T}}{A_\nu}\right),$$

where

$$Re_T = \frac{f_W^2 k_T}{\nu\omega}$$

is the effective turbulence Reynolds number.

The shear-sheltering effect is included through the following damping function:

$$f_{SS} = \exp\left[-\left(\frac{C_{SS}\nu\Omega}{k_T}\right)^2\right] \ , \tag{19}$$

where $\Omega = \sqrt{2\Omega_{ij}\Omega_{ij}}$ is the vorticity magnitude.

In order to satisfy the realizability constraint suggested by Shih *et al.* [6], the turbulent viscosity coefficient, C_μ, takes the form

$$C_\mu = \frac{1}{A_0 + A_S\left(\frac{S}{\omega}\right)} \ . \tag{20}$$

Intermittency effects on the production of turbulence are included through the intermittency damping function

$$f_{INT} = \min\left(\frac{k_T}{C_{INT} k_{TOT}}\right) \ , \tag{21}$$

where $k_{TOT} = k_T + k_L$ is the total fluctuation kinetic energy.

The production of laminar kinetic energy, k_L, is assumed to be given by large-scale near wall turbulence, based on the correlation of pre-transitional fluctuation growth with free-stream

low-frequency wall-normal turbulent fluctuations [7]. The large-scale turbulence contribution is defined as

$$k_{T,l} = k_T - k_{T,s} \ , \tag{22}$$

whereas the small-scale contribution is defined by Eq. 18. The production term is then given by Eq. 15, where

$$\nu_{T,l} = \min\left\{ f_{\tau l} C_{l1} \left(\frac{\Omega \lambda_{eff}^2}{\nu} \right) \sqrt{k_{T,l}} \lambda_{eff} + \beta_{TS} C_{l2} Re_\Omega d^2 \Omega, \frac{k_L + k_{T,l}}{2S} \right\}, \tag{23}$$

where

$$\begin{aligned} Re_\Omega &= \frac{d^2 \Omega}{\nu} \ , \\ \beta_{TS} &= 1 - \exp\left(\frac{-\max\left(Re_\Omega - C_{TS,crit}, 0\right)^2}{A_{TS}} \right) \ , \\ f_{\tau,l} &= 1 - \exp\left(-C_{\tau,l} \frac{k_{T,l}}{\lambda_{eff}^2 \Omega^2} \right) \ . \end{aligned}$$

The limit is applied to ensure satisfaction of the realizability constraint for the total Reynolds stress contribution. The production term is comprised of two parts: the first addresses the development of Klebanoff modes and the second addresses self-excited (*i.e.* natural) modes.

The dissipation is divided into an isotropic ($k_T\omega$) and an anisotropic (D_T and D_L) part, as in the low-Reynolds Launder-Sharma $k-\epsilon$ model [8], with

$$D_T = \nu \frac{\partial \sqrt{k_T}}{\partial x_i} \frac{\partial \sqrt{k_T}}{\partial x_i} \ , \tag{24}$$

$$D_L = \nu \frac{\partial \sqrt{k_L}}{\partial x_i} \frac{\partial \sqrt{k_L}}{\partial x_i} \ . \tag{25}$$

The turbulent transport terms in the k_T and ω equations include a turbulent effective diffusivity, α_T, defined as

$$\alpha_T = f_\nu C_{\mu,std} \sqrt{k_{T,s}} \lambda_{eff} \ . \tag{26}$$

The boundary layer production term, P_{BL}, is added to reproduce properly the behaviour of the boundary layer wake region, and is defined as:

$$P_{BL} = C_{\omega 3} f_\omega \alpha_T f_W^2 \frac{\sqrt{k_T}}{d^3} \ , \tag{27}$$

where

$$f_\omega = 1 - \exp\left[-0.41 \left(\frac{\lambda_{eff}}{\lambda_T} \right)^4 \right] \tag{28}$$

is the kinematic damping function.

The remaining terms in the transport equations are related to the laminr-turbulent transition mechanism in the model. As mentioned above, transition occurs as a transfer of energy from k_L

to k_T, with a concurrent reduction in turbulence length scale from the free-stream value to the value found in an equilibrium turbulent boundary layer.

The transport equation for turbulent kinetic energy, k_T, and the transport equation for laminar kinetic energy, k_L, are linked by the terms R_{BP} and R_{NAT}, which represent the break-down process of the pre-transitional fluctuations. Once a certain threshold is reached (β_{BP} and β_{NAT}), these terms become a sink to the k_L transport equation and a source to the k_T equation. R_{BP} and R_{NAT} account for the bypass and natural transition mechanism, respectively, and are defined as:

$$R_{BP} = C_R \beta_{BP} k_L \omega / f_W \ , \tag{29}$$

$$R_{NAT} = C_{R,NAT} \beta_{NAT} k_L \Omega \ , \tag{30}$$

where

$$\beta_{BP} = 1 - \exp\left(-\frac{\phi_{BP}}{A_{BP}}\right) \ , \tag{31}$$

$$\phi_{BP} = \max\left[\left(\frac{k_T}{\nu\Omega} - C_{BP,crit}\right), 0\right] \ , \tag{32}$$

$$\beta_{NAT} = 1 - \exp\left(-\frac{\phi_{NAT}}{A_{NAT}}\right) \ , \tag{33}$$

$$\phi_{NAT} = \max\left(Re_\Omega - \frac{C_{NAT,crit}}{f_{NAT,crit}}, 0\right) \ , \tag{34}$$

$$f_{NAT,crit} = 1 - \exp\left(-C_{NC}\frac{\sqrt{k_L}d}{\nu}\right) \ , \tag{35}$$

Transition in both cases is assumed to initiate when the characteristic time-scale for turbulence production is smaller than the viscous diffusion time-scale of the pre-transitional fluctuations.

The turbulent kinematic viscosity used in the momentum equations is the sum of the small-scale and large-scale contributions defined above

$$\nu_T = \nu_{T,s} + \nu_{T,l} \ . \tag{36}$$

When including heat transfer effects, the turbulent thermal diffusivity, α_θ, is modelled as

$$\alpha_\theta = f_W \frac{k_T}{k_{TOT}} \frac{\nu_{T,s}}{Pr} + (1 - f_W)\, C_{\alpha,\theta} \sqrt{k_T} \lambda_{eff} \ . \tag{37}$$

As regards the dissiparion rate transport equation, the first term on the right hand side of Eq. 6 represents an increase in dissipation rate due to turbulence production. The damping function of the third term is of the following form

$$f_2 = C_{\omega 2} f_W^2 \ , \tag{38}$$

while all the other model constants are reported in Tab. 1.

Table 1: Summary of model constants

$A_0 = 4.04$	$A_S = 2.12$	$A_\nu = 6.75$
$A_{BP} = 0.6$	$A_{NAT} = 200$	$A_{TS} = 200$
$C_{BP,crit} = 1.2$	$C_{NC} = 0.1$	$C_{NAT,crit} = 1250$
$C_{INT} = 0.75$	$C_{TS,crit} = 1000$	$C_{R,NAT} = 0.02$
$C_{l1} = 3.4 \times 10^{-6}$	$C_{l2} = 10^{-10}$	$C_R = 0.12$
$C_{\alpha,\theta} = 0.035$	$C_{SS} = 1.5$	$C_{\tau,l} = 4360$
$C_{\omega 1} = 0.44$	$C_{\omega 2} = 0.92$	$C_{\omega 3} = 0.3$
$C_{\omega R} = 1.5$	$C_\lambda = 2.495$	$C_{\mu,std} = 0.09$
$Pr = 0.85$	$\sigma_k = 1$	$\sigma_\omega = 1.17$

3 DG SPACE and TIME DISCRETIZATION

The governing equations can be written in compact form as

$$\frac{\partial \mathbf{q}}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{F}_c(\mathbf{q}) + \boldsymbol{\nabla} \cdot \mathbf{F}_v(\mathbf{q}, \boldsymbol{\nabla}\mathbf{q}) + \mathbf{s}(\mathbf{q}, \boldsymbol{\nabla}\mathbf{q}) = \mathbf{0}, \tag{39}$$

where $\mathbf{q} \in \mathbb{R}^m$ denotes the vector of the m primitive variables, $\mathbf{s} \in \mathbb{R}^m$ the source term, d the space dimension, $\mathbf{F}_c, \mathbf{F}_v \in \mathbb{R}^M \otimes \mathbb{R}^N$ the inviscid and viscous flux functions.

A weak formulation of Eq. (39) is obtained multiplying each scalar law by an arbitrary smooth test function $v_j \in \mathbf{v}$, $1 \leq j \leq m$, and integrating by parts:

$$\int_\Omega v_j \frac{\partial q_j}{\partial t} \, \mathrm{d}\mathbf{x} - \int_\Omega \boldsymbol{\nabla} v_j \cdot \mathbf{F}_j(\mathbf{q}, \boldsymbol{\nabla}\mathbf{q}) \, \mathrm{d}\mathbf{x} + \int_{\partial\Omega} v_j \mathbf{F}_j(\mathbf{q}, \boldsymbol{\nabla}\mathbf{q}) \cdot \mathbf{n} \, \mathrm{d}\sigma + \int_\Omega v_j \mathbf{s}_j(\mathbf{q}, \boldsymbol{\nabla}\mathbf{q}) \, \mathrm{d}\mathbf{x} = 0, \tag{40}$$

where $\mathbf{F}_j$ is the sum of the inviscid and viscous flux vectors, Ω the computational domain, $\partial\Omega$ the boundary of Ω, $\mathbf{n}$ the unit normal vector to the boundary.

Let Ω_h be an approximation of the domain $\Omega \in \mathbb{R}^d$, $\mathcal{T}_h = \{K\}$ a mesh of Ω_h, i.e. a collection of "finite elements" K, $\mathcal{F}_h = \{F\}$ the mesh faces, and let $\mathbf{V}_h$ denotes a discontinuous finite element space spanned by polynomial functions continuous only inside each element K, i.e.

$$\mathbf{V}_h \stackrel{\text{def}}{=} \left[\mathbb{P}_d^l(\mathcal{T}_h)\right]^m, \tag{41}$$

where

$$\mathbb{P}_d^l \stackrel{\text{def}}{=} \{v_h \in L^2(\Omega_h) : v_h|_K \in \mathbb{P}_d^l, \forall K \in \mathcal{T}_h\} \tag{42}$$

is the space of polynomials of degree at most l on the element K. Hierarchical and orthogonal shape functions are adopted and are obtained using a modified Gram-Schmidt procedure, assuming as a starting point a set of monomial functions [12]. The solution $\mathbf{q}$, the test function $\mathbf{v}$ are replaced with finite element approximations $\mathbf{q}_h$ and $\mathbf{v}_h$, belonging to the space $\mathbf{V}_h$. The DG

formulation of the problem (40) requires to find $\mathbf{q}_h \in \mathbf{V}_h$ such that

$$\sum_{K\in\mathcal{T}_h}\int_K v_{h,j}\frac{\partial q_{h,j}}{\partial t}\,\mathrm{d}\mathbf{x} - \sum_{K\in\mathcal{T}_h}\int_K \boldsymbol{\nabla}_h v_{h,j}\cdot\mathbf{F}_j\left(\mathbf{q}_h,\boldsymbol{\nabla}_h\mathbf{q}_h+\mathbf{r}(\llbracket\mathbf{q}_h\rrbracket)\right)\,\mathrm{d}\mathbf{x} + \sum_{F\in\mathcal{F}_h}\int_F \llbracket v_{h,j}\rrbracket\cdot\widehat{\mathbf{f}}_j\left(\mathbf{q}_h^{\pm},(\boldsymbol{\nabla}_h\mathbf{q}_h+\eta_F\mathbf{r}_F(\llbracket\mathbf{q}_h\rrbracket))^{\pm}\right)\,\mathrm{d}\sigma + \sum_{K\in\mathcal{T}_h}\int_K v_{h,j}\mathbf{s}_j\left(\mathbf{q}_h,\boldsymbol{\nabla}_h\mathbf{q}_h+\mathbf{r}(\llbracket\mathbf{q}_h\rrbracket)\right)\,\mathrm{d}\mathbf{x}=\mathbf{0} \qquad \mathbf{v}_h\in\mathbf{V}_h, \quad (43)$$

where $\mathbf{r}$ and $\mathbf{r}_F$ are the global and the local lifting operators, η_F a stability parameter, and $\widehat{\mathbf{f}}$ is the sum of the inviscid and the viscous numerical fluxes.

The numerical flux function $\widehat{\mathbf{f}}$, appearing in the boundary integral of Eq. (43), is introduced in order to uniquely define the flux at the elements interfaces, to obtain a consistent and conservative approximation of Eq. (40), and to prescribe the boundary data. The flux $\widehat{\mathbf{f}}$ is the sum of an inviscid, $\widehat{\mathbf{f}}_c$, and a viscous, $\widehat{\mathbf{f}}_v$, part. The former is based on the Godunov flux computed with an exact Riemann solver. For the latter the BR2 scheme, proposed in [11] and theoretically analyzed in [9, 10], is employed.

The DG space discretization of Eq. (43) results in the following system of (nonlinear) ODEs in time

$$\mathbf{M}\frac{d\mathbf{Q}}{dt}+\mathbf{R}\left(\mathbf{Q}\right)=0, \quad (44)$$

where $\mathbf{Q}$ is the global vector of unknown degrees of freedom, $\mathbf{M}$ is a global block diagonal matrix and $\mathbf{R}\left(\mathbf{Q}\right)$ is the vector of "residuals".

In the case of steady state computations the semi-discrete problem in Eq. (44) is discretized in time by means of the classical backward Euler scheme coupled with the pseudo-transient continuation strategy proposed in [14]. The resulting linear system is solved at each time step by means of GMRES algorithm. Linear algebra and parallelization are handled through PETSc library [13].

4 RESULTS

The implementation of the transition model was verified and assessed in the computation of two test cases, representative of the main transition modes: the compressible turbulent flow over a flat plate (T3A and T3B configurations of the ERCOFTAC SIG 10) and through the T106A turbine cascade. The influence of the polynomial degree, and of coarse grids on the solution accuracy were investigated, comparing results with available experimental data. All the computations were run in parallel, initializing the $\mathbb{P}^0$ solution from the uniform flow at inflow conditions and the higher-order solutions from the lower-order ones.

4.1 Zero-pressure-gradient flat plate

The transition model is first verified and validated on a zero-pressure-gradient flat plate. The T3A and T3B test cases of the ERCOFTAC were used, which are characterized by different values of the velocity and turbulence intensity at the leading edge, as reported in Tab. 2. Inlet turbulent quantities are chosen in order to match the value of Tu at the plate leading edge and the correct decay of the turbulence kinetic energy along the plate (see Tab. 3).

Table 2: Flow conditions of the T3A and T3B test cases

	T3A	T3B
Upstream Velocity (m/s)	5.4	9.4
Leading edge Tu (%)	3.0	6.0
Pressure Gradient	0	0

Table 3: Inlet turbulent quantities of the T3A and T3B test cases

	T3A	T3B
Tu_1	10%,	8.2%
$(\mu_T/\mu)_1$	0.35	114

In order to verify the implementation of the model, a grid sensitivity analysis was performed for the T3A test case on three meshes with 3600 (coarse), 8800 (medium) and 18000 (fine) quadrilateral elements. The solution approximations were $\mathbb{P}^{1\to4}$ for all the grids, while the maximum number of DoF was 54000 for the coarse mesh, 132000 for the medium mesh and 270000 for the fine mesh.

Figures 3a-3c show the skin friction coefficient, C_f, distributions on each mesh. In Fig. 3d a comparison of the predicted C_f distribution on the three meshes for the highest polynomial order with available experimental data [15], and with a standard k-$\widetilde{\omega}$ model is presented.

For every mesh, increasing the polynomial order of the solution approximation, all the predicted C_f curves are almost overlapped, and, starting from $\mathbb{P}^1$ approximation, the solution is in agreement with experimental data. Notice that the $\mathbb{P}^1$ solution on the coarse mesh corresponds only to 10800 DoF.

Also for the T3B test case, a grid sensitivity analysis was performed, using the same meshes (see Figs. 1 and 2).

Figures 4a-4c show the skin friction coefficient, C_f, distributions on each mesh. In Fig. 4d a comparison of the predicted C_f distribution on the three meshes for the highest polynomial order with available experimental data [15], and with a standard k-$\widetilde{\omega}$ model is presented. Also in this case the predicted curves are in good agreement with experimental data. The transition onset is slightly moved up, while the end of the transition is captured correctly.

For both test cases, the C_f distributions obtained with the standard k-$\widetilde{\omega}$ is also depicted in Figs. 3d and 4d, showing a completely wrong behaviour, *i.e.* a fast transition at the beginning of the plate.

4.2 MTU T106A

In this section the subsonic turbulent flow through the MTU T106A turbine cascade is considered. This cascade, designed by MTU Aero Engines, has been extensively investigated in experimental and computational studies [19, 18] and is characterized by a complex transitional turbulent flow.

An adiabatic wall boundary condition was imposed on blade surface. At the inflow, the total temperature, total pressure, flow angle $\alpha_1 = 37.7°$, and turbulence intensity $Tu_1 = 4\%$ were

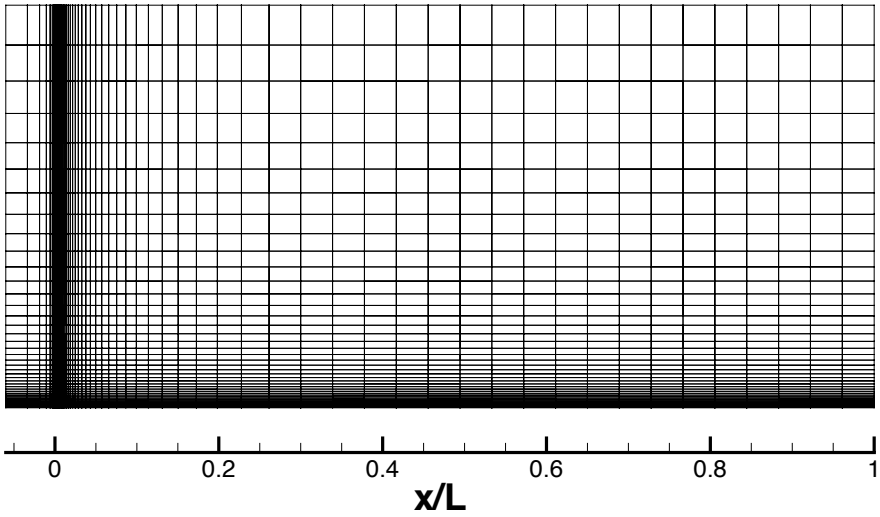
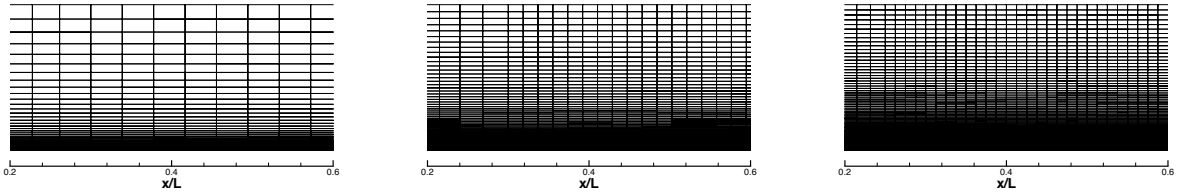
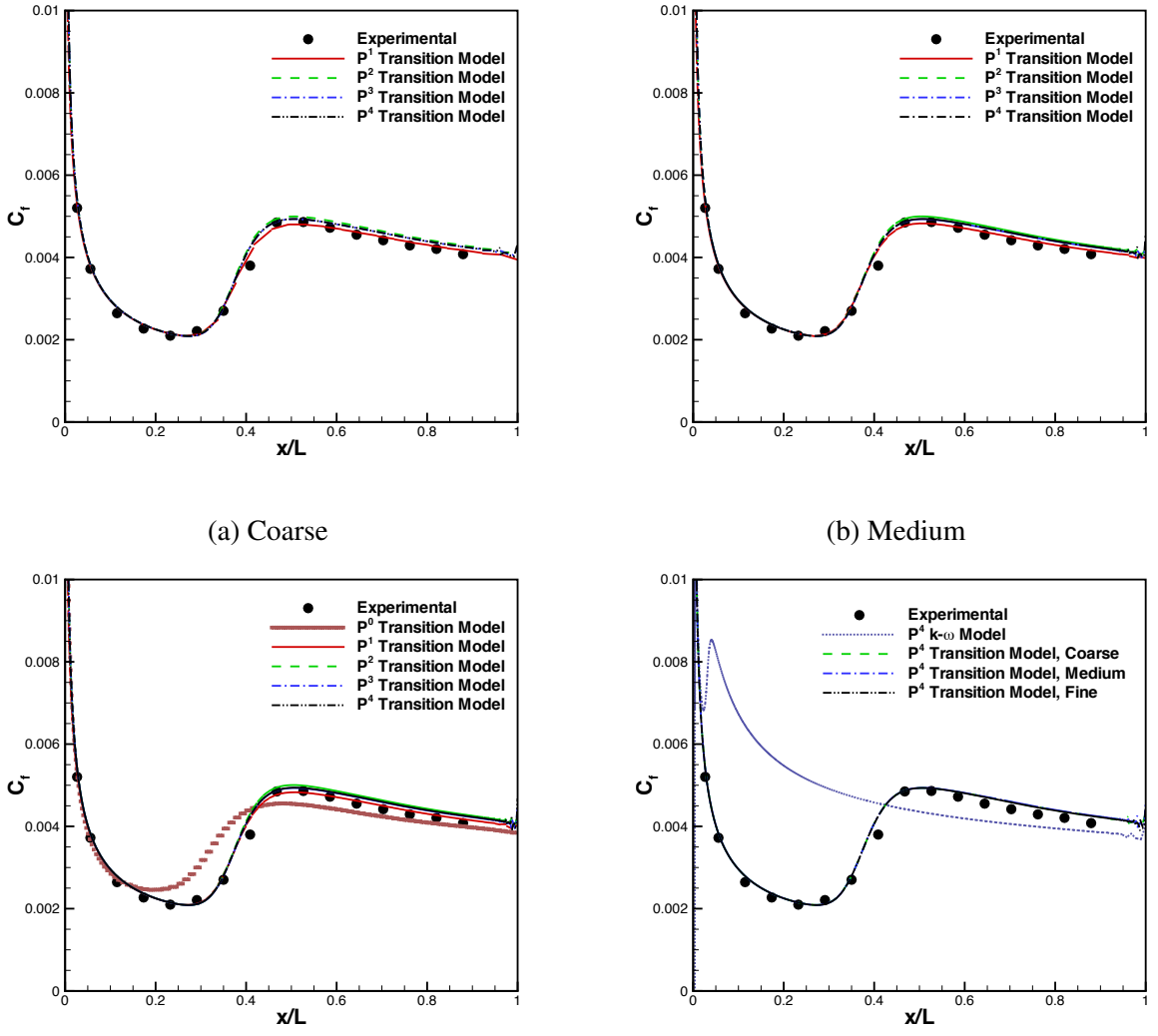


Figure 1: Flat plate - 3600 (coarse) quadrilateral elements mesh with linear edges for the flat plate test cases.



(a) Coarse (b) Medium (c) Fine

Figure 2: Flat plate - zoom in on the transition region for the coarse (a), medium (b), and fine (c) meshes.



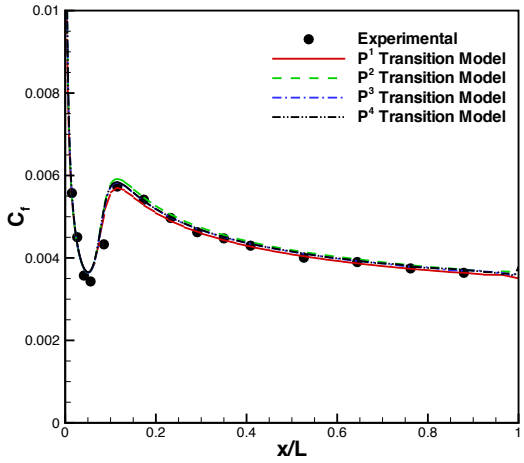
(a) Coarse

(b) Medium

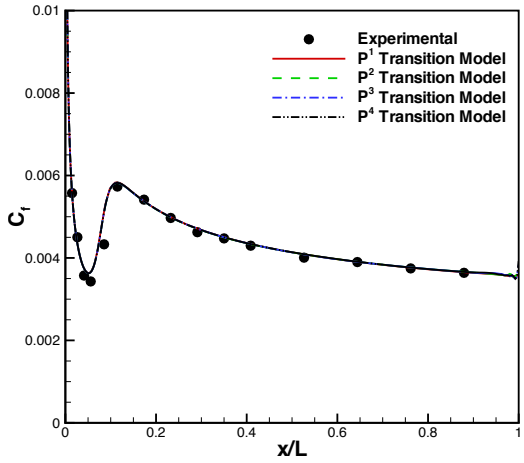
(c) Fine

(d) Comparison

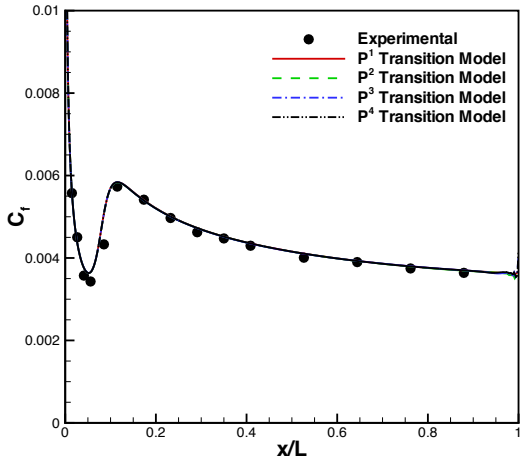
Figure 3: Flat plate T3A - distribution of the skin friction coefficient for the coarse (a), medium (b), fine (c) meshes, and comparison of the predicted skin friction distributions on the considered meshes for the highest polynomial order ($\mathbb{P}^4$) with a fully turbulent solution and experimental data (d).



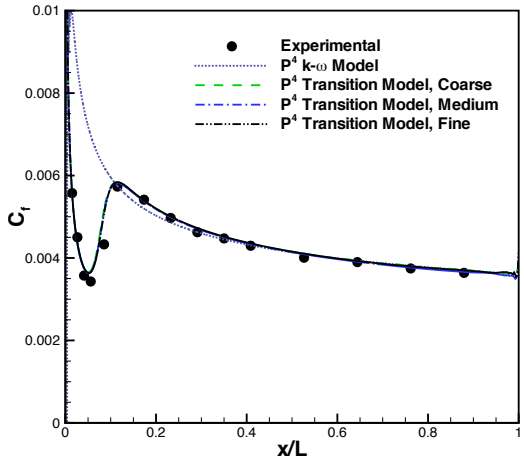
(a) Coarse



(b) Medium



(c) Fine



(d) Comparison

Figure 4: Flat plate T3B - distribution of the skin friction coefficient for the coarse (a), medium (b), fine (c) meshes, and comparison of the predicted skin friction distributions on the considered meshes for the highest polynomial order ($\mathbb{P}^4$) with a fully turbulent solution and experimental data (d)

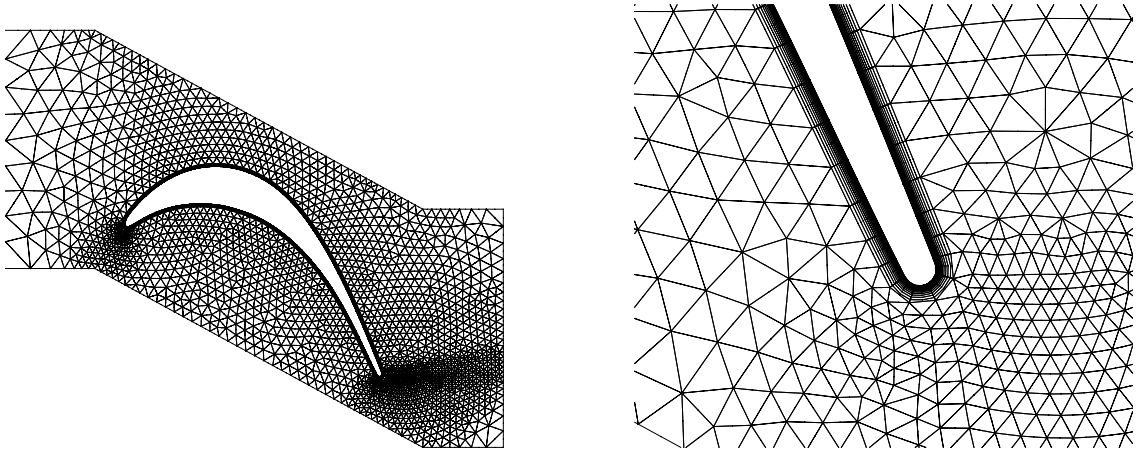


Figure 5: T106A, $Re_{2,is} = 1.1 \times 10^6$ - $y^+ = 5$ grid (left) and detail of the grid near the trailing edge (right)

prescribed. At the outflow, the static pressure was set, resulting in a downstream isentropic Mach number $M_{2,is} = 0.59$. The Reynolds number based on the downstream isentropic conditions and on the blade chord is $Re_{2,is} = 1.1 \times 10^6$. To assess the predicting capability of the model, the pressure coefficient, C_p, and the skin friction, C_f, distribution along the blade and the loss coefficient, ζ, were compared with experimental data. The flow field is attached for this flow condition and the prediction capabilities of the natural and bypass transition modes can be investigated.

All computations were performed by using a hybrid grid with a height of the first cell adjacent to the wall $y^+ = 5$ and 5743 quadratic elements. The grids employed for the simulations were generated with a 2D high-order version of a fully automated in-house hybrid mesh generator based on the advancing-Delaunay strategy [16]. The full grid and a detail of the trailing edge region can be seen in Fig. 5.

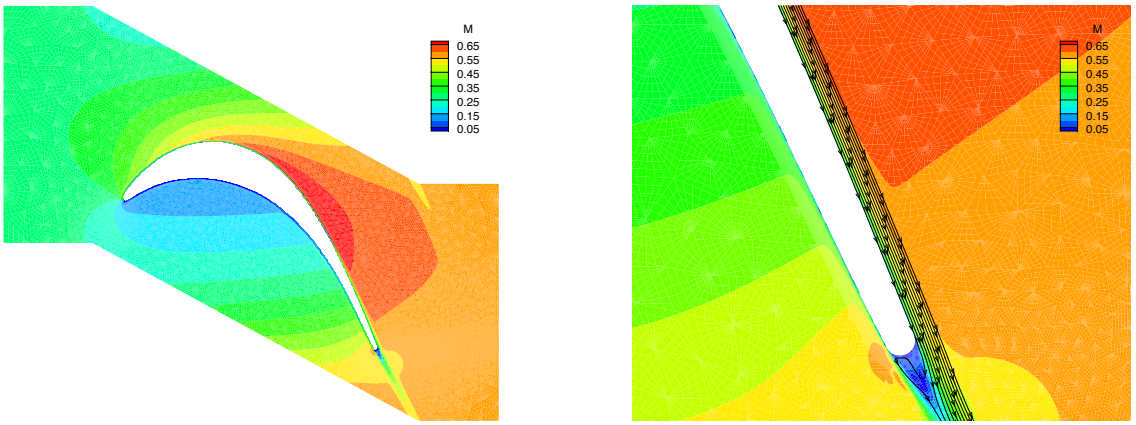


Figure 6: T106A, $Re_{2,is} = 1.1 \times 10^6$ - Mach number contours, $y^+ = 5$ grid, $\mathbb{P}^4$ approximation

Fig. 6 (left) shows the computed $\mathbb{P}^4$ Mach number contours on the $y^+ = 5$ mesh, while in Fig. 6 (right) the streamlines in the transition region confirms the absence of the flow separation.

Figure 7 shows the C_p distribution along the blade for $\mathbb{P}^{1\to4}$ approximation. Starting from $\mathbb{P}^2$ approximation, the curves on pressure and suction sides are in good agreement with the

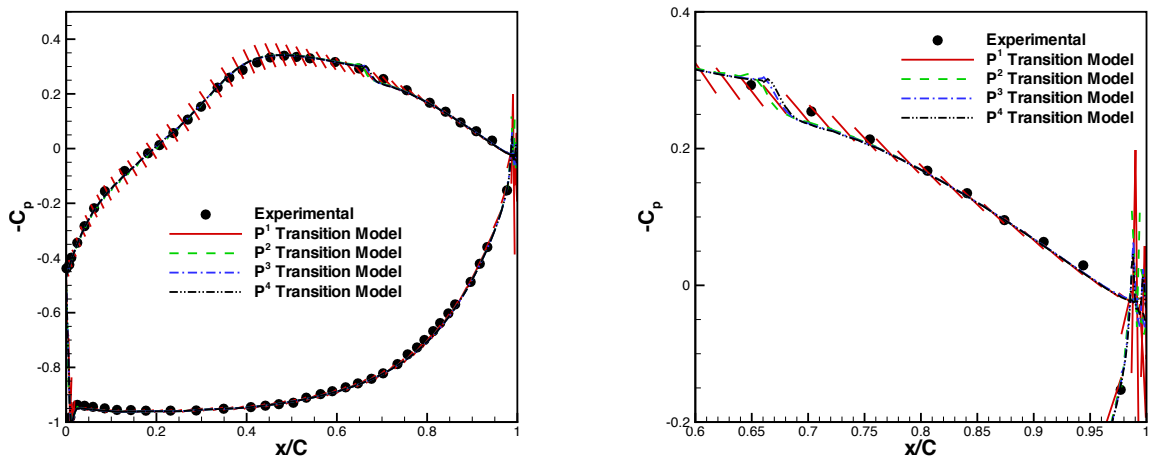


Figure 7: T106A, $Re_{2,is} = 1.1 \times 10^6$ - distribution of the pressure coefficient along the blade (left), and zoom in on the transition region (right), $y^+ = 5$ grid, $\mathbb{P}^{1\to4}$ approximations

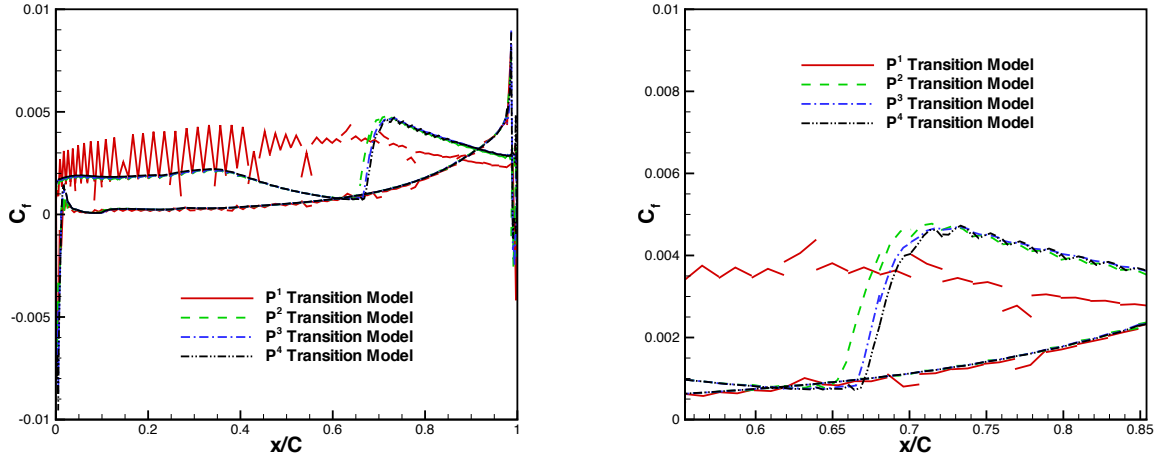


Figure 8: T106A, $Re_{2,is} = 1.1 \times 10^6$ - distribution of the skin friction coefficient along the blade (left), and zoom in on the transition region (right), $y^+ = 5$ grid, $\mathbb{P}^{1\to4}$ approximations

experimental data.

Fig. 8 shows the C_f distribution along the blade for all the polynomial approximations on the $y^+ = 5$ grid. The transition location moves slightly downstream, increasing the polynomial degree of the approximation. The $\mathbb{P}^1$ curve has rather large oscillations, which suggest that the spatial approximation need to be increased. Starting from $\mathbb{P}^2$ approximation, the transition location is predicted at approximately 65% of the chord, where a sudden increase of the C_f value can be observed.

Finally, in Tab. 4 the loss coefficient, ζ, is reported for each polynomial degree. As expected, it was overestimated by the low-order approximation, while the transition model allowed to predict almost the correct value starting from the $\mathbb{P}^2$ approximation.

5 CONCLUSIONS

The parallel implementation of a local transition model based on the laminar kinetic energy concept into a high-order accurate DG code was presented. The model was validated on the flat plate test case (T3A and T3B configurations) and we assessed the capability of the model to

Table 4: Loss coefficient, ζ, for the T106A test case.

	$\mathbb{P}^1$	$\mathbb{P}^2$	$\mathbb{P}^3$	$\mathbb{P}^4$	Experimental
$Re = 1.1 \times 10^6$, $y^+ = 5$	0.028	0.0204	0.0203	0.0202	0.02

compute complex transitional turbomachinery flows (MTU T106A turbine cascade). The beneficial effect of the transition model on the solution accuracy was demonstrated by comparing numerical results for increasing order of accuracy with fully turbulent solution and the available experimental data. Ongoing work is devoted to improve prediction capabilities of the transition model, considering also the laminar separation mode.

REFERENCES

[1] F.R. Menter, R.B. Langtry, S.R. Likki, Y.B. Suzen, P.G. Huang, S. Völker, A correlation-based transition model using local variables - Part I: Model formulation. *Journal of Turbomachinery*, **128**, 413–422, 2006.

[2] F.R. Menter, R.B. Langtry, S.R. Likki, Y.B. Suzen, P.G. Huang, S. Völker, A correlation-based transition model using local variables - Part II: Test cases and industrial applications. *Journal of Turbomachinery*, **128**, 423–434, 2006.

[3] D.K. Walters, D. Cokljat, A three-equation eddy-viscosity model for Reynolds-averaged Navier-Stokes simulations of transitional flow. *Journal of Fluids Engineering, Transactions of the ASME*, **130**, 121401-1–121401-14, 2008.

[4] F. Bassi, L. Botti, A. Colombo, A. Crivellini, C. De Bartolo, N. Franchina, A. Ghidoni, S. Rebay, Time integration in the discontinuous Galerkin code MIGALE – Steady problems. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, **128**, 179–204, 2015.

[5] F. Bassi, L. Botti, A. Colombo, A. Crivellini, A. Ghidoni, A. Nigro, S. Rebay, Time integration in the discontinuous Galerkin code MIGALE – Unsteady problems . *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, **128**, 179–204, 2015.

[6] T.H. Shih, W.W. Liou, A. Shabbir, Z. Yang, J. Zhu, New k-eddy viscosity model for high Reynolds number turbulent flows. *Comput. & Fluids*, **24**, 227–238, 1995.

[7] R.J. Volino, T.W. Simon, Boundary Layer Transition Under High Free-Stream Turbulence and Strong Acceleration Conditions. *Journal of Turbomachinery*, **119**, 420–432, 1997.

[8] B.E. Launder, B.I. Sharma, Application of the Energy-dissipation Model of Turbulence to the Calculation of Flow Near a Spinning Disc. *Letters in Heat and Mass Transfer*, **2**, 131–138, 1974.

[9] F. Brezzi, G. Manzini, D. Marini, P. Pietra, A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations*, **16**, 365–378, 2000.

[10] D. N. Arnold, F. Brezzi, B. Cockburn, D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, **39**, 1749–1779, 2002.

[11] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In: R. Decuypere and G. Dibelius. *Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, Technologisch Instituut, Antwerpen, Belgium, 99–108, 1997.

[12] F. Bassi, L. Botti, A. Colombo, D. A. Di Pietro, P. Tesini. On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. *Journal of Computational Physics*, **231**, 45–65, 2012.

[13] S. Balay, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, H. Zhang, 2001. PETSc Web page. Http://www.mcs.anl.gov/petsc.

[14] F. Bassi F, L. Botti, A. Colombo, A. Crivellini, N. Franchina, A. Ghidoni, S. Rebay. Very high-order accurate discontinuous Galerkin computation of transonic turbulent flows on aeronautical configurations. *ADIGMA - A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*, *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, vol. 113, Kroll N, Bieler H, Deconinck H, Couaillier V, van der Ven H, Sørensen K (eds.). Springer Berlin / Heidelberg, 2010; 25–38.

[15] J. Coupland. ERCOFTAC Special Interest Group on Laminar to Turbulent Transition and Retransition: T3A and T3B test cases, *Technical Report*. ERCOFTAC, 1990.

[16] A. Ghidoni, E. Pelizzari, S. Rebay, V. Selmin, V. 3D anisotropic unstructured grid generation. *International Journal for Numerical Methods in Fluids*, **51**, 1097-1115, 2006.

[17] M. Lorini, F. Bassi, A. Colombo, A. Ghidoni. High-order implementation of a non-local transition model in a DG solver for turbomachinery applications. *Computers & Fluids*, **127**, 115-130, 2016.

[18] A. Ghidoni, A. Colombo, S. Rebay, F. Bassi. Simulation of the Transitional Flow in a Low Pressure Gas Turbine Cascade with a High-Order Discontinuous Galerkin Method. *Journal of Fluids Engineering*, **135**, 2012.

[19] H. Hoheisel. Entwicklung neuer Entwurfskonzepte fur zwei Turbinengitter, Teil III, Ergebnisse T106. *Technical Report*, Institut fur Entwurfsaerodynamik, Braunschweig,1981.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A STUDY OF MULTIGRID SMOOTHERS USED IN COMPRESSIBLE CFD BASED ON THE CONVECTION DIFFUSION EQUATION

Philipp Birken[1], **Jonathan Bull**[2], **and Antony Jameson**[3]

[1]Lund University, Centre for the Mathematical Sciences, Numerical Analysis
Box 118, 22100 Lund
e-mail: philipp.birken@na.lu.se

[2] Uppsala University, Division of Scientific Computing, Dept. of Information Technology
Box 337, 75105 Uppsala
e-mail: jonathan.bull@it.uu.se

[3] Stanford University, Department of Aeronautics & Astronautics
Stanford, CA 94305
e-mail: jameson@baboon.stanford.edu

Keywords: Unsteady flows, Multigrid

Abstract. *We look at multigrid methods for unsteady viscous compressible flows. We specifically target smoothers that can be used in parallel and without computation of a Jacobian, which are particlarly attractive candidates in the context of Discontinuous Galerkin discretizations. In CFD, a plethora of nonlinear smoothers have been suggested which are hard to analyze. Our methodology is to use a linear model problem, here the convection diffusion equation, to be able to classify and compare smoothers better. Specifically, we consider explicit and implicit pseudo time iterations, GMRES as a smoother, SGS and implicit line smoothers. We relate GMRES to explicit Runge-Kutta smoothers, identify implicit line smoothers as Block Jacobi and analyze the potential of implicit pseudo time iterations. Finally, we discuss the relation between methods for steady and unsteady flows. Numerical results show that GMRES is a very attractive smoother in this context.*

1 INTRODUCTION

After decades of research, numerical methods for solving steady fluid flows have reached an appreciable level of maturity. It has been demonstrated that steady Euler flows can be solved with 'textbook' efficiency in three to five multigrid iterations [3]. The added stiffness of the steady RANS equations means that convergence of the multigrid method is somewhat slower - between 50 and 100 iterations [11] - but still within acceptable bounds for CFD applications.

For unsteady problems, the expectation would be to significantly reduce this number, but this is not the case for the available methods [2]. When looking at high order methods, which will be necessary to do LES simulations in an industrial setting, there is the situation that a fast multigrid method is quite simply missing. On reason for this is the lack of guiding theory for nonelliptic equations, which makes designing a multigrid method difficult.

Multigrid consists of two components, the smoother and the coarse grid correction. For finite volume methods, the coarse grid correction is based on agglomeration of neighboring cells. As smoothers for compressible flows, a plethora of both linear and nonlinear methods has been suggested. This ranges from LU-SGS [3], explicit Runge-Kutta methods [13], additive Runge-Kutta methods [5], implicit line smoothers [12, 8], and more recently point implicit and SDIRK smoothers [7, 9], GMRES [4], GMRES within a Rosenbrock pseudo time iteration [10] and many others. In particular, implicit smoothers, that are variants of implicit pseudo time iteration smoothers, allow for a huge variety of methods.

When thinking of LES simulations with discontinuous Galerkin methods, we have to keep in mind that the Jacobians will consist of blocks that have a size of a few hundred, instead of just five as it would be with finite volume methods. Our strategy to obtain a parallel scaling, fast and low memory multigrid method is therefore to not use a Jacobian, but only function evaluations, meaning evaluations of the spatial discretization. This can happen in one of two ways: There is a nonlinear version of a linear iterative method that avoids use of the Jacobian. This is for example the case for pseudo time iterations where a matrix vector product corresponds to a function evaluation. The other option is to replace matrix vector products by a Jacobian free finite difference.

Previous work by one of the authors [1] demonstrated that the smoothers used in the multigrid iteration could be optimized for fastest convergence of the unsteady linear advection equation. Specifically, it was suggested to numerically minimize the spectral radius of the multigrid iteration matrix over the pseudo timestep and coefficients of explicit Runge-Kutta (RK) schemes. Significant improvements were demonstrated in both cases compared to non-optimized smoothers.

In this paper we aim at getting a better understanding of the differences between these methods for both the steady and the unsteady case. To this end, we classify these methods, The findings will be backed up by numerical results based on the convection diffusion equation as a model problem.

2 GOVERNING EQUATIONS AND DISCRETIZATION

We consider the linear advection diffusion equation

$$u_t + au_x - bu_{xx} = 0. \tag{1}$$

with $a, b > 0$ on the interval $x \in [0, 2]$ with periodic boundary conditions.

An equidistant FV discretization for (1) with mesh width Δx leads to the evolution equation

for the cell average u_i in one cell i:

$$u_{i_t} + \frac{a}{\Delta x}(u_i - u_{i-1}) + \frac{b}{\Delta x^2}(-u_{i+1} + 2u_i - u_{i-1}) = 0.$$

Using the vector $\mathbf{u} = (u_1, ..., u_m)^T$ and the matrices

$$\mathbf{B} = \begin{pmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} 2 & -1 & & & -1 \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ -1 & & & -1 & 2 \end{pmatrix}$$

we obtain the system of ODEs

$$\mathbf{u}_t + \frac{a}{\Delta x}\mathbf{B}\mathbf{u}(t) + \frac{b}{\Delta x^2}\mathbf{C}\mathbf{u}(t) = \mathbf{0}. \tag{2}$$

In 2D, we have the following equation for the unknown function $u(x, y, t)$:

$$u_t + a \cdot \nabla u - b\Delta u = 0, \tag{3}$$

where $b > 0$ and

$$a = \tilde{a} \begin{pmatrix} \sin\gamma \\ \cos\gamma \end{pmatrix}$$

with $\tilde{\beta} \in \mathbf{R}$ being another parameter and γ the angle of the direction of forced convection. Regarding boundary conditions, we use periodic ones.

An equidistant FV discretization for (3) with mesh width $\Delta x = \Delta y$ leads to the evolution equation for the cell average $u_{i,j}$ in one cell (i, j):

$$u_{i,j_t} + \frac{\tilde{a}}{\Delta x}((c+s)u_{i,j} - su_{i-1,j} - cu_{i,j-1})$$
$$+\frac{b}{\Delta x^2}(-u_{i+1,j} - u_{i,j+1} + 4u_{i,j} - u_{i-1,j} - u_{i,j-1}) = 0.$$

Using the vector $\mathbf{u} = (u_1, ..., u_m)^T$ and the matrices

$$\mathbf{B} = \begin{pmatrix} \bar{\mathbf{B}} & -s\mathbf{I} & & \\ & \bar{\mathbf{B}} & \ddots & \\ & & \ddots & -s\mathbf{I} \\ -s\mathbf{I} & & & \bar{\mathbf{B}} \end{pmatrix},$$

where

$$\bar{\mathbf{B}} = \begin{pmatrix} c+s & & & & -c \\ -c & c+s & & & \\ & -c & c+s & & \\ & & \ddots & \ddots & \\ & & & -c & c+s \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} \bar{\mathbf{C}} & -\mathbf{I} & & -\mathbf{I} \\ -\mathbf{I} & \bar{\mathbf{C}} & -\mathbf{I} & \\ & \ddots & \ddots & \ddots \\ -\mathbf{I} & & -\mathbf{I} & \bar{\mathbf{C}} \end{pmatrix},$$

where

$$\bar{\mathbf{C}} = \begin{pmatrix} 4 & -1 & & & -1 \\ -1 & 4 & -1 & & \\ & -1 & 4 & \ddots & \\ & & \ddots & \ddots & -1 \\ -1 & & & -1 & 4 \end{pmatrix}.$$

Finally, we obtain the system of ODEs

$$\mathbf{u}_t + \frac{\tilde{a}}{\Delta x}\mathbf{B}\mathbf{u}(t) + \frac{b}{\Delta x^2}\mathbf{C}\mathbf{u}(t) = \mathbf{0}. \tag{4}$$

We discretize this in time using implicit Euler with time step size Δt, which is also a building block for the more general diagonally implicit Runge-Kutta (DIRK) methods. Thus, in each time step, a linear system has to be solved. Using the notation $\mathbf{u}^n \approx \mathbf{u}(t_n)$, this can be written as

$$\begin{aligned} &\mathbf{u}^{n+1} - \mathbf{u}^n + \frac{\tilde{a}\Delta t}{\Delta x}\mathbf{B}\mathbf{u}^{n+1} + \frac{b\Delta t}{\Delta x^2}\mathbf{C}\mathbf{u}^{n+1} = \mathbf{0} \\ &\Leftrightarrow \mathbf{u}^n - \mathbf{A}\mathbf{u}^{n+1} = 0 \end{aligned} \tag{5}$$

where

$$\mathbf{A} = \left(\mathbf{I} + \frac{\nu}{\Delta x}\mathbf{B} + \frac{\mu}{\Delta x^2}\mathbf{C}\right) \tag{6}$$

with $\nu = \tilde{a}\Delta t$ and $\mu = b\Delta t$. Here, $CFL := \tilde{a}\Delta t/\Delta x$ corresponds to the CFL number in the implicit Euler method. If we consider nonperiodic boundary conditions, the entry in the upper right corner of $\mathbf{B}$ becomes zero. Otherwise, additional terms appear on the right hand side, but this does not affect multigrid convergence.

In a steady state, the time dependence vanishes and we obtain the equation system

$$\frac{a}{\Delta x}\mathbf{B}\mathbf{u} = \mathbf{0}. \tag{7}$$

3 LINEAR ITERATIVE METHODS FOR LINEAR EQUATION SYSTEMS

A very important type of iterative method for the linear equation system

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{A} \in \mathbf{R}^{m\times m}, \mathbf{x}, \mathbf{b} \in \mathbf{R}^m \tag{8}$$

are linear iterative methods which can be written as

$$\mathbf{x}^{k+1} = \mathbf{M}\mathbf{x}^k + \mathbf{N}^{-1}\mathbf{b},\ \mathbf{M}, \mathbf{N} \in \mathbf{R}^{m\times m}.$$

It can be shown that this converges if and only of $\rho(\mathbf{M}) < 1$, which is why $\mathbf{M}$ is called the iteration matrix. The second matrix determines the limit in case of convergence. To get convergence to the solution of the linear equation system 8, it is required that $\mathbf{M} = \mathbf{I} - \mathbf{N}^{-1}\mathbf{A}$.

4 MULTIGRID METHOD

As a solver, an agglomeration multigrid method is used, which corresponds best to finite volume discretizations. Thus, in the one dimensional case, the restriction and prolongation correspond to joining and dividing neighboring cells and are given by

$$\mathbf{R} = \frac{1}{2}\begin{pmatrix} 1 & 1 & & & & \\ & & 1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & 1 \end{pmatrix} \text{ and } \mathbf{P} = 2\mathbf{R}^T = \begin{pmatrix} 1 & & & & \\ 1 & & & & \\ & 1 & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & 1 & \end{pmatrix}.$$

Hereby, it is assumed that we have an even number of cells.

The coarse grid matrix is obtained by discretizing the problem on that grid. On the coarsest level, the smoother is applied instead of the usual direct solver, since this better corresponds to the Full Approximation Scheme used for the nonlinear equations. We use a V-cycle and presmoothing only. Thus we obtain the scheme:

Function MG$(\mathbf{x}_l, \mathbf{b}_l, l)$

- $\mathbf{x}_l = \mathbf{S}_l^{\nu_1}(\mathbf{x}_l, \mathbf{b}_l)$ (Presmoothing)
- if $(l > 0)$
  - $\mathbf{r}_{l-1} = \mathbf{R}_{l-1,l}(\mathbf{b}_l - \mathbf{A}_l\mathbf{x}_l)$ (Restriction)
  - $\mathbf{v}_{l-1} = 0$
  - Call $MG(\mathbf{v}_{l-1}, \mathbf{r}_{l-1}, l-1)$ (Computation of the coarse grid correction)
  - $\mathbf{x}_l = \mathbf{x}_l + \mathbf{P}_{l,l-1}\mathbf{v}_{l-1}$ (Correction via Prolongation)
- end if

Thus, the iteration matrix of the corresponding two grid scheme is given by

$$\mathbf{M} = (\mathbf{I} - \mathbf{P}_{l,l-1}(\mathbf{N}_S^{l-1})^{-1}\mathbf{R}_{l-1,l}\mathbf{A}_l)\mathbf{M}_S^l \tag{9}$$

with $\mathbf{M}_S^l$ and $(\mathbf{N}_S^l)^{-1}$ being the matrices defining the smoother. In a standard two grid method, the matrix $\mathbf{N}_S^{-1}$ would be $\mathbf{A}_{l-1}^{-1}$ instead.

	α_1	α_2	α_3	c
ERK2	0.28			1.1
ERK4	0.04	0.12	0.32	2.0

Table 1: Coefficients for 2- and 4-stage ERK smoothers.

4.1 Smoothers

We consider a smoother to be good if the spectral radius of the iteration matrix is small. As a design criterion, this results in good smoothers, but it is expensive to use [1]. The main alternative is the smoothing factor, which is the slowest rate at which the error in one of the high frequency components is reduced by the smoother. To this end, the eigenvectors of $\mathbf{A}$ are determined and split into low and high frequency components.

4.1.1 Pseudo time iterations

The first class of smoothers we consider are time integration schemes. These approximate the solution of an initial value problem

$$\mathbf{u}_t = \mathbf{f}(\mathbf{u}), \quad \mathbf{u}_n = \mathbf{u}(t_n).$$

To apply them as iterative solvers for a linear equation system, the residual of that is used as the right hand side $\mathbf{f}(\mathbf{u})$. The differential equation resulting from (5) is given in a pseudo or dual time t^*:

$$\mathbf{u}_{t^*} = \mathbf{u}^n - \mathbf{A}\mathbf{u}(t^*), \quad \mathbf{u}(t_0^*) = \mathbf{u}^n. \tag{10}$$

One step of a pseudo time iteration thus consists of performing one step of the RK scheme for the solution of equation (10). The pseudo time step Δt^* is a parameter of this class of smoothers, as are their coefficients. However, instead of Δt^*, we will use the CFL number in pseudo time $c = a\Delta t^*/\Delta x$ as a parameter. This implies that on coarser grids, larger time steps will be chosen, leading to faster multigrid methods.

The simplest smoothers of this type are low storage explicit Runge-Kutta methods, which can be implemented using only three vectors. These are described by

$$\begin{gathered}\mathbf{u}_0 = \mathbf{u}_n \\ \mathbf{u}_j = \mathbf{u}_n + \alpha_j \Delta t^* \mathbf{f}(\mathbf{u}_{j-1}),\ j = 1, ..., s-1 \\ \mathbf{u}_{n+1} = \mathbf{u}_n + \Delta t^* \mathbf{f}(\mathbf{u}_{s-1}),\end{gathered}$$

For $s = 1$, we obtain he explicit Euler method, which is nothing but a Richardson smoother:

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \Delta t^*(\mathbf{u}^n - \mathbf{A}\mathbf{u}^k) = \mathbf{u}^k + \Delta t^* \mathbf{r}^k. \tag{11}$$

In the linear case, a general Runge-Kutta method can be represented by its stability function $R(z)$, which for explicit Runge-Kutta methods is a polynomial $P_s(z)$ of degree s. Thus, the iteration matrix is for our case given by

$$\mathbf{M} = P_s(-\Delta t \mathbf{A}).$$

The specific schemes used here are a 2-stage and a 4-stage scheme taken from [1], where the coefficients for the case of a large outer time step are taken. These can be seen in Table 1.

	i	1	2	3	4	5	c
ARK4	α_i	1/3	4/15	5/9	1	-	0.5
	β_i	1	1/2	0	0	-	
ARK 5	α_i	1/4	1/6	3/8	1/2	1	0.2
	β_i	1	0	0.56	0	0.44	

Table 2: Coefficients of additive Runge-Kutta smoothers, 4-stage and 5-stage method.

In [6], Jameson, Schmidt and Turkel introduced additive Runge-Kutta methods, where different coefficients are used for the convective and the diffusive parts. This is done to reduce the number of evaluations of the expensive diffusive terms, but also to increase the degrees of freedom in choosing a good smoother. Given a splitting in the convective and diffusive part

$$\mathbf{f}(\mathbf{u}) = \mathbf{f}^c(\mathbf{u}) + \mathbf{f}^v(\mathbf{u}),$$

Jameson suggests to implement these schemes in the following equivalent form

$$\begin{aligned} \mathbf{u}_l^{(0)} &= \mathbf{u}_l \\ \mathbf{u}_l^{(j)} &= \mathbf{u}_l - \alpha_j \Delta t_l (\mathbf{f}^{c,(j-1)} + \mathbf{f}^{v,(j-1)}), \quad j = 1, ..., s \\ \mathbf{u}_l^{n+1} &= \mathbf{u}_l^{(s)}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{f}^{c,(j)} &= \mathbf{f}^c(\mathbf{u}_l^{(j)}), \quad j = 0, ..., s-1 \\ \mathbf{f}^{v,(0)} &= \mathbf{f}^v(\mathbf{u}_l^{(0)}), \\ \mathbf{f}^{v,(j)} &= \beta_j \mathbf{f}^v(\mathbf{u}_l^{(j)}) + (1-\beta_j)\mathbf{f}^{v,(j-1)}, \quad j = 1, ..., s-1. \end{aligned}$$

For our model problems, we have $\mathbf{f}^c(\mathbf{u}) = \mathbf{b} - \mathbf{B}\mathbf{u}$ and $\mathbf{f}^v(\mathbf{u}) = -\mathbf{C}\mathbf{u}$. The iteration matrix now involves both powers of the single matrices, as well as products, making it very difficult to analyze. The schemes we consider were designed in [5] and the coefficients are given in Table 2.

The smoothing factor is the maximum of the stability function over the high frequency eigenvalues:

$$\max_{\lambda_{HF}} |R(\lambda)|.$$

Thus we look for methods with an optimal smoothing factor, found by

$$\min_{\alpha, \Delta t^*} \max_{\lambda_{HF}} |R(\lambda)|.$$

For an ERK scheme, this gives

$$\min_{\alpha, \Delta t^*} \max_{\lambda_{HF}} |P_s(\lambda)|. \tag{12}$$

The vector α contains the coefficients of the time integration method. Here, we will, instead of the pseudo time step Δt^*, use the CFL number in pseudo time, c^*, giving

$$\min_{\alpha, c^*} \max_{\lambda_{HF}} |R(\lambda)|. \tag{13}$$

In case of an implicit RK method, the stability function is rational, e.g. $R(z) = 1/(1+z)$ for the implicit Euler method. However, this implies that the application of the smoother consists of solving the original linear equation system, defying the purpose of an iterative method. Therefore, the arising system is solved using another iterative method, resulting in new families of smoothers. We will consider the particular case of the implicit Euler method, combined with GMRES.

4.2 GMRES

Given the matrix $\mathbf{A}$ and a residual vector $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0$, the space

$$\mathcal{K}_l = \text{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, ..., \mathbf{A}^{l-1}\mathbf{r}_0\}$$

is called the l-th Krylov subspace. Using this, the k-th iteration of GMRES is the solution of the problem

$$\min_{x \in x^{(0)} + \mathcal{K}_k} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2, \tag{14}$$

If $x \in x^{(0)} + \mathcal{K}_k$, then

$$\mathbf{x} = \mathbf{x}^{(0)} + \sum_{j=0}^{k-1} \gamma_j \mathbf{A}^j \mathbf{r}_0 = \mathbf{x}^{(0)} + \sum_{j=0}^{k-1} \gamma_j \mathbf{A}^j (\mathbf{b} - \mathbf{A}\mathbf{x}^0).$$

Thus, the iteration can be written as

$$\mathbf{x}^k = \sum_{j=0}^{k-1} \bar{\gamma}_j \mathbf{A}^j \mathbf{b} + \sum_{j=0}^{k} \bar{\gamma}_j \mathbf{A}^j \mathbf{x}^0$$

for some coefficients $\bar{\gamma}_j$ and the iteration matrix is

$$\mathbf{M}_k^{GMRES} = \sum_{j=0}^{k} \bar{\gamma}_j \mathbf{A}^j = p_k(\mathbf{A}).$$

Thus, the iteration for a GMRES method with fixed k steps is represented by a polynomial of degree k in $\mathbf{A}$ or otherwise put, a rational function is approximated by a polynomial. The difference to an explicit RK iteration is that the coefficients are not fixed a priori, but depend on the choice of Krylov subspace, in this case on the residual $\mathbf{r}_0$. This means that the smoother will behave differently for different initial guesses. This also means that it is difficult to analyze the smoothing factor. At least, from the definition of GMRES, we see that

$$\|\mathbf{r}_k\|_2 = \min_{p \in \Pi_k, p(0)=1} \|p(\mathbf{A})\mathbf{r}_0\|_2 \leq \|\bar{p}(\mathbf{A})\mathbf{r}_0\|_2,$$

which is similar to (12). The first difference is that here, the optimization takes into account all eigenvalues, whereas the previous one looks only at the high frequency ones. Then, the optimization here is dynamic in the sense that it is done automatically by GMRES for each linear system seperate, which makes this approach potentially more robust. For the optimized ERK schemes, it could happen that when operating in off design conditions, the performance significantly deteriorates.

When considering the use of GMRES within an implicit Runge-Kutta smoother, the following property is important. For $\mathbf{A} = \mathbf{I} + \alpha\mathbf{B}$,

$$\text{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, ..., \mathbf{A}^{l-1}\mathbf{r}_0\} = \text{span}\{\mathbf{r}_0, \mathbf{B}\mathbf{r}_0, ..., \mathbf{B}^{l-1}\mathbf{r}_0\}$$

and thus, for the same vector $\mathbf{r}_0$ the Krylov subspaces are identical. Now we look at the following two approaches:

1. Use k-step GMRES as smoother

2. Use a smoother which consists of one implicit Euler step, where the linear system is solved using k-step GMRES

In the first case, we look at the system

$$(\mathbf{I} + \Delta t\mathbf{B})\mathbf{u} = \mathbf{u}^n$$

with residual

$$\mathbf{r} = \mathbf{u}^n - (\mathbf{I} + \Delta t\mathbf{B})\mathbf{u}. \tag{15}$$

In the second case, the system is

$$[\mathbf{I} + \Delta t^*(\mathbf{I} + \Delta t\mathbf{B})]\mathbf{u} = \mathbf{u}^k + \Delta t^*\mathbf{u}^n$$

with residual

$$\mathbf{r}^* = \mathbf{u}^k + \Delta t^*\mathbf{u}^n - [\mathbf{I} + \Delta t^*(\mathbf{I} + \Delta t\mathbf{B})]\mathbf{u} = \mathbf{u}^k - \mathbf{u} + \Delta t^*\mathbf{r}. \tag{16}$$

By the property named above, the Krylov subspaces generated will be $\mathcal{K}_k(\mathbf{B}, \mathbf{r})$ and $\mathcal{K}_k(\mathbf{B}, \mathbf{u}^k - \mathbf{u} + \Delta t^*\mathbf{r})$, respectively. On the fine grid, the initial guess for the smoother is $\mathbf{u}^k$ and thus $\mathbf{r}_0^* = \mathbf{u}^k - \mathbf{u} + \Delta t^*\mathbf{r} = \Delta t^*\mathbf{r}_0$ and therefore, the two spaces are identical. However, the residuals (15) and (16) minimized by GMRES are different and therefore, unless the solution is in that space, a different result will be obtained. However, for a large value of Δt^*, this difference will vanish.

On the coarse grid, we have a zero initial guess for the equation $(\mathbf{I}+\Delta t\mathbf{B})\mathbf{e} = \mathbf{r}$ and therefore in the first case,

$$\mathbf{r}_0 = \mathbf{r} - (\mathbf{I} + \Delta t\mathbf{B})\mathbf{0} = \mathbf{r}.$$

For the second variant we have the equation

$$[\mathbf{I} + \Delta t^*(\mathbf{I} + \Delta t\mathbf{B})]\mathbf{e} = \mathbf{u}^k + \Delta t^*\mathbf{u}^n - [\mathbf{I} + \Delta t^*(\mathbf{I} + \Delta t\mathbf{B})]\mathbf{u}^{k+1}$$

and thus

$$\hat{\mathbf{r}}_0 = \mathbf{u}^k - \mathbf{u}^{k+1} + \Delta t^*\mathbf{r}.$$

Thus, on the coarse grids, the smoothers no longer operate in the same Krylov subspace. Again, for a large value of Δt^*, this difference will vanish. Thus, we expect the two smoothers to behave the same way for a sufficiently large pseudo time step.

4.3 Symmetric Gauss-Seidel

Given a splitting of the matrix $\mathbf{A}$ into its strictly lower, diagonal and strictly upper part, $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$, the symmetric Gauss-Seidel (SGS) iteration is given by

$$\mathbf{u}^{k+1} = (\mathbf{I} - (\mathbf{D} + \mathbf{U})^{-1}\mathbf{D}(\mathbf{L} + \mathbf{D})^{-1}\mathbf{A})\mathbf{u}^k + (\mathbf{D} + \mathbf{U})^{-1}\mathbf{D}(\mathbf{L} + \mathbf{D})^{-1}\mathbf{b},$$

thus the iteration matrix is

$$\mathbf{M}^{SGS} = \mathbf{I} - (\mathbf{D} + \mathbf{U})^{-1}\mathbf{D}(\mathbf{L} + \mathbf{D})^{-1}\mathbf{A}.$$

This method is parameter free and thus its smoothing factor depends on the problem only. SGS is implemented by solving two equation systems with triangular system matrices.

s-ARK	s-ERK	GMRES-k	SGS	ILS
$s+1$	s+1	$k+2$	3.5	5

Table 3: Computational effort of multigrid method in terms of function evaluations, respectively MatVecs.

4.4 Implicit line smoothers

For compressible viscous flows, implicit point smoothers and implicit line smoothers (ILS) are sometimes used. Both are in effect variants of block Jacobi smoothers, meaning that the following iteration is used

$$\mathbf{u}^{k+1} = (\mathbf{I} - \mathbf{D}^{-1}\mathbf{A})\mathbf{u}^k + \mathbf{D}^{-1}\mathbf{b},$$

where $\mathbf{D}$ is a certain block diagonal matrix. For the point implicit smoother, the small blocks of the size of the number of equations in the PDE system (four for 2D Navier-Stokes or Euler, 5 in 3D). Since the convection diffusion equation is scalar, this reduces to the normal Jacobi smoother, which is not a good method. Therefore, we only consider implicit line smoothers here.

These are used typically in boundary layers with lines going a finite length in a direction normal to the boundary. Then the block is defined from the unknowns in that line and has the size as the number of equations corresponding to these unknowns. Here, we will use in the 2D case lines going in x-direction through the the whole domain, meaning that there is exactly one line for each discrete y value and their length is the number of unknowns in x-direction. We denote this smoother as ILS.

4.5 Computational effort

The convergence rate of a multigrid method with a specific smoother is important, but it needs to be evaluated with regards to the computational effort needed. In the nonlinear case, the cost should be measured by the number of evaluations of the right hand side, involving computation of numerical fluxes and so on. In the linear case, this corresponds to matrix vector products (MatVecs). For an s-step ERK or a k-step GMRES smoother, the number of these is easy to determine, namely s and $k+1$. For an ARK method, it should be noted that while we actually increase computational effort in the linear case, this is not so in the nonlinear case, where the computation of viscous and inviscid fluxes is essentially independent. Thus, we consider the ARK smoother to have a computational effort proportional to s MatVecs as well.

With regards to SGS, there is one MatVec, one multiplication with a diagonal and two solves with tridiagonal equation systems. The latter take together slightly more than on MatVec, which is why we use 2.5 MatVecs as a measure for SGS. As for ILS, this requires solving a number of equation systems of smaller dimension. This is hard to predict, which is why we compared the CPU time for the use of that to one MatVec, arriving at a factor of four. Added to cost of a smoother in the present context is the cost of the MatVec needed for the computation of the residual on the fine grid.

4.6 Steady versus unsteady

We now consider the case of a parameter dependent smoother with parameter ω for the linear equation system $\mathbf{Ax} = \mathbf{b}$. This could be a damping parameter or a pseudo time step size, respectively CFL number in pseudotime. The iteration matrix is then of the form $\mathbf{M}^s(\omega\mathbf{A})$. Assume that we have determined an optimal coefficient ω^* for a steady state problem. For an

c	ImpE	GMRES-2
1	0.4903	0.3772
10	0.3798	0.3772
100	0.3774	0.3772
1000	0.3773	0.3772

Table 4: Measured convergence rate of the multigrid method for 1D convection diffusion using 2-step GMRES directly vs. 2-step GMRES inside implicit Euler smoother, $N = 128$.

c	ImpE	GMRES-2
1	0.5835	0.4870
10	0.4937	0.4870
100	0.4876	0.4870
1000	0.4871	0.4870

Table 5: Measured convergence rate of the multigrid method for 2D convection diffusion using 2-step GMRES directly vs. 2-step GMRES inside implicit Euler smoother, N=128.

unsteady problem, we obtain the system matrix $\mathbf{I} - \Delta t\mathbf{A}$. Under the assumption of a large outer CFL number, the identity matrix can be neglected and the matrix to consider is $\Delta t\mathbf{A}$ and the smoother is of the form $\mathbf{M}^S(\omega\Delta t\mathbf{A})$. To obtain the optimal coefficient for this case, we thus have to divide ω^* by Δt. This gives us a smoothing factor that is approximately the same as for the steady state and means that it is sufficient to design the smoother based on the steady state problem only.

If we instead of the pseudotime consider a CFL number, then from the condition $c_S^* = c_U^*$ on the optimal parameter, we obtain

$$c_U^* = c\Delta t_U^* \Rightarrow \Delta t_U^* = c_U^*/c = \Delta t_S^*/\Delta t.$$

This implies that in the unsteady case, the larger the outer time step, the smaller the pseudo time step has to be. However, this does not decrease the smoothing factor, as seen above.

5 NUMERICAL RESULTS

5.1 Implicit pseudo time iterations

We first test the difference between using an implicit pseudo time iteration with a certain smoother compared to using that smoother directly on the example of GMRES with 2 steps. The first test case is the one dimensional convection diffusion equation with $a = 1$, $b = 0.001$, $\Delta t = 0.5$. We now vary the pseudo time CFL number and compare the convergence rates, as seen in Table 4. The second test case is the two dimensional convection diffusion equation with convection speed $\tilde{a} = 1$ and angle $\gamma = \pi/4$, $b = 0.001$, $\Delta t = 0.5$. Again, the pseudo time CFL number c is varied and the results can be seen in table 5.

As can be seen, for about value for c of 1000, there is no essentially no difference between convergence rates, as predicted by the theory. Furthermore, the convergence rate of the method with pseudotime iteration is decidedly smaller for small c. Thus, we do not see a benefit in this class of methods.

5.2 Variants of GMRES

We now test the results of section 4.2. The first test case is the one-dimensional convection diffusion equation with $a = 1$, $b = 0.001$, $\Delta t = 0.5$. As a smoother k-step GMRES is used.

Δx	1	2	3	4
1/64	0.6643	0.3853	0.2955	0.1801
1/128	0.6979	0.427	0.3345	0.2120
1/256	0.7484	0.4522	0.3645	0.2322

Table 6: Measured convergence rate of the multigrid method for 1D convection diffusion when using k-step GMRES as a smoother.

Δx	1	2	3	4
1/64	0.8725	0.7879	0.7836	0.7097
1/128	0.8870	0.8084	0.8033	0.7333
1/256	0.9079	0.8200	0.8172	0.7467

Table 7: The $k+2$-nd root of the measured convergence rate of the multigrid method for 1D convection diffusion when using k-step GMRES as a smoother.

In Table 6, the convergence rates observed in practice of the multigrid method can be seen for different Δx and different k. As can be seen, increasing the dimension of the Krylov subspace used, respectively the degree of the polynomial, decreases the convergence rate. Furthermore, the convergence rate increases with a finer mesh width, but not terribly so.

To get a better picture of the efficiency of the smoother, we now look at the $k+2$-nd root of the convergence rate measured. As can be seen in Table 7, increasing the number of smoothing steps shows an improving trend. It is also noticeable that an even number of steps perform slightly better than an odd number.

As a second test case, we use the two-dimensional convection diffusion equation with $\tilde{a} = 1$, $\gamma = \pi/4$, $b = 0.001$, $\Delta t = 0.5$. As a smoother k-step GMRES is used. In Table 8, the convergence rates observed in practice of the multigrid method can be seen for different Δx. As can be seen, increasing the dimension of the Krylov subspace used, respectively the degree of the polynomial, decreases the convergence rate. Furthermore, the convergence rate increases with a finer mesh width, but not terribly so. Again, the $k+2$-nd root of the convergence rate measured is shown in Table 9. Now the root is approximately constant on a particular mesh, but goes up from $k = 3$ to $k = 4$. Thus, 2- and 3-step GMRES perform about the same. Since the cost per iteration in GMRES increases with due to the increasing number of scalar products, we consider 2-step GMRES to be the most promising method.

5.3 Comparison of smoothers

Now all the smoothers under consideration are compared. The first test case is the two-dimensional convection diffusion equation with $\tilde{a} = 1$, $\gamma = \pi/4$, $b = 0.001$, $\Delta t = 0.5$ and a two-level multigrid cycle. Table 10 lists the convergence rates obtained with all smoothers. We test the effect of using the smoother vs. a direct solver on the coarsest multigrid level. In Table 11, the m-th root of the convergence rates is listed; the values of m are given in Table 3. Using

Δx	1	2	3	4
1/32	0.5192	0.3853	0.2838	0.2726
1/64	0.5913	0.4506	0.3576	0.3531
1/128	0.6326	0.4870	0.4076	0.3975

Table 8: Measured convergence rate of the multigrid method for 2D convection diffusion when using k-step GMRES as a smoother.

Δx	1	2	3	4
1/32	0.8037	0.7879	0.7773	0.8052
1/64	0.8393	0.8193	0.8141	0.8407
1/128	0.8584	0.8354	0.8357	0.8575

Table 9: The $k+2$-nd root of the measured convergence rate of the multigrid method for 2D convection diffusion when using k-step GMRES as a smoother.

	direct			smoother		
Scheme	1/32	1/64	1/128	1/32	1/64	1/128
ARK4	0.5330	0.5430	0.5397	0.8034	0.8977	0.9471
ARK5	0.6473	0.6412	0.6266	0.9169	0.9580	0.9787
RK2	0.4312	0.4646	0.4789	0.7049	0.8452	0.9198
RK4	0.3505	0.4130	0.4428	0.4749	0.7129	0.8479
ImpE	0.2654	0.3541	0.4165	0.3539	0.5675	0.7608
GMRES	0.2658	0.3541	0.4158	0.3576	0.5679	0.7583
SGS	0.0875	0.2013	0.3489	0.1019	0.2854	0.5998
ILS	0.7814	0.8334	0.8494	0.7344	0.8496	0.9246

Table 10: The measured convergence rate of the 2-level multigrid method for 2D convection diffusion with $\gamma = \pi/4$ when using smoother vs. direct solve on coarsest multigrid level.

a smoother is less effective than a direct solver as a coarse-grid method. The best smoother in terms of computational effort is SGS. Explicit multistep smoothers (RK and ARK) achieve relatively poor convergence rates. The implicit line smoother achieves the slowest convergence because the lines are not aligned with the convection direction.

In the next test, the number of multigrid levels is increased to the maximum possible for each N: i.e. $\log_2(N)$. Tables 12 and 13 list the convergence rates and their m-th roots. The choice of smoother/direct solver on the coarsest grid is insignificant. Overall, convergence rates are improved by the increased number of multigrid levels. As in the two-level test, SGS is the fastest-converging smoother. However, the m-th root of convergence rate increases with N and on the finest grid RK2 has comparable values. On still finer grids SGS may not be the clear winner.

As a final test the convection angle is set to $\gamma = \pi/2$ so that the lines of ILS are aligned with the flow. The number of multigrid cycles is kept at $\log_2(N)$. Tables 14 and 15 list the convergence rates and their m-th roots respectively. Most notable is the ILS convergence rate is greatly improved at lower N. The other schemes follow the opposite trend, their convergence rates getting larger at $\gamma = \pi/2$.

6 CONCLUSIONS

We considered different smoothers in an agglomeration multigrid method for a finite volume discretization of the linear convection diffusion equation. For these, we considered different smoothers, namely explicit and additive Runge-Kutta methods, GMRES with a fixed number of steps, SGS and implicit line smoothers. We related GMRES to optimized explicit RK methods in that GMRES automatically chooses a polynomial that optimizes residual reduction over all eigenvalues, not just the ones relevant for smoothing. In the case of GMRES, we analyze and demonstrate by numerical experiments that implicit pseudo time iterations do not offer a benefit compared to using the inner method directly as a smoother. Implicit line smoothing turns out to

	direct			smoother		
Scheme	1/32	1/64	1/128	1/32	1/64	1/128
ARK4	0.8817	0.8850	0.8840	0.9572	0.9786	0.9892
ARK5	0.9301	0.9286	0.9251	0.9856	0.9929	0.9964
RK2	0.7555	0.7745	0.7824	0.8900	0.9455	0.9725
RK4	0.8109	0.8379	0.8497	0.8616	0.9346	0.9675
ImpE	0.8016	0.8411	0.8642	0.8410	0.9099	0.9555
GMRES	0.8018	0.8411	0.8639	0.8425	0.9100	0.9549
SGS	0.4985	0.6325	0.7402	0.5208	0.6989	0.8641
ILS	0.9519	0.9642	0.9679	0.9401	0.9679	0.9845

Table 11: The m-th root of the measured convergence rate of the 2-level multigrid method for 2D convection diffusion with $\gamma = \pi/4$ when using smoother vs. direct solve on coarsest multigrid level. Table 3 lists the value of m for each scheme.

	direct			smoother		
Scheme	1/32	1/64	1/128	1/32	1/64	1/128
ARK4	0.5562	0.5918	0.6143	0.5624	0.5907	0.6105
ARK5	0.6780	0.7035	0.7295	0.6876	0.7101	0.7364
RK2	0.4130	0.4275	0.4497	0.4144	0.4332	0.4558
RK4	0.2933	0.3550	0.3868	0.2931	0.3539	0.3867
ImpE	0.2716	0.3527	0.3969	0.2717	0.3527	0.3969
GMRES	0.2726	0.3533	0.3974	0.2726	0.3533	0.3974
SGS	0.0817	0.1938	0.3358	0.0818	0.1938	0.3358
ILS	0.9019	1.1877	1.4010	0.9787	1.2581	1.4712

Table 12: The measured convergence rate of the $\log_2(N)$-level multigrid method for 2D convection diffusion with $\gamma = \pi/4$ when using smoother vs. direct solve on coarsest multigrid level.

	direct			smoother		
Scheme	1/32	1/64	1/128	1/32	1/64	1/128
ARK4	0.8893	0.9004	0.9071	0.8913	0.9001	0.9060
ARK5	0.9373	0.9431	0.9488	0.9395	0.9445	0.9503
RK2	0.7447	0.7533	0.7662	0.7455	0.7566	0.7696
RK4	0.7824	0.8129	0.8270	0.7824	0.8124	0.8270
ImpE	0.8048	0.8406	0.8573	0.8048	0.8406	0.8573
GMRES	0.8052	0.8408	0.8574	0.8052	0.8408	0.8574
SGS	0.4889	0.6258	0.7321	0.4890	0.6258	0.7321
ILS	0.9796	1.0350	1.0698	0.9957	1.0470	1.0803

Table 13: The m-th root of the measured convergence rate of the $\log_2(N)$-level multigrid method for 2D convection diffusion with $\gamma = \pi/4$ when using smoother vs. direct solve on coarsest multigrid level.

	direct			smoother		
Scheme	1/32	1/64	1/128	1/32	1/64	1/128
ARK4	0.8964	0.8893	0.8539	0.8966	0.8897	0.8538
ARK5	0.9440	0.9378	0.9158	0.9443	0.9388	0.9179
RK2	0.8506	0.8469	0.7996	0.8506	0.8469	0.7995
RK4	0.7352	0.7427	0.6857	0.7352	0.7427	0.6857
ImpE	0.4314	0.5400	0.5652	0.4314	0.5400	0.5652
GMRES	0.5183	0.5719	0.5634	0.5183	0.5719	0.5634
SGS	0.1156	0.2422	0.4414	0.1156	0.2422	0.4414
ILS	0.1913	0.4768	0.7437	0.1913	0.4768	0.7437

Table 14: The measured convergence rate of the $\log_2(N)$-level multigrid method for 2D convection diffusion with $\gamma = \pi/2$ when using smoother vs. direct solve on coarsest multigrid level.

	direct			smoother		
Scheme	1/32	1/64	1/128	1/32	1/64	1/128
ARK4	0.9784	0.9768	0.9869	0.9784	0.9769	0.9869
ARK5	0.9904	0.9893	0.9855	0.9905	0.9895	0.9858
RK2	0.9475	0.9461	0.9282	0.9475	0.9461	0.9281
RK4	0.9403	0.9422	0.9273	0.9403	0.9422	0.9273
ImpE	0.8693	0.9024	0.9093	0.8693	0.9024	0.9093
GMRES	0.8963	0.9111	0.9088	0.8963	0.9111	0.9088
SGS	0.5398	0.6668	0.7917	0.5398	0.6668	0.7917
ILS	0.7183	0.8623	0.9425	0.7183	0.8623	0.9425

Table 15: The m-th root of the measured convergence rate of the $\log_2(N)$-level multigrid method for 2D convection diffusion with $\gamma = \pi/2$ when using smoother vs. direct solve on coarsest multigrid level.

be a specific block Jacobi method.

When looking at computational effort measured in terms of matrix vector products, 2-step GMRES and SGS perform best. Since the latter is not parallel, 2-step GMRES is a very interesting candidate for further study.

REFERENCES

[1] P. BIRKEN, *Optimizing Runge-Kutta smoothers for unsteady flow problems*, ETNA, 39 (2012), pp. 298–312.

[2] P. BIRKEN, *Solving nonlinear systems inside implicit time integration schemes for unsteady viscous flows*, in Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws, R. Ansorge, H. Bijl, A. Meister, and T. Sonar, eds., Springer, 2013, pp. 57–71.

[3] D. CAUGHEY AND A. JAMESON, *How many steps are required to solve the Euler equations of steady compressible flow: In search of a fast solution algorithm*, AIAA Paper 2001-2673, (2001).

[4] H. C. ELMAN, O. G. ERNST, AND D. P. O'LEARY, *A Multigrid Method Enhanced by Krylov Subspace Iteration for Discrete Helmholtz Equations*, SIAM J. Sci. Comput., 23 (2001), pp. 1291–1315.

[5] A. JAMESON, *Transonic flow calculations for aircraft*, in Numerical Methods in Fluid Dynamics, F. Brezzi, ed., Lecture Notes in Mathematics, Springer, 1985, pp. 156–242.

[6] A. JAMESON, W. SCHMIDT, AND E. TURKEL, *Numerical Solution of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time-Stepping Schemes*, in AIAA Paper 1981-1259, 1981.

[7] S. LANGER, *Investigation and application of point implicit Runge-Kutta methods to inviscid flow problems*, Int. J. Num. Meth. Fluids, 69 (2012), pp. 332–352.

[8] S. LANGER, *Application of a line implicit method to fully coupled system of equations for turbulent flow problems*, Int. J. CFD, 27 (2013), pp. 131–150.

[9] S. LANGER AND D. LI, *Application of point implicit Runge-Kutta methods to inviscid and laminar flow problems using AUSM and AUSM + upwinding*, International Journal of Computational Fluid Dynamics, 25 (2011), pp. 255–269.

[10] S. LANGER, A. SCHWÖPPE, AND N. KROLL, *Investigation and Comparison of Implicit Smoothers Applied in Agglomeration Multigrid*, AIAA Journal, 53 (2015), pp. 2080–2096.

[11] D. MAVRIPLIS, *An assessment of linear versus nonlinear multigrid methods for unstructured mesh solvers*, Journal of Computational Physics, 175 (2002), pp. 302–325.

[12] D. J. MAVRIPLIS, *Directional Agglomeration Multigrid Techniques for High-Reynolds Number Viscous Flows*, Tech. Rep. 98-7, ICASE, 1998.

[13] B. VAN LEER, C.-H. TAI, AND K. G. POWELL, *Design of Optimally Smoothing Multi-Stage Schemes for the Euler Equations*, in AIAA 89-1933-CP, 1989, pp. 40–59.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

HIGH-ORDER LINEARLY IMPLICIT TWO-STEP PEER METHODS FOR THE DISCONTINUOUS GALERKIN SOLUTION OF THE INCOMPRESSIBLE RANS EQUATIONS

Francesco Carlo Massa[2], Gianmaria Noventa[1], Francesco Bassi[2], Alessandro Colombo[2], Antonio Ghidoni[1] and Marco Lorini[1]

[1]Dipartimento di Ingegneria Meccanica e Inustriale, Universita degli Studi di Brescia
via Branze 38, 25123 Brescia, Italy
e-mail: {antonio.ghidoni, m.lorini006, g.noventa002}@unibs.it

[2] Dipartimento di Ingegneria Industriale, Universita degli studi di Bergamo
Viale Marconi 5, 24044 Dalmine (BG), Italy
e-mail: {francescocarlo.massa, francesco.bassi, alessandro.colombo}@unibg.it

Keywords: Discontinuous Galerkin discretization, linearly implicit Rosenbrock-type two-step peer methods, incompressible flows, unsteady simulations, NS, RANS, time-step adaptation strategy

Abstract. *In recent years the increasing attention to high-order Finite Volume (FV), Finite Element (FE) and spectral methods and the growth of computing power promote the development of high-order temporal schemes to perform robust, accurate and efficient unsteady long-time simulations. In this context, some features of the Discontinuous Galerkin finite element (DG) methods, e.g. compactness and flexibility, can be advantageous both for explicit and implicit time integration approaches. Explicit schemes can achieve very high accuracy, but are limited by time-step restrictions, while implicit schemes, even if memory consuming, can overcome time-step limitations, thus improving the time integration efficiency. During last decades several high order implicit temporal schemes have been developed, and some of them have been successfully coupled with DG methods. However these schemes can show the order reduction if applied to very stiff problems or problems with time-dependent boundary conditions. To overcome these limitations, high-order linearly implicit two-step peer methods have been proposed and successfully applied to the numerical solution of differential-algebraic equations. The aim of the present work is to implement high-order two-step peer methods in a DG code and assess their performance for the unsteady solution of the incompressible Navier-Stokes (INS) and Reynolds Averaged Navier-Stokes equations (RANS) equations.*

1 Introduction

In recent years, due to the increasing computational power and accuracy requirement, Discontinuous Galerkin (DG) method has become one of the most promising approaches to high-fidelity fluid dynamic computations in many technical areas, such as aeronautics, aeroacoustics and turbomachinery [17, 3, 2]. DG method proved to be very well suited for the Direct Numerical Simulation (DNS) [16, 15, 32] and the Large Eddy Simulation (LES) [31] of turbulent flows. Recently, the DG method has also been applied by Bassi *et al.* [3] to hybrid RANS-LES approaches in an attempt to overcome the poor predictive capabilities of the Reynolds-averaged Navier-Stokes (RANS) equations with first-moment closures in case of particular flow conditions, *e.g.*, massively separated flows.

These simulations require robust, accurate and efficient long time integration of unsteady flows, characterized by a wide range of temporal scales. Explicit schemes can achieve a very high accuracy [18, 19, 22] but are limited by time-step restriction. Otherwise, implicit schemes, although memory consuming due to the need of the Jacobian matrix, can be *A*-stable and *L*-stable even for high order of accuracy. Most popular implicit schemes are Backward Differentiation Formulae (BDF) [9], which are only *A*-stable up to the second-order, and their low accuracy is not well suited to match the spatial accuracy of DG methods. In the attempt to couple a high-order discretization both in space and time, several temporal schemes have been used to advance in time the DG space discretized equations [6, 26, 25, 5].

Among these schemes, Rosenbrock-type Runge-Kutta schemes [5] have received particular attention. In fact, being linearly implicit, they solve only one linear system per stage, showing a greater computational efficiency if compared to non linear methods [26, 25], which require to solve several non-linear systems at each time-step. Unfortunately, these Rosenbrock one-step methods can suffer from order reduction when they are applied to stiff problems. Additional order conditions can be derived to avoid the order reduction, but methods with higher order of convergence satisfying these conditions are difficult to find. For this reason, the class of the linearly implicit two-step peer methods [20] has been proposed for the numerical solution of time dependent Differential Algebric (DAE) and Partial Differential (PDE) equations. All solutions computed by Peer methods within a time-step have the same order of accuracy (this feature explains the attribute peer), and this is in contrast to one-step methods, where stage values have lower order than the final solution. Peer methods up to order eight are available in literature, and are characterized by good stability properties, *i.e.*, optimal zero-stability and $L\,(\alpha)$-stability [27], strong damping properties at infinity without further restrictions, and robustness with respect to stepsize changes. As all multi-step schemes, Peer methods are non self-starting and thus need a starting procedure in order to obtain required initial solutions. They are also multi-stage schemes and for each stage they require the solution of one linear system making them very attractive for practical computations. Since all stage solutions have the same accuracy and stability properties a continuous output of high-order is available with no extra-cost.

The aim of the present paper is to describe the implementation of high-order linearly implicit two-step peer methods in the DG MIGALE code for the integration in time of incompressible Navier-Stokes and RANS equations closed by the $k-\tilde{\omega}$ turbulence model [8]. Linearly implicit two-step peer methods up to sixth order of accuracy are assessed on two incompressible unsteady test cases: (i) the laminar travelling waves solution on a doubly periodic unit square domain and (ii) the turbulent flow around a circular cylinder for a Reynolds number $Re = 5 \times 10^4$. The travelling waves test case was used to investigate (i) the Peer methods order of accuracy, (ii) different types of starting procedure and their accuracy and (iii) the computational efficiency

of the schemes to obtain a given accuracy with the adaptive time-step strategy. Also the circular cylinder test case has been used to investigate the computational efficiency of Peer methods with the adaptive time-step strategy. A comparative assessment in terms of accuracy and performance with the traditional fourth order/six stages ESDIRK scheme [13] and third order/three stages (ROS3PL) [23] and fourth order/six stages (RODASP) [30] linearly implicit one-step Rosenbrock schemes has been done.

The paper is organized as follow: Section 2 describes the DG space discretization method for the Incompressible Navier-Stokes, RANS and turbulence model equations, while Section 3 introduces the class of the linearly implicit two-step peer methods and the adaptive time-step strategy. Numerical results are shown in Section 4, and the final Section 5 summarizes the results and presents the conclusion.

2 DG space discretization

Incompressible Navier-Stokes, RANS and turbulence model equations can be written in compact form as

$$\frac{\partial \widehat{\mathbf{q}}}{\partial t} + \nabla \cdot \mathbf{F}_c(\mathbf{q}) + \nabla \cdot \mathbf{F}_v(\mathbf{q}, \nabla \mathbf{q}) + \mathbf{s}(\mathbf{q}, \nabla \mathbf{q}) = \mathbf{0}, \tag{1}$$

where $\mathbf{q} \in \mathbb{R}^m$ denotes the vector of the m primitive variables $(p, u_i, k, \omega)^T$ (for $i = 1, d$), $\mathbf{s} \in \mathbb{R}^m$ the source term, d the space dimension. $\mathbf{F}_c, \mathbf{F}_v \in \mathbb{R}^m \otimes \mathbb{R}^d$ denote the inviscid and viscous flux functions, while $\widehat{\mathbf{q}} \in \mathbb{R}^m$ is defined as $(0, u_i, k, \omega)^T$.

To discretize the governing equations in space the system (1) is firstly multiplied by an arbitrary smooth test function $\mathbf{v} = \{v_1, \ldots, v_m\}$ and then integrated by parts, thus obtaining its weak form. The solution $\mathbf{q}$ and the test function $\mathbf{v}$ are then replaced with a finite element approximation $\mathbf{q}_h$ and a discrete test function $\mathbf{v}_h$ both belonging to $\mathbf{V}_h \stackrel{\text{def}}{=} [\mathbb{P}_d^k(\mathcal{T}_h)]^m$, where

$$\mathbb{P}_d^k(\mathcal{T}_h) \stackrel{\text{def}}{=} \left\{ v_h \in L^2(\Omega) \,|\, v_{h|K} \in \mathbb{P}_d^k(K),\ \forall K \in \mathcal{T}_h \right\} \tag{2}$$

is the discrete polynomial space in physical coordinates. $\mathbb{P}_d^k(K)$ denotes the restriction of the polynomial functions of d variables and total degree $\leq k$ to the element K belonging to the triangulation $\mathcal{T}_h = \{K\}$, consisting of a set of non-overlapping arbitrarily shaped and possibly curved elements, built on an approximation Ω_h of the domain Ω. We also define as $\mathcal{F}_h$ the set of the mesh faces $\mathcal{F}_h \stackrel{\text{def}}{=} \mathcal{F}_h^i \cup \mathcal{F}_h^b$, where $\mathcal{F}_h^b$ collects the faces located on the boundary of Ω_h. For any internal face $F \in \mathcal{F}_h^i$ there exist two elements $K^+, K^- \in \mathcal{T}_h$ such that $F \in \partial K^+ \cap \partial K^-$, where ∂ denotes cell boundaries. Moreover, for all $F \in \mathcal{F}_h^b$, $\mathbf{n}_F$ is the unit outward normal to Ω_h, whereas, for all $F \in \mathcal{F}_h^i$, $\mathbf{n}_F^-$ and $\mathbf{n}_F^+$ are the unit outward normals pointing to K^+ and K^-, respectively. To deal with discontinuous functions over the internal faces $F \in \mathcal{F}_h^i$ we introduce the jump $[\![\cdot]\!]$ and average $\{\cdot\}$ trace operators, that is

$$[\![v_h]\!] \stackrel{\text{def}}{=} v_{h|K^+}\mathbf{n}_F^+ + v_{h|K^-}\mathbf{n}_F^-, \qquad \{v_h\} \stackrel{\text{def}}{=} \frac{v_{h|K^+} + v_{h|K^-}}{2}. \tag{3}$$

When applied to vector functions these operators act componentwise.

Following the approach presented in [4], for each equation of the system, and without loss of generality, we choose the set of test and shape functions in any element K coincident with the set $\{\phi\}$ of N_{dof}^K orthogonal and hierarchical basis functions in that element. Such basis is

built by means of the modified Gram-Schmidt (MGS) algorithm, starting from a set of monomials defined over each elementary space $\mathbb{P}_d^k(K)$ in a reference frame relocated in the element barycenter and aligned with the principal axes of inertia of K.

Each component $q_{h,j}$, $j = 1, \ldots, m$, of the numerical solution $\mathbf{q}_h \in \mathbf{V}_h$ can be expressed, in terms of the elements of the global vector $\mathbf{Q}$ of unknown degrees of freedom, as $q_{h,j} = \phi_l Q_{j,l}$, $l = 1, \ldots, N_{dof}^K$, $\forall K \in \mathcal{T}_h$.

Accounting for these aspects, the DG discretization of the RANS and turbulence model equations consists in seeking, for $j = 1, \ldots, m$, the elements of $\mathbf{Q}$ such that

$$\sum_{K\in\mathcal{T}_h} \int_K \phi_i \phi_l \frac{dQ_{k,l}}{dt} d\mathbf{x} - \sum_{K\in\mathcal{T}_h} \int_K \frac{\partial \phi_i}{\partial x_n} F_{j,n}\left(\mathbf{q}_h, \nabla_h \mathbf{q}_h + \mathbf{r}\left([\![\mathbf{q}_h]\!]\right)\right) d\mathbf{x} + \sum_{F\in\mathcal{F}_h} \int_F [\![\phi_i]\!]_n \widehat{F}_{j,n}\left(\mathbf{q}_h^{\pm}, \left(\nabla_h \mathbf{q}_h + \eta_F \mathbf{r}_F\left([\![\mathbf{q}_h]\!]\right)\right)^{\pm}\right) d\sigma + \sum_{K\in\mathcal{T}_h} \int_K \phi_i s_j\left(\mathbf{q}_h, \nabla_h \mathbf{q}_h + \mathbf{r}\left([\![\mathbf{q}_h]\!]\right)\right) d\mathbf{x} = 0, \quad (4)$$

for $i = 1, \ldots, N_{dof}^K$ and where repeated indices imply summation over the ranges $k = 1, \ldots, m$, $l = 1, \ldots, N_{dof}^K$ and $n = 1, \ldots, d$.

In Eq. (4) $\mathbf{F}$ denotes the sum of the convective and viscous flux functions, and $\widehat{\mathbf{F}}$ the sum of their numerical counterparts. For the former the flux computation is based on the exact solution of the Riemann problem for the artificial compressibility perturbation of the locally 1D inviscid Euler equations, as suggested in [7, 8], while for the latter the BR2 scheme is employed, proposed in [9] and theoretically analyzed in [11, 1].

3 Time discretization

The discrete problem corresponding to Eq. (4) can be written as:

$$\widehat{\mathbf{M}} \frac{d\mathbf{Q}}{dt} + \mathbf{R}\left(\mathbf{Q}\right) = \mathbf{0}, \quad (5)$$

where $\mathbf{Q}$ is the global vector of unknown degrees of freedom, $\mathbf{R}$ the residuals vector, and $\widehat{\mathbf{M}}$ the global block diagonal mass matrix. Thanks to the use of orthonormal base functions defined in physical space, the matrix $\widehat{\mathbf{M}}$ is simply the identity matrix with null entries corresponding to the pressure degrees of freedom due to the lack of pressure time derivatives.

In this work the Eq. (5) is advanced in time in an implicit sense, by using the high-order linearly implicit two-step peer methods.

3.1 Linearly implicit two-step peer methods

The s-stage linearly implicit two-step peer method computes solution approximations $\mathbf{Q}_i^n$, $i = 1, \ldots, s$, of the Eq. (5) at times $t_i^n = t^n + c_i \Delta t^n$ by means of s linear systems:

$$\left(\frac{\widehat{\mathbf{M}}}{\gamma \Delta t^n} + \mathbf{J}_s^{n-1}\right)\left(\mathbf{Q}_i^n - \mathbf{Q}_{0i}^{\ n}\right) = -\mathbf{R}_{0i}^{\ n} + \frac{\widehat{\mathbf{M}}}{\gamma \Delta t^n}\left(\mathbf{Y}_i - \mathbf{Q}_{0i}^{\ n}\right) \qquad i = 1, ..., s, \quad (6)$$

where

$$\mathbf{Y}_i = \sum_{j=1}^{i-1} \frac{a_{ij}}{\gamma}\left(\mathbf{Q}_j^n - \mathbf{Y}_j\right) + \sum_{j=1}^{s} u_{ij} \mathbf{Q}_j^{n-1}, \quad (7)$$

$$\mathbf{Q}_{0_i}^{\ n} = \sum_{j=1}^{i-1} \frac{a_{0ij}}{\gamma} \left(\mathbf{Q}_j^n - \mathbf{Y}_j \right) + \sum_{j=1}^{s} u_{0ij} \mathbf{Q}_j^{n-1}, \tag{8}$$

with

$$\mathbf{J}_s^{n-1} = \frac{\partial \mathbf{R}(\mathbf{Q}_s^{n-1})}{\partial \mathbf{Q}}, \tag{9}$$

At each step, an embedded solution can be computed as

$$\hat{\mathbf{Q}}_s^{\ n} = \sum_{i=1}^{s-1} \alpha_i \mathbf{Q}_i^n. \tag{10}$$

$\mathbf{A} = \{a_{ij}\}$, $\mathbf{U} = \{u_{ij}\}$, $\mathbf{A}_0 = \left\{a_{0ij}\right\}$ and $\mathbf{U}_0 = \left\{u_{0ij}\right\}$ are $s \times s$ matrices of real coefficients and α_i, c_i and γ are real parameters. $\mathbf{U}$ and $\mathbf{U}_0$ depend on the step size ratio $\sigma^n = \Delta t^n / \Delta t^{n-1}$ between current and previous steps in order to ensure accuracy and stability properties for variable step sizes.

The Jacobian matrix of the DG space discretization $\mathbf{J}_s^{n-1}$ is computed only once per step with the last stage solution $\mathbf{Q}_s^{n-1}$ of the previous step. Thanks to the peculiar treatment of the convective numerical flux, $\mathbf{J}_s^{n-1}$ has non null entries for pressure degrees of freedom; it follows that, despite the singularity of matrix $\widehat{\mathbf{M}}$, Eq. (6) can be solved with standard algorithms.

In order to compute step $n = 1$ of the two-step method described in Eq. (6), initial solutions $\mathbf{Q}_i^0$, $i = 1, \ldots, s$, at times $t_i^0 = t^{st} + c_i \Delta t^{st}$ must be available, where superscript st denotes the starting step. The initialization algorithm to compute $\mathbf{Q}_i^0$ will be described in Sec. 3.3.

3.2 Accuracy and stability properties

Matrices $\mathbf{A}$, $\mathbf{U}$, $\mathbf{A}_0$ and $\mathbf{U}_0$ and parameters γ and c_i uniquely define accuracy and stability properties of a s-stage Peer method. Consistency and zero stability conditions that ensure order $q = s - 1$ for all step size ratio $\sigma > 0$ and for all Peer stages are provided in [27, 20]; it follows that all quantities computed by Eqs. (6), (7) and (8) have the same accuracy and stability properties. Furthermore, these conditions are derived independently from the Jacobian matrix, which thus can be computed numerically, e.g. by means of finite differences. As proposed in Ref. [27], matrices $\mathbf{A}$, $\mathbf{U}$, $\mathbf{A}_0$ and $\mathbf{U}_0$ are functions of parameters c_i and γ, and coefficients c_i are chosen to be stretched Chebychev nodes in the interval $[-1, 1]$:

$$c_i = -\frac{\cos\left(\left(i - \frac{1}{2}\right)\frac{\pi}{s}\right)}{\cos\left(\frac{\pi}{2s}\right)}, \qquad i = 1, \ldots, s. \tag{11}$$

The γ parameter specifies the θ angle of $A\,(\theta)$- and $L\,(\theta)$-stability, and, as demonstrated in [27], can guarantees (if conveniently chosen) an order $q = s$ for a s-stage Peer method. with constant time-step. Finally, coefficients α_i, $i = 1, \ldots, s - 1$, are only function of nodes c_i,and are computed such that the order of the embedded solution is equal to $\hat{q} = s - 2$ (see Ref. [20]).

Peer methods implemented have three, four, five and six stages and in this work are referred as peer3A, peer4A, peer5A and peer6A. They are A-stable and L-stable, *i.e.* $\theta = 90$, and have order $q = s - 1$ and $q = s$ for variable and constant time-step, respectively. The γ values adopted for each scheme are reported in Tab. 1.

Table 1: List of γ values and corresponding θ angles of $A(\theta)$-stability and $L(\theta)$-stability for each Peer method used in this work.

	s	γ	θ
peer3A	3	$2.165162598341552E+00$	90
peer4A	4	$1.038881828680110E+00$	90
peer5A	5	$5.614731292097847E-01$	90
peer6A	6	$3.476356828221134E-01$	90

3.3 Starting procedure

Peer methods are multi-step and multi-stage time integration schemes. As all multi-step schemes, they are not self-starting, that means they need a set of starting solutions $\mathbf{Q}_i^0$ at times:

$$t_i^0 = t^{st} + c_i \Delta t^{st}, \qquad i = 1, \dots, s \tag{12}$$

where

$$\Delta t^{st} = t^1 - t^{st}, \tag{13}$$

in order to perform the first step $n = 1$.

Defining

$$c^- = \min_{i \in [1,s]} (c_i), \qquad t^- = t^{st} + c^- \Delta t^{st} \tag{14}$$

as the minimum value of nodes c_i and the smallest time point of the starting procedure, respectively, and imposing that t^- is equal to initial time t^0, it follows from Eqs. (12) and (13) that:

$$t_i^0 = t^0 + \left(c_i - c^-\right) \Delta t^{st}, \qquad i = 1, \dots, s, \tag{15}$$

$$t^1 = t^0 + \left(1 - c^-\right) \Delta t^{st}, \tag{16}$$

where Δt^{st} is the starting procedure time-step here imposed equal to the user-defined initial time-step of Peer methods,

$$\Delta t^{st} = \Delta t^1, \tag{17}$$

in order to ensure $\sigma^1 = 1$.

Starting solutions at corresponding times t_i^0 can be obtained by means of a one-step time integration scheme, *e.g.* Rosenbrock or ESDIRK scheme. It is mandatory that the accuracy of $\mathbf{Q}_i^0$ is proportional to the accuracy of the Peer method used for the simulation, otherwise a reduction of the accuracy order is introduced. Basing on the asymptotic behaviour of time integration errors, this request leads to:

$$T^{st} \left(\overline{\Delta t'}\right)^{q'} \propto \left(\overline{\Delta t}\right)^{q}, \tag{18}$$

where $\overline{\Delta t'}$ and $\overline{\Delta t}$ are the mean time-steps used by the one-step scheme of order q' and Peer method of order q, respectively and T^{st} is the starting procedure period that can be computed from Eq. (15) as:

$$T^{st} = \left(c^+ - c^-\right) \Delta t^{st}, \tag{19}$$

with $c^+ = \max_{i \in [1,s]} (c_i)$. Since $\overline{\Delta t} \propto \Delta t^1$, from Eq. (17) it follows that:

$$\overline{\Delta t'} = \psi \left(\Delta t^1\right)^{\frac{q-1}{q'}}, \tag{20}$$

where the constant value must be defined by the user. For simplicity we consider $\psi = 1$. After sorting the starting time points such that:

$$t_j^0 < t_{j+1}^0, \qquad j = 1, \ldots, s-1, \tag{21}$$

with in general $t_j^0 \neq t_i^0$, it can be found the number of steps,

$$N_j' = \text{int}\left[\frac{\left(t_{j+1}^0 - t_j^0\right)}{\left(\overline{\Delta t'}\right)}\right] + 1, \tag{22}$$

needed to integrate in time from t_j^0 to t_{j+1}^0 by means of the one-step scheme and, as a consequence, the corresponding fixed time-step:

$$\Delta t'_j = \frac{\left(t_{j+1}^0 - t_j^0\right)}{N_j'}. \tag{23}$$

Notice that $\Delta t'_j \leq \overline{\Delta t'}$. Finally, starting from $\mathbf{Q}_j^0$, the solution $\mathbf{Q}_{j+1}^0$ can be achieved after N_j' steps of dimension $\Delta t_j'$ performed with a one-step temporal scheme.

3.4 Time-step adaptation algorithm

The time-step control is an important feature to increase the efficiency and robustness of a time integration method. The adaptive technique is based on the local error estimator, which exploit the lower order embedded solution to approximate the local truncation error of the considered temporal scheme:

$$r^{n+1} = \frac{\|\mathbf{Q}^n - \hat{\mathbf{Q}^n}\|_2}{d}, \tag{24}$$

where d is the threshold tolerance, defined as:

$$d = ATOL + RTOL\|\mathbf{Q}^n\|_2. \tag{25}$$

with $ATOL$ and $RTOL$ user-defined absolute and relative tolerances, respectively. In this work is set $ATOL = RTOL = TOL$, that leads to:

$$d = TOL + TOL\|\mathbf{Q}^{n+1}\|_2. \tag{26}$$

Considering relations proposed by Söderlind et al., in [28, 29], the next time-step Δt^{n+1} can be computed as:

$$\Delta t^{n+1} = \min\left(\left(1 + 2\text{atan}\left[\left(\rho^n - 1\right)/2\right]\right)\Delta t^n, \Delta t_{max}\right), \tag{27}$$

where

$$\rho^n = \left(\frac{\Delta t^n}{\Delta t^{n-1}}\right)^{-\alpha_2}\left(\frac{1}{r^{n+1}}\right)^{\beta_1}\left(\frac{1}{r^n}\right)^{\beta_2}, \tag{28}$$

$\hat{q}\beta_1 = \hat{q}\beta_2 = \alpha_2 = 1/4$, Δt^n is the current time-step, Δt^{n-1} the previous time-step, $\hat{q}$ the order of the embedded scheme, r^n the previous local error estimator and Δt_{max} the user-defined maximum allowable time-step which can be defined as a fraction of the characteristic temporal scale of the considered flow problem (*e.g.* the vortex shedding period).

If $r_{n+1} > \delta$, the current time-step Δt^n is rejected and the iteration recomputed with Δt^{n+1}. $\delta > 1$ is a user-defined safety factor, here set to 1.5.

The time-step adaptation algorithm should operate in a tolerance proportional mode, *i.e.* a change of the user defined tolerance TOL by one order of magnitude should correspond to a change of the solution error by one order of magnitude. Furthermore, the different time integration schemes should deliver the same accuracy for the same tolerance setting. These features can be obtained with a proper scaling and calibration of the adaptation algorithm by using a modified input tolerance TOL' inside Eq. (26), where:

$$TOL' = \beta TOL_0^{(k-1)/k} TOL^{1/k}. \tag{29}$$

TOL_0 is the equivalence point determined during the calibration, β is a constant to have similar accuracies for the same TOL and different time integration schemes at the end of the simulations and k is the measured order of the curve $error - TOL'$ of the reference computations used for the calibration. The calibration was done on the travelling waves test case (see Section 4.1), where the analytical solution is known, and on the circular cylinder (see Section 4.2), computing a reference solution [29].

Finally, the tolerance of the iterative method used for the solution of linear systems (6) is set as

$$TOL_{GMRES} = \eta TOL', \tag{30}$$

where η is a safety factor. The choice of η will be numerically investigated in the Section 4.1.

4 Results

In this section the robustness and the accuracy of the proposed coupling between high-order DG method and high-order linearly implicit two-step peer schemes are demonstrated by means of the analysis of two incompressible test cases.

The first test case is the laminar solution of travelling dumped waves on a doubly periodic unit square domain. In the second test case, the unsteady turbulent flow around a circular cylinder is performed for $Re = 5 \times 10^4$ based on the circle diameter and on the freestream conditions [10]. The convergence order of Peer temporal schemes up to six stages, the correctness of the starting procedure and the influence of the system tolerance TOL_{System} have been assessed for the laminar test case. Afterwards, the efficiency of the proposed DG-Peer coupling was investigated for a turbulent test case by means of the adaptive time-step strategy and compared with the traditional one-step temporal schemes, *i.e.* the fourth order/six stages ESDIRK scheme [13] and the third order/three stages (ROS3PL) [23] and fourth order/six stages (RODASP) [30] linearly implicit one-step Rosenbrock schemes.

Computations have been run on a Linux cluster with 12 AMD 6220 CPUs (8 cores per CPU). Computing times t_{CPU} reported in this work are normalized values with respect to the *TauBenchmark* [12] value, $t_{TauBench}$, obtained on a full node of the cluster used for the CFD simulation[1]. The normalized computing time is measured as work units (wu) and is defined as wu=$(t_{CPU} n_{cores})/t_{TauBench}$, where t_{CPU} is the wall clock time and n_{cores} the numbers of cores.

[1] -n 250000 -s 10 define the reference TauBench workload for the hardware benchmark.

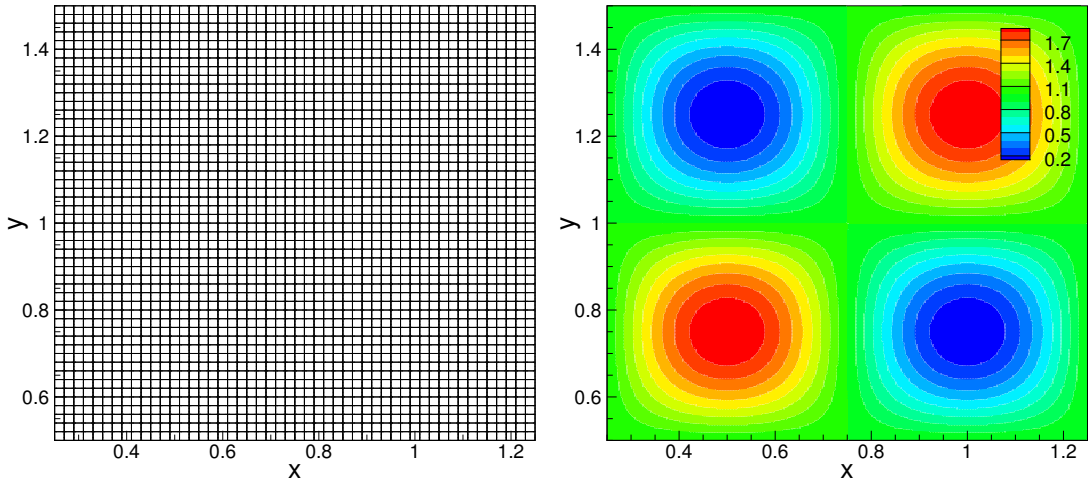


Figure 1: Travelling waves. Computational mesh (left), 2500 quadrilateral elements with linear edges, and velocity (x-component) contours after a simulation time $\Delta t_{sim} = T$ (right), $\mathbb{P}^6$ solution approximation

4.1 Travelling waves

The analytical solution of the laminar travelling dumbed waves test case is defined as:

$$u_1(x,y,t) = 1 + 2\cos(2\pi(x-t))\sin(2\pi(y-t))\,e^{-8\pi^2\nu t}, \tag{31}$$

$$u_2(x,y,t) = 1 - 2\sin(2\pi(x-t))\cos(2\pi(y-t))\,e^{-8\pi^2\nu t}, \tag{32}$$

$$p(x,y,t) = -(\cos(4\pi(x-t)) + \cos(4\pi(y-t)))\,e^{-16\pi^2\nu t}, \tag{33}$$

with $\nu = 1\mathrm{e}{-2}$. The domain is the doubly periodic unit square $[0.25, 1.25] \times [0.5, 1.5]$. A $\mathbb{P}^6$ solution approximation is used to keep the space discretization error below the time integration error (see Fig. 1 for the mesh and the velocity contours). The time integration period T is set equal to one non-dimensional convective time.

The convergence order of Peer schemes has been first verified, considering the following time-steps: $T/40$, $T/80$, $T/160$, $T/320$ and $T/400$ and a system tolerance $TOL_{GMRES} = 10^{-15}$. All schemes verify the expected convergence order $q = s$ for constant stepsize, as shown in Tab. 2. The errors of peer6A at smallest time-steps are affected by the machine error, thus preventing the verification of the convergence order. The initial solution $\mathbf{Q}_i^0$ is simply defined by the analytical solution of the problem at times $t^0_{i=1,\dots,s}$.

In general, the analytical solution is not available and starting values $\mathbf{Q}^0_{i=1,\dots,s}$ must be computed numerically by means of one-step time integration schemes. The starting procedure described in Section 3.3 is here tested adopting linearly-implicit Rosenbrock-type Runge-Kutta schemes ROS3PL [23] ($q' = 3$), RODASP [30] ($q' = 4$) and RODAS-ROD5_1 [24] ($q' = 5$). In particular, higher order RODASP and RODAS-ROD5_1 schemes are used only for peer5A and peer6A, respectively, while lower order ROS3PL is applied to all Peer methods.

Tab. 3 shows the L^2 norm of the pressure error ($\|err_p\|_2$) and the computational time obtained with numerical starting procedure. It can be seen that the Peer methods order of convergence is preserved. However, using a low order one-step scheme, *i.e.* ROS3PL, to initialize a high order Peer method is not efficient because the time-step needed by the one-step method to preserve the order of the Peer methods is too small.

The parameter η, used in Eq. (30), has been first chosen for each schemes with the adaptive time-step strategy, in order to set the system tolerance TOL_{GMRES} as a function of the adaptive

Table 2: Travelling waves. L^2 norm of the pressure ($\|err_p\|_2$) and velocity components ($\|err_u\|_2$, $\|err_v\|_2$) errors, and convergence order for Peer schemes with exact initialization, $\mathbb{P}^6$ solution approximation

	Δt	CFL	$\|err_p\|_2$	order	$\|err_u\|_2$	order	$\|err_v\|_2$	order
	T/40	7.96	1.78E-01	1.57	2.76E-01	1.57	2.76E-01	1.57
	T/80	3.98	3.74E-02	2.92	5.82E-02	2.92	5.82E-02	2.92
peer3A	T/160	1.99	4.84E-03	2.98	7.54E-03	2.98	7.54E-03	2.98
	T/320	1.00	6.13E-04	2.99	9.54E-04	2.99	9.54E-04	2.99
	T/400	0.80	3.15E-04	-	4.90E-04	-	4.90E-04	-
	T/40	7.96	8.53E-03	3.74	1.33E-02	3.73	1.33E-02	3.73
	T/80	3.98	6.01E-04	3.93	9.36E-04	3.93	9.36E-04	3.93
peer4A	T/160	1.99	3.88E-05	3.98	6.04E-05	3.98	6.04E-05	3.98
	T/320	1.00	2.45E-06	3.99	3.82E-06	3.99	3.82E-06	3.99
	T/400	0.80	1.01E-06	-	1.57E-06	-	1.57E-06	-
	T/40	7.96	2.75E-04	4.89	4.28E-04	4.89	4.28E-04	4.89
	T/80	3.98	9.07E-06	4.97	1.41E-05	4.97	1.41E-05	4.97
peer5A	T/160	1.99	2.87E-07	4.99	4.47E-07	4.99	4.47E-07	4.99
	T/320	1.00	9.02E-09	5.00	1.40E-08	5.00	1.40E-08	5.00
	T/400	0.80	2.96E-09	-	4.60E-09	-	4.60E-09	-
	T/40	7.96	7.66E-06	5.98	1.19E-05	5.98	1.19E-05	5.98
	T/80	3.98	1.20E-07	6.01	1.86E-07	6.01	1.86E-07	6.01
peer6A	T/160	1.99	1.87E-09	5.80	2.89E-09	6.03	2.89E-09	6.03
	T/320	1.00	1.78E-10	-	4.06E-11	-	4.06E-11	-
	T/400	0.80	1.76E-10	-	1.81E-11	-	1.81E-11	-

Table 3: Travelling waves. L^2 norm of the pressure ($\|err_p\|_2$) error, convergence order and computational time for Peer schemes with numerical starting procedure, $\mathbb{P}^6$ solution approximation

	Δt	CFL	$\|err_p\|_2$	order	$\|err_u\|_2$	order	$\|err_v\|_2$	order	CPU time [wu]
	T/40	7.96	1.78E-01	1.57	2.76E-01	1.57	2.76E-01	1.57	2.66E+03
	T/80	3.98	3.74E-02	2.92	5.82E-02	2.92	5.82E-02	2.92	3.02E+03
peer3A (ROS3PL)	T/160	1.99	4.84E-03	2.98	7.54E-03	2.98	7.54E-03	2.98	3.88E+03
	T/320	1.00	6.13E-04	2.99	9.54E-04	2.99	9.54E-04	2.99	5.58E+03
	T/400	0.80	3.15E-04	-	4.90E-04	-	4.90E-04	-	6.55E+03
	T/40	7.96	8.52E-03	3.74	1.33E-02	3.73	1.33E-02	3.73	1.78E+03
	T/80	3.98	6.01E-04	3.93	9.35E-04	3.93	9.35E-04	3.93	2.22E+03
peer4A (ROS3PL)	T/160	1.99	3.88E-05	3.98	6.04E-05	3.98	6.04E-05	3.98	3.28E+03
	T/320	1.00	2.45E-06	3.99	3.81E-06	3.99	3.81E-06	3.99	6.16E+03
	T/400	0.80	1.01E-06	-	1.57E-06	-	1.57E-06	-	7.15E+03
	T/40	7.96	2.75E-04	4.89	4.29E-04	4.89	4.29E-04	4.89	1.60E+03
	T/80	3.98	9.07E-06	4.97	1.41E-05	4.97	1.41E-05	4.97	2.14E+03
peer5A (RODASP)	T/160	1.99	2.87E-07	4.99	4.47E-07	4.99	4.47E-07	4.99	3.62E+03
	T/320	1.00	9.02E-09	5.00	1.40E-08	5.00	1.40E-08	5.00	6.03E+03
	T/400	0.80	2.96E-09	-	4.60E-09	-	4.60E-09	-	7.63E+03
	T/40	7.96	2.75E-04	4.89	4.29E-04	4.89	4.29E-04	4.89	1.54E+03
	T/80	3.98	9.08E-06	4.97	1.41E-05	4.97	1.41E-05	4.97	2.25E+03
peer5A (ROS3PL)	T/160	1.99	2.87E-07	4.99	4.48E-07	4.99	4.48E-07	4.99	3.89E+03
	T/320	1.00	9.03E-09	4.99	1.41E-08	5.00	1.41E-08	5.00	7.50E+03
	T/400	0.80	2.96E-09	-	4.61E-09	-	4.61E-09	-	7.78E+03
	T/40	7.96	7.66E-06	5.98	1.19E-05	5.98	1.19E-05	5.98	1.66E+03
	T/80	3.98	1.20E-07	6.01	1.86E-07	6.01	1.86E-07	6.01	2.24E+03
peer6A (ROD5_1)	T/160	1.99	1.87E-09	5.80	2.89E-09	6.03	2.89E-09	6.03	3.81E+03
	T/320	1.00	1.78E-10	-	4.06E-11	-	4.06E-11	-	6.16E+03
	T/400	0.80	1.79E-10	-	1.81E-11	-	1.81E-11	-	6.75E+03
	T/40	7.96	7.65E-06	5.99	1.19E-05	5.98	1.19E-05	5.98	2.32E+03
	T/80	3.98	1.20E-07	6.00	1.86E-07	6.00	1.86E-07	6.00	4.72E+03
peer6A (ROS3PL)	T/160	1.99	1.86E-09	5.79	2.89E-09	6.02	2.89E-09	6.02	9.03E+03
	T/320	1.00	1.79E-10	-	4.05E-11	-	4.05E-11	-	1.98E+04
	T/400	0.80	1.74E-10	-	1.80E-11	-	1.80E-11	-	2.42E+04

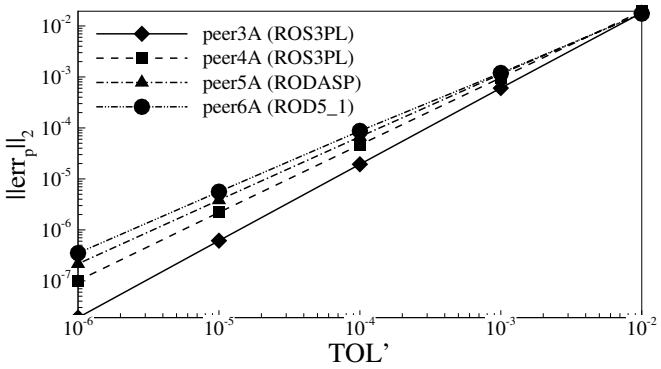


Figure 2: Travelling waves. L^2 norm of the pressure error ($\|err_p\|_2$) as a function of the adaptive tolerance TOL' for adaptive Peer schemes, $TOL_{GMRES} = 10^{-15}$ and $\mathbb{P}^6$ solution approximation

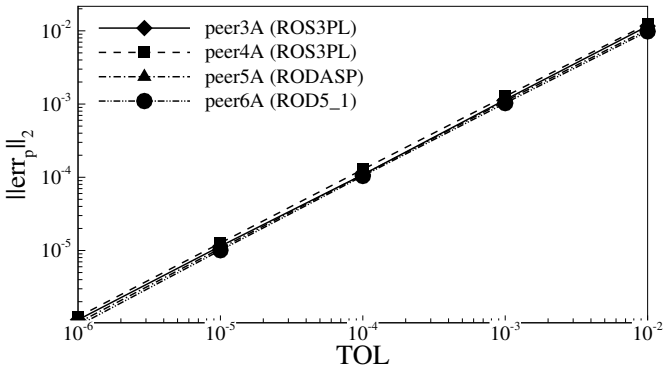


Figure 3: Travelling waves. L^2 norm of the pressure error ($\|err_p\|_2$) as a function of the user defined tolerance TOL for adaptive Peer schemes, $TOL_{GMRES} = \eta TOL$ and $\mathbb{P}^6$ solution approximation

tolerance TOL'. The results, summarized in Tab. 4, show the different errors with respect to the reference solution with $TOL_{GMRES} = 10^{-15}$. The pressure error starts to change only when $TOL_{GMRES} \geq 0.1\,TOL'$, that leads to set $\eta = 0.1$.

Fig. 2 shows the L^2 norm of the pressure error ($\|err_p\|_2$) as a function of the adaptive tolerance TOL'. Without scaling procedure and for a given tolerance TOL', the adaptive time-step strategy provides different errors for different Peer methods. However, the tolerance calibration defined by Eq. (29) allows to achieve the same error for different Peer schemes, as shown in Fig. 3. After the calibration each scheme shows comparable errors for the same user defined tolerance TOL. Tab. 5 summarizes the parameters of the calibration procedure.

Fig. 4 compare the performance of the Peer methods with the one-step traditional methods. The efficiency of the Peer methods increases with the order of convergence. The performance of peer6A and peer5A is comparable with RODASP scheme and peer4A with ROS3PL scheme. Only peer3A achieves a low computational efficiency.

4.2 Turbulent flow around a circular cylinder

The coupling between DG method and Peer schemes has been investigated also computing the turbulent flow around a circular cylinder for a Reynolds number $Re = 5 \times 10^4$, based on the cylinder diameter and the freestream conditions.

The far-field boundary is at 50 chords from the circle and the first cell height is equivalent to $y^+ \sim 4$ (see Fig. 5 for the mesh and Fig. 6 for the pressure and turbulent intensity contours).

Table 4: Travelling waves. Average time-step Δt_{avg} and Courant-Friedrichs-Lewy number CFL_{avg}, L^2 norm of the pressure ($\|err_p\|_2$) and velocity components ($\|err_u\|_2$, $\|err_v\|_2$) errors and the computational time for adaptive Peer schemes and different system tolerances TOL_{GMRES}, $TOL' = 10^{-3}$ and $\mathbb{P}^6$ solution approximation

	TOL_{GMRES}	Δt_{avg}	CFL_{avg}	$\|err_p\|_2$	$\|err_u\|_2$	$\|err_v\|_2$	CPU time [wu]
peer3A (ROS3PL)	1.00E-06	3.06E-03	0.97	6.05E-04	9.42E-04	9.42E-04	3.67E+03
	1.00E-05	3.06E-03	0.97	6.05E-04	9.42E-04	9.42E-04	3.28E+03
	1.00E-04	3.06E-03	0.97	6.00E-04	9.35E-04	9.35E-04	2.47E+03
	1.00E-03	3.06E-03	0.97	5.89E-04	9.13E-04	9.14E-04	2.14E+03
	1.00E-02	3.06E-03	0.97	5.89E-04	9.13E-04	9.14E-04	2.08E+03
	1.00E-01	3.05E-03	0.91	2.69E-03	4.23E-03	4.23E-03	1.76E+03
peer4A (ROS3PL)	1.00E-06	1.37E-02	4.36	9.66E-04	1.50E-03	1.50E-03	1.32E+03
	1.00E-05	1.37E-02	4.36	9.66E-04	1.50E-03	1.50E-03	1.10E+03
	1.00E-04	1.37E-02	4.36	9.66E-04	1.50E-03	1.50E-03	9.88E+02
	1.00E-03	1.37E-02	4.36	9.78E-04	1.52E-03	1.52E-03	8.51E+02
	1.00E-02	1.39E-02	4.42	1.24E-03	1.98E-03	1.98E-03	6.52E+02
	1.00E-01	5.92E-03	1.88	7.89E-03	8.46E-03	7.41E-03	1.71E+03
peer5A (RODASP)	1.00E-06	3.23E-02	10.28	1.14E-03	1.78E-03	1.78E-03	7.48E+02
	1.00E-05	3.23E-02	10.28	1.14E-03	1.78E-03	1.78E-03	6.73E+02
	1.00E-04	3.23E-02	10.28	1.14E-03	1.78E-03	1.78E-03	5.69E+02
	1.00E-03	3.23E-02	10.28	1.13E-03	1.77E-03	1.77E-03	4.50E+02
	1.00E-02	3.23E-02	10.28	1.21E-03	1.62E-03	1.71E-03	3.67E+02
peer6A (ROD5_1)	1.00E-06	5.56E-02	17.70	1.20E-03	1.86E-03	1.86E-03	6.24E+02
	1.00E-05	5.56E-02	17.70	1.20E-03	1.86E-03	1.86E-03	5.09E+02
	1.00E-04	5.56E-02	17.70	1.20E-03	1.87E-03	1.87E-03	4.49E+02
	1.00E-03	5.26E-02	16.77	9.75E-04	1.71E-03	1.94E-03	5.27E+02

Table 5: Travelling waves. Parameters for the calibration procedure

	k	β	TOL_0
peer3A	1.49	5.16	10^{-4}
peer4A	1.32	2.14	10^{-4}
peer5A	1.23	1.49	10^{-4}
peer6A	1.17	1.14	10^{-4}

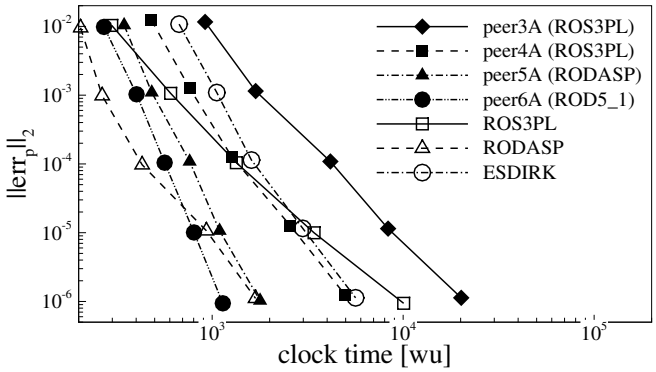


Figure 4: Travelling waves. L^2 norm of the pressure error ($\|err_p\|_2$) as a function of the computational time for adaptive Peer schemes, $TOL_{GMRES} = \eta\, TOL$ and $\mathbb{P}^6$ solution approximation

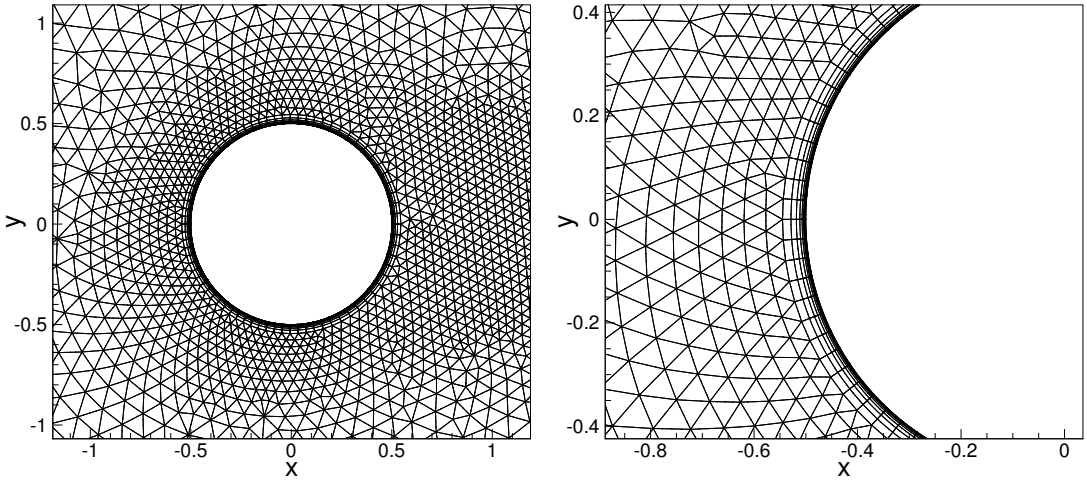


Figure 5: Cylinder. Computational mesh (left) of a circular cylinder, 8753 quadratic elements, and detail of the boundary layer (right)

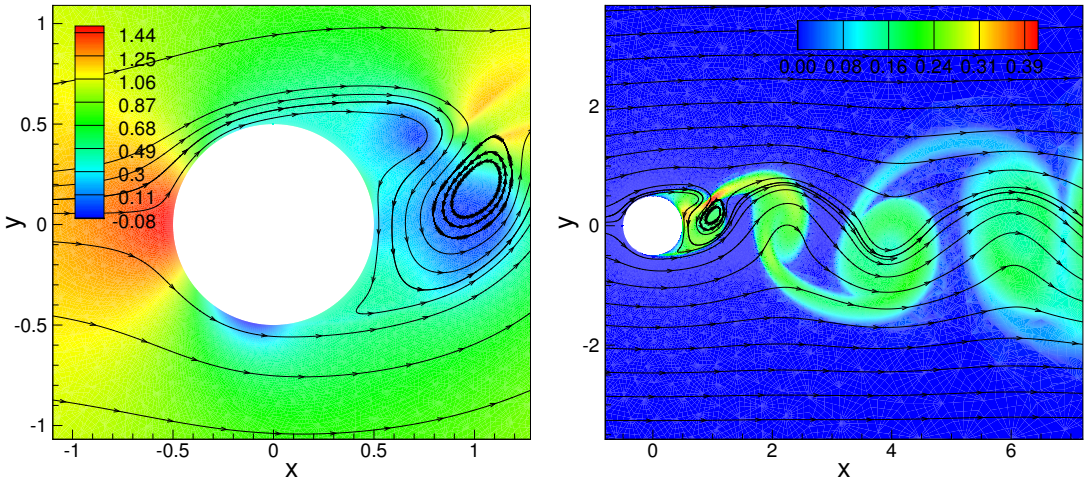


Figure 6: Cylinder. Pressure (left) and turbulence intensity (right) contours, $\mathbb{P}^3$ solution approximation

A $\mathbb{P}^3$ solution approximation provides a number of DOF, 87530, comparable with the computations on a fine mesh proposed by Carpenter *et. al.* on the same test case with a Finite Volume code [14]. The grid has been generated with a $2D$ high-order version of a fully automated in-house hybrid mesh generator based on the Advancing-Delaunay strategy [21].

Time integration error has been computed using the error of the lift coefficient:

$$err_{C_L} = C_L - C_{L,ref}. \tag{34}$$

The time integration period is set equal to one forth of the vortex shedding period. All computations start from the same initial flow field and the reference solution and the $C_{L,ref}$ have been obtained with the RODASP scheme and the following parameters for the adaptive time-step strategy: $TOL' = 10^{-9}$, $TOL_{GMRES} = \eta TOL'$.

The convergence order analysis has not been considered for this test case, due to the not optimal convergence results shown by Carpenter *et. al.* [14], which will certainly characterize also the schemes investigated in this work. Simulations have been carried out only with the adaptive time-step strategy and for the starting procedure the following one-step time integration

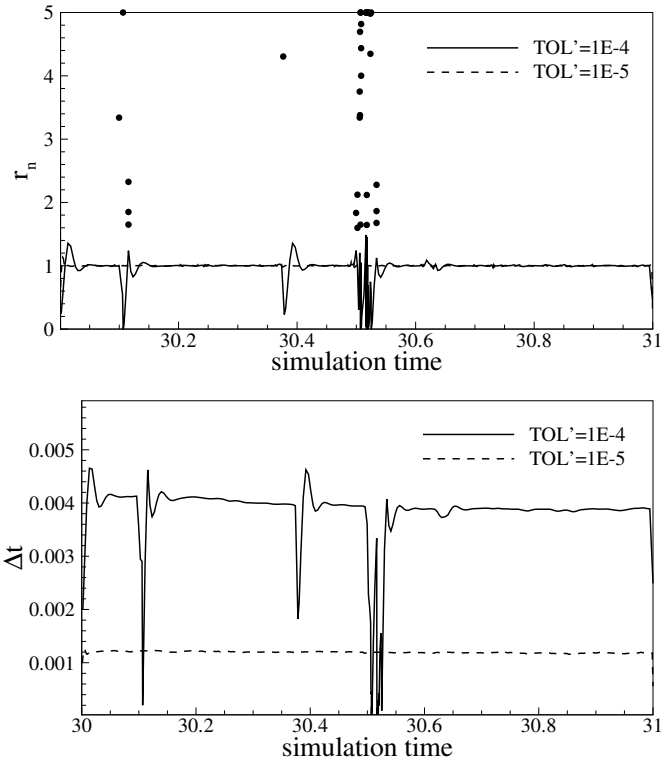


Figure 7: Cylinder. Evolution of the error estimator of the adaptive time-step strategy, r^{n+1}, (on the left) and of the time-step, Δt, (on the right) for adaptive peer3A scheme with $TOL' = 10^{-4}$ and $TOL' = 10^{-5}$, $TOL_{GMRES} = \eta TOL'$ and $\mathbb{P}^3$ solution approximation

schemes were used: ROS3PL for peer3 and peer4A, RODASP for peer5A and RODAS-ROD5_1 for peer6A.

The error on the lift coefficient, the computational time, the average time-step and the Courant-Friedrichs-Lewy number for different adaptive tolerances are reported in Tab. 6. Lower adaptive tolerance values can greatly increase the total number of the time-step and consequently the computational cost of simulations that thus can be beyond the available resources. For this reason very small adaptive tolerances, *i.e.* $TOL' = 10^{-7}$ and $TOL' = 10^{-8}$ for peer3A and $TOL' = 10^{-8}$ for peer4A, have not been considered in this work.

On the other hand, with higher adaptive tolerance values numerical instabilities appear, as depicted in Fig. 7, which shows the evolutions of the error estimator r^{n+1} and of the time-step for peer3A scheme ($TOL' = 10^{-4}$ and $TOL' = 10^{-5}$). The simulation with $TOL' = 10^{-5}$ is more stable, the error estimator slightly fluctuates around δ and there is no rejected time-step. Instead, the adaptive time-step strategy with $TOL' = 10^{-4}$ shows strong variations of the error estimator, leading to a high number of rejected time-steps, approximately 10%, and, as a consequence, to a higher computational cost. The origin of these instabilities could be related to the loss of the linearisation properties, an well known issue that can also affect Rosenbrock schemes.

As a consequence, optimal adaptive tolerances can be found for each Peer method: $TOL' = 10^{-4}$ for peer3A, $TOL' = 10^{-5}$ for peer4A and $TOL' = 10^{-7}$ for peer5A. Simulations with Peer6A have not been performed due to excessive computing time required by the small adaptive tolerance ($TOL' = 10^{-8}$), necessary to avoid numerical instabilities.

A performance comparison between the peer methods and the traditional one-step methods is shown in Figs. 8 and 9. As seen in the first test case, peer3A scheme must only be taken

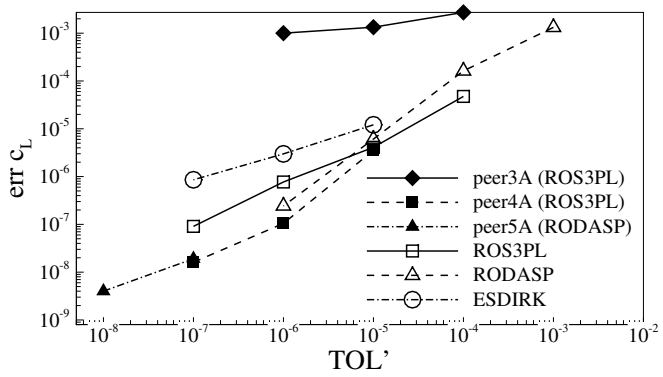


Figure 8: Cylinder. Error of the lift coefficient, err_{C_L}, as a function of the adaptive tolerance TOL' for adaptive Peer schemes, $TOL_{GMRES} = \eta TOL'$ and $\mathbb{P}^3$ solution approximation

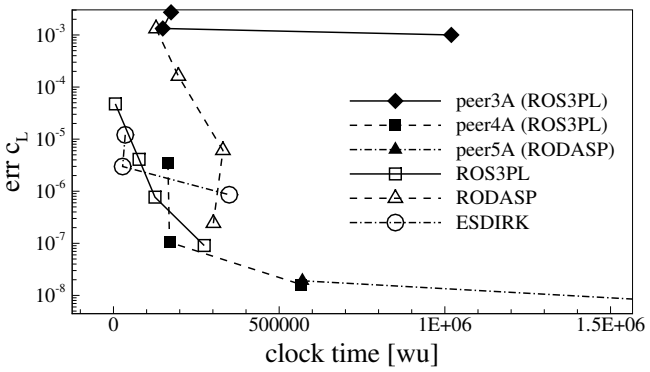


Figure 9: Cylinder. Error of the lift coefficient, err_{C_L}, as a function of the computational time for adaptive Peer schemes, $TOL_{GMRES} = \eta TOL'$ and $\mathbb{P}^3$ solution approximation

into account for very low level of accuracy, *i.e.* $err_{C_L} \geq 10^{-4}$. Instead, peer4A scheme with the optimal values of the adaptive tolerance TOL' is more efficient than ROS3PL scheme. Finally, because of numerical instabilities, peer5A scheme can become competitive in terms of efficiency only for very high accuracy, *i.e.* $err_{C_L} \leq 10^{-7}$

As in the previous test case, the tolerance calibration procedure can be used to reduce the difference in accuracy between time integration schemes. Calibration parameters are reported in Tab. 7.

5 Conclusions

In this work we have implemented the linearly implicit two-step Peer methods in a high-order discontinuous Galerkin solver for the numerical solution of incompressible unsteady flows. The performance, in terms of accuracy and computational efficiency, have been compared on laminar and turbulent test cases with one-step traditional methods.

The assessment showed different behaviours for the considered schemes depending on the flow regime of the test case, *i.e.* laminar or turbulent. On the laminar test case the Peer accuracy increases with the order of convergence leading to an efficiency improvement of time integration other traditional one-step implicit methods for high accuracy requirement. Instead, for the turbulent test case, even if Peer methods show a very high accuracy, their efficiency is comparable to Rosenbrock and ESDIRK ones.

Future works will assess Peer methods on other incompressible laminar and turbulent test

Table 6: Cylinder. Error of the lift coefficient, err_{C_L}, the average time-step Δt_{avg}, the average Courant-Friedrichs-Lewy number CFL_{avg}, the number of the rejected time-step [%] and the computational time for adaptive Peer schemes and different adaptive tolerance TOL', $TOL_{GMRES} = \eta TOL'$ and $\mathbb{P}^3$ solution approximation

	TOL'	err_{C_L}	Δt_{avg}	CFL_{avg}	% rejected time-step	CPU time [wu]
peer3A	1.00E-04	2.72E-03	2.95E-03	241.68	11.27%	1.73E+05
	1.00E-05	1.33E-03	1.19E-03	97.65	0%	1.48E+05
	1.00E-06	1.00E-03	3.77E-04	30.91	0%	1.02E+06
	1.00E-07	-	-	-	-	-
	1.00E-08	-	-	-	-	-
peer4A	1.00E-04	-	-	-	-	-
	1.00E-05	3.53E-06	3.12E-03	255.23	28.04%	1.64E+05
	1.00E-06	1.05E-07	1.70E-03	139.34	6.29%	1.69E+05
	1.00E-07	1.61E-08	7.03E-04	57.58	8.78%	5.67E+05
	1.00E-08	-	-	-	-	-
peer5A	1.00E-04	-	-	-	-	-
	1.00E-05	-	-	-	-	-
	1.00E-06	-	-	-	-	-
	1.00E-07	1.91E-08	1.09E-03	89.25	14.49%	5.70E+05
	1.00E-08	4.00E-09	4.24E-04	34.73	23.07%	2.50E+06

Table 7: Cylinder. Parameters for the calibration procedure

	k	β	TOL_0
peer3A	0.22	10^{-3}	10^{-6}
peer4A	1.17	9.55	10^{-6}
peer5A	0.68	5.24	10^{-7}

cases, to have a deeper understanding of their behaviour and performance, considering also time dependent boundary conditions, which can be handled by Peer schemes without order reduction. Furthermore, the coupling between DG method and Peer schemes will be evaluate also for compressible test cases.

Acknowledgements

HPC resources provided by CINECA within the "Convenzione di Ateneo Università degli Studi di Brescia" and "Convenzione di Ateneo Università degli Studi di Bergamo" are gratefully acknowledged.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[2] F. Bassi, L. Botti, A. Colombo, A. Crivellini, A. Ghidoni, and F. Massa. On the development of an implicit high-order Discontinuous Galerkin method for DNS and implicit LES of turbulent flows. *European Journal of Mechanics - B/Fluids*, pages –, 2015.

[3] F. Bassi, L. Botti, A. Colombo, A. Crivellini, A. Ghidoni, A. Nigro, and S. Rebay. Time Integration in the Discontinuous Galerkin Code MIGALE - Unsteady Problems. In Norbert Kroll, Charles Hirsch, Francesco Bassi, Craig Johnston, and Koen Hillewaert, editors, *IDIHOM: Industrialization of High-Order Methods - A Top-Down Approach*, volume 128 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 205–230. Springer International Publishing, 2015.

[4] F. Bassi, L. Botti, A. Colombo, D.A. Di Pietro, and P. Tesini. On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. *Journal of Computational Physics*, 231(1):45 – 65, 2012.

[5] F. Bassi, L. Botti, A. Colombo, A. Ghidoni, and F. Massa. Linearly implicit Rosenbrock-type RungeKutta schemes applied to the Discontinuous Galerkin solution of compressible and incompressible unsteady flows. *Computers & Fluids*, 118:305 – 320, 2015.

[6] F. Bassi, A. Colombo, C. De Bartolo, N. Franchina, A. Ghidoni, and A. Nigro. Investigation of high-order temporal schemes for the discontinuous Galerkin solution of the Navier-Stokes equations. *11th World Congress on Computational Mechanics, WCCM 2014, 5th European Conference on Computational Mechanics, ECCM 2014 and 6th European Conference on Computational Fluid Dynamics, ECFD 2014*, pages 5651–5662, 2014.

[7] F. Bassi, A. Crivellini, D. A. Di Pietro, and S. Rebay. An artificial compressibility flux for the discontinuous Galerkin solution of the incompressible Navier–Stokes equations. *J. Comput. Phys.*, 218(2):794–815, 2006.

[8] F. Bassi, A. Ghidoni, A. Perbellini, S. Rebay, A. Crivellini, N. Franchina, and M. Savini. A high-order discontinuous galerkin solver for the incompressible RANS and $k-\omega$ turbulence model equations. *Computers & Fluids*, 98:54 – 68, 2014.

[9] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997. Technologisch Instituut.

[10] Andrea D. Beck, Thomas Bolemann, David Flad, Hannes Frank, Gregor J. Gassner, Florian Hindenlang, and Claus-Dieter Munz. High-order discontinuous Galerkin spectral element methods for transitional and turbulent flow simulations. *International Journal for Numerical Methods in Fluids*, 76(8):522–548, 2014.

[11] F. Brezzi, M. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations*, 16:365–378, 2000.

[12] Simmendinger C. and Versick D. Integrated performance analysis of computer systems (ipacs). benchmarks for distributed computer systems. *Praxis der Informationsverarbeitung und Kommunikation*, 28(3), 2005.

[13] M. H. Carpenter, C. A. Kennedy, Hester Bijl, S. A. Viken, and Veer N. Vatsa. Fourth-order Runge—Kutta schemes for fluid mechanics applications. *J. Sci. Comput.*, 25(1):157–194, October 2005.

[14] M.H. Carpenter, S.A. Viken, and E.J. Nielsen. The efficiency of high order temporal schemes. In *41st Aerospace Sciences Meeting and Exhibit*, 2003.

[15] C. Carton de Wiart, K. Hillewaert, M. Duponcheel, and G. Winckelmans. Assessment of a discontinuous Galerkin method for the simulation of vortical flows at high Reynolds number. *Int. J. Numer. Methods Fluids*, 74(7):469–493, 2014.

[16] J.-B. Chapelier, M. de la Llave Plata, F. Renac, and E. Lamballais. Evaluation of a high-order discontinuous Galerkin method for the DNS of turbulent flows. *Comput. Fluids*, 95(0):210 – 226, 2014.

[17] A. Crivellini, V. D'Alessandro, and F. Bassi. Assessment of a high-order discontinuous Galerkin method for incompressible three-dimensional Navier-Stokes equations: Benchmark results for the flow past a sphere up to Re=500. *Computers and Fluids*, 86:442–458, 2013.

[18] E. Fehlberg. Classical fifth-, sixth-, seventh-, and eighth-order runge-kutta formulas with stepsize control. *NASA TR R-287*, 1968.

[19] Gregor Gassner, Michael Dumbser, Florian Hindenlang, and Claus-Dieter Munz. Explicit one-step time discretizations for discontinuous galerkin and finite volume schemes based on local predictors. *Journal of Computational Physics*, 230(11):4232 – 4247, 2011. Special issue High Order Methods for {CFD} Problems.

[20] A. Gerisch, J. Lang, H. Podhaisky, and R. Weiner. High-order linearly implicit two-step peer - finite element methods for time-dependent pdes. *Applied Numerical Mathematics*, 59(34):624 – 638, 2009. Selected Papers from NUMDIFF-11.

[21] A. Ghidoni, E. Pelizzari, S. Rebay, and V. Selmin. 3D anisotropic unstructured grid generation. *Int. J. Numer. Meth. Fl.*, 51(9-10):1097–1115, 2006.

[22] Florian Hindenlang, Gregor J. Gassner, Christoph Altmann, Andrea Beck, Marc Staudenmaier, and Claus-Dieter Munz. Explicit discontinuous Galerkin methods for unsteady problems. *Computers & Fluids*, 61:86 – 93, 2012.

[23] J. Lang and D. Teleaga. Towards a fully space-time adaptive FEM for magnetoquasistatics. *Magnetics, IEEE Transactions on*, 44(6):1238–1241, June 2008.

[24] G. Di Marzo. *RODAS5(4)-Mthodes de Rosenbrock d'ordre 5(4) adaptes aux problemes diffrentiels-algbriques*. 1993.

[25] A. Nigro, C. De Bartolo, F. Bassi, and A. Ghidoni. Up to sixth-order accurate A-stable implicit schemes applied to the discontinuous Galerkin discretized Navier-Stokes equations. *J. Comput. Phys.*, 276(0):136 – 162, 2014.

[26] A. Nigro, A. Ghidoni, S. Rebay, and F. Bassi. Modified Extended BDF scheme for the discontinuous Galerkin solution of unsteady compressible flows. *Int. J. Numer. Methods Fluids*, 76(9):549–574, 2014.

[27] H. Podhaisky, R. Weiner, and B.A. Schmitt. Rosenbrock-type 'peer' two-step methods. *Applied Numerical Mathematics*, 53:409 – 420, 2005. Dedicated to Profesor Karl Strehmel, on the occasion of his 70th birthday.

[28] Gustaf Soderlind. Digital filters in adaptive time-stepping. *ACM Trans. Math. Softw.*, 29(1):1–26, March 2003.

[29] Gustaf Soderlind and Lina Wang. Adaptive time-stepping and computational stability. *Journal of Computational and Applied Mathematics*, 185(2):225 – 243, 2006. Special Issue: International Workshop on the Technological Aspects of Mathematics.

[30] F. Steinebach. Order reduction of ROW methods for DAEs and method of lines applications. *Techn. Hochsch., Fachbereich Mathematik*, 1995.

[31] Fedderik van der Bos and Bernard J. Geurts. Computational error-analysis of a discontinuous galerkin discretization applied to large-eddy simulation of homogeneous turbulence. *Computer Methods in Applied Mechanics and Engineering*, 199(1316):903 – 915, 2010. Turbulence Modeling for Large Eddy Simulations.

[32] Liang Wei and Andrew Pollard. Direct numerical simulation of compressible turbulent channel flows using the discontinuous Galerkin method. *Comput. Fluids*, 47(1):85 – 100, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

OUTPUT-BASED SPACE-TIME ADAPTATION WITH NON-VARIATIONAL TIME INTEGRATION

Krzysztof J. Fidkowski[1]

[1]University of Michigan
1320 Beal Avenue, Ann Arbor, MI, 48109
e-mail: kfid@umich.edu

Keywords: Error Estimation, Adaptation, Unsteady, Non-Variational, Discontinuous Galerkin.

Abstract. *In this paper we present an output-based method for estimating numerical errors in unsteady simulations. This method applies to both variational and non-variational time integrators, and we focus on the latter. Key to the method is the solution of a continuous-in-time adjoint; that is, the unsteady adjoint system is derived from the semi-discrete form of the primal using a discretization of choice that needs not be the same as the primal. This provides flexibility in choosing the most appropriate and efficient time integration methods for the primal and adjoint problems separately. We also present a general approach for reconstructing the solution within a time interval, which is required for an accurate adjoint solution when using high-order methods. We then show how the output error estimate can be separated into contributions from spatial and temporal discretizations using a spatial down-projection of the adjoint. Finally, we present a space-time mesh adaptation procedure that appropriately targets spatial and temporal errors by setting equal ratios of marginal error to marginal cost of each refinement option, space or time. We demonstrate the error estimation and adaptation method for a simple prototypical problem: unsteady linear advection in one spatial dimension.*

1 INTRODUCTION

In an output-based setting, mesh resolution is automatically dictated by an error estimate of an output of interest. Much work has been done in this area for steady problems with finite volume and finite element methods [13, 2, 9, 16, 15, 12, 6, 17]. However, unsteady problems pose additional challenges and computational costs, namely in the unsteady adjoint solution. Yet output-based adaptive methods have also been explored for such problems, with various mechanics of adaptation, including static-mesh, dynamic-mesh, space-only, and combined space-time [14, 1, 4, 11, 3, 8, 7, 5, 10].

A variational discretization, such as a finite element method in space and time, puts output error estimation on a sound theoretical foundation and simplifies the error calculation and localization [6]. In particular, the adjoint solution, which is at the core of output-based methods, can be solved using a *discrete* approach in which the adjoint equations are obtained systematically from the primal discrete system by transposing the operator. This is the approach taken in many previous works [14, 1, 7, 5, 10]. While the same approach can be applied to non-variational discretizations, it has pitfalls as the relationship between the resulting discrete adjoint coefficients and the underlying continuous adjoint solution may not be clear or simple to work with, especially if employing approximate adjoint solvers that rely on "smoothness" of the adjoint solution.

On the other hand, non-variational methods are ubiquitous for time integration of unsteady problems. Multi-step and multi-stage methods dominate such simulations, due to their simplicity and generally lower computational cost compared to variational methods. There is therefore a need to extend output-based methods to such discretizations, to increase their utility and prevalence, and this is the topic of the present work. Some work in this area has already been done [11, 8], with scheme-specific algorithms for computing or approximating the fine-space discrete adjoint and linear or spline-based interpolation of solutions between time nodes. In this work we take a step back and derive a general approach for error estimation, through a *continuous*-in-time adjoint solution, that accommodates nearly arbitrary time integration methods, even of high order, and fine-space choices. We also show how to separate the error into contributions of the spatial and temporal discretization.

In the remainder of this paper we first introduce our discretization in Section 2, which is variational in space but could be non-variational in time. In Section 3 we derive a continuous-in-time adjoint solution and show how to use it to estimate temporal errors, and spatial errors when also using spatial refinement. A key part of the adjoint solution and the error estimate is a reconstruction of the primal and adjoint solution in time, for which we present a scheme-independent approach in Section 3.2. Section 3.4 outlines the space-time adaptive procedure, and Section 4 presents results that demonstrate the methods for an advection problem.

2 DISCRETIZATION

We are interested in the simulation of phenomena governed by partial differential equations in space and time. We write the PDE in canonical form as

$$\frac{\partial u}{\partial t} + L(u) = 0, \tag{1}$$

where u is the state, and $L(\cdot)$ is the continuous spatial residual operator. We discretize this equation in space and time separately, using a semi-discrete approach, as outlined in this section.

2.1 Spatial

We employ a discontinuous Galerkin (DG) finite-element method in space. Denote by T_h the set of N_{elem} elements in a non-overlapping tessellation of the domain Ω. In DG, the state is approximated by polynomials of order p on each element, with no continuity constraints imposed on the approximations on adjacent elements. Formally, for a scalar equation, we write that $u_h \in \mathcal{V}_h$, where $\mathcal{V}_h = \left\{u \in L_2(\Omega) : u|_{\Omega_e} \in \mathcal{P}^p \; \forall \Omega_e \in T_h\right\}$, and $\mathcal{P}^p$ denotes polynomials of order p on an element Ω_e. The weak-form of (1) follows from multiplying the PDE by test functions in the same approximation space, integrating by parts, and coupling elements via common fluxes. Choosing a basis for the test and trial spaces yields a system of ordinary differential equations,

$$\bar{\mathbf{R}}(\mathbf{U}) \equiv \mathbf{M}\frac{d\mathbf{U}}{dt} + \mathbf{R}(\mathbf{U}) \;\; = \;\; \mathbf{0}, \tag{2}$$

where $\mathbf{M}$ is the mass matrix, $\mathbf{U} \in \mathbb{R}^N$ is the discrete state vector of basis function coefficients, $\mathbf{R}$ is the discrete spatial residual vector, and $\bar{\mathbf{R}}$ is the strong-form unsteady residual.

2.2 Temporal

We consider general marching schemes for advancing the system of ODEs in (2) in time. We do not assume a specific form for the time discretization and only require that the method advances the state one time step, Δt: from time node n to $n+1$, i.e. $\mathbf{U}^n \rightarrow \mathbf{U}^{n+1}$. This encompasses both *variational* time integrators, based on a weak-form in time, and non-variational integrators such as traditional multi-step and multi-stage methods. In this work we show results for a DG-in-time variational discretization [7] (DG1 and DG2), standard backwards-difference multi-step methods (BDF1 and BDF2), and diagonally-implicit Runge-Kutta methods (DIRK3 and DIRK4). For example, an n_{stage} DIRK method takes the form

$$\begin{aligned}
&\text{for} \quad i = 1 : n_{\text{stage}} \\
&\qquad \mathbf{S}_i = -\frac{\mathbf{M}}{\Delta t}\mathbf{W}^0 + \sum_{j=1}^{i-1} a_{ij}\mathbf{R}(\mathbf{W}^j, t_j) \\
&\qquad \text{solve: } \frac{\mathbf{M}}{\Delta t}\mathbf{W}^i + a_{ii}\mathbf{R}(\mathbf{W}^i, t_i) + \mathbf{S}_i = \mathbf{0} \\
&\text{end}
\end{aligned}$$

where $\mathbf{W}^0 = \mathbf{U}^n$ is the state at the start of the time interval, $\mathbf{U}^{n+1} = \mathbf{W}^{n_{\text{stage}}}$ is the desired result, and $t_i = t^n + b_i \Delta t$ are the stage times. The coefficients a_{ij} and b_i define the method. For example, a fourth-order accurate scheme (DIRK4) with $n_{\text{stage}} = 5$ has

$$a_{ij} = \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{17}{50} & -\frac{1}{25} & \frac{1}{4} & 0 & 0 \\ \frac{371}{1360} & -\frac{137}{2720} & \frac{15}{544} & \frac{1}{4} & 0 \\ \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4} \end{bmatrix}, \quad b_i = \begin{bmatrix} \frac{1}{4} \\ \frac{3}{4} \\ \frac{11}{20} \\ \frac{1}{2} \\ 1 \end{bmatrix}.$$

3 OUTPUT ERROR ESTIMATION

In previous work, the author and collaborators have presented output error estimation for unsteady problems using the discrete space-time adjoint and a variational discretization in time. In the present work, to accommodate general time integration methods that do not necessarily possess a temporal weak form, we rely on a continuous-in-time adjoint, as described in this section.

3.1 The continuous-in-time adjoint

Consider a time-integral output of a simulation,

$$\bar{J} \equiv \int_0^T J(\mathbf{U}, t)\, dt, \tag{3}$$

where $J(\mathbf{U}, t)$ is a function of the spatial distribution of the state, via the discrete coefficients $\mathbf{U}$, and of time [1]. The continuous-in-time adjoint $\mathbf{\Psi}(t)$ is the sensitivity of $\bar{J}$ to source perturbations in the unsteady residual, $\bar{\mathbf{R}}$, (2). To derive the equation for $\mathbf{\Psi}(t)$, we define a Lagrangian as

$$\mathcal{L} \equiv \bar{J} + \int_0^T \mathbf{\Psi}^T \bar{\mathbf{R}}\, dt = \bar{J} + \int_0^T \mathbf{\Psi}^T \left(\mathbf{M} \frac{d\mathbf{U}}{dt} + \mathbf{R}(\mathbf{U}) \right) dt. \tag{4}$$

Integrating the first term in the integral by parts, we have

$$\mathcal{L} = \bar{J} + \left[\mathbf{\Psi} \mathbf{M} \mathbf{U} \right]_0^T + \int_0^T \left[-\frac{d\mathbf{\Psi}^T}{dt} \mathbf{M} \mathbf{U} + \mathbf{\Psi}^T \mathbf{R}(\mathbf{U}) \right] dt. \tag{5}$$

Now requiring stationarity of the Lagrangian with respect to permissible state variations gives

$$\mathcal{L}_{\mathbf{U}}(\delta\mathbf{U}) = \bar{J}_{\mathbf{U}}(\delta\mathbf{U}) + \left[\mathbf{\Psi} \mathbf{M} \delta\mathbf{U} \right]_0^T + \int_0^T \left[-\frac{d\mathbf{\Psi}^T}{dt} \mathbf{M} \delta\mathbf{U} + \mathbf{\Psi}^T \mathbf{R}_{\mathbf{U}}(\delta\mathbf{U}) \right] dt = \mathbf{0}, \tag{6}$$

where the subscript denotes differentiation with respect to the discrete state vector $\mathbf{U}$. The middle term drops out since $\delta\mathbf{U} = 0$ at $t = 0$, and $\mathbf{\Psi} = 0$ at $t = T$ ($\bar{J}$ does not depend on the terminal state). Combining the remaining terms,

$$\int_0^T \left[J_{\mathbf{U}} - \frac{d\mathbf{\Psi}^T}{dt} \mathbf{M} + \mathbf{\Psi}^T \mathbf{R}_{\mathbf{U}} \right] \delta\mathbf{U}\, dt = \mathbf{0}. \tag{7}$$

This equation must hold for all $\delta\mathbf{U}$, which forces the term in the square brackets to zero. Transposing this term produces the continuous-in-time adjoint equation,

$$-\mathbf{M} \frac{d\mathbf{\Psi}}{dt} + \mathbf{R}_{\mathbf{U}}^T \mathbf{\Psi} + J_{\mathbf{U}}^T = \mathbf{0}, \tag{8}$$

with the terminal condition $\mathbf{\Psi}(T) = \mathbf{0}$. Due to the terminal condition, the adjoint is solved backwards in time, from $t = T$ to $t = 0$. In the present work, we do not restrict the choice of time integration for the adjoint equation to be variational or even the same as the primal. We note that when applying standard time integrators for marching backwards in time, it is useful to define $\tau = T - t$ and to rewrite (8) in a form similar to the primal,

$$\mathbf{M} \frac{d\mathbf{\Psi}}{d\tau} + \underbrace{\mathbf{R}_{\mathbf{U}}^T \mathbf{\Psi} + J_{\mathbf{U}}^T}_{\mathbf{R}^{\mathbf{\Psi}}} = \mathbf{0}, \tag{9}$$

where $\mathbf{R}^{\mathbf{\Psi}}$ is the adjoint residual.

[1] More general outputs that involve end-of-time states could also be considered in a straightforward manner.

3.2 Temporal reconstruction

Solving the adjoint using (9) requires evaluating the adjoint residual, which in general depends on the primal state, at not only time nodes but possibly also on the interval in between – for example, if using a multi-stage method. These in-between evaluations are not directly available for general non-variational time integrators. The same situation occurs in error estimation, where both the primal and adjoint are required in between time nodes for integrating the adjoint-weighted residual. In this section, we describe how to reconstruct the solution in time, independent of the time scheme used.

The temporal reconstruction is based on the semi-discrete form of the PDE, (2) [2], which we can re-arrange to solve for the state slope,

$$\frac{d\mathbf{U}}{dt} = -\mathbf{M}^{-1}\mathbf{R}(\mathbf{U}). \tag{10}$$

That is, we can obtain the slope in time through a residual evaluation at a known state. Suppose that we know the states at the endpoints of a time interval, t^n and t^{n+1} – this is what a general time-marching scheme would give us. A very simple reconstruction would be to just connect these states linearly, but that would only give us second-order accuracy. We can do much better if we apply (10) to evaluate the slopes at the endpoints, and using these four pieces of information reconstruct a cubic solution in time, as shown in Figure 1.

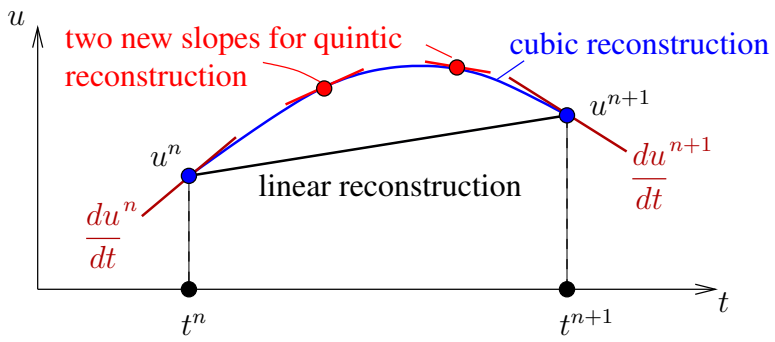


Figure 1: Illustration of solution reconstruction on a time interval. High-order representations can be constructed using slope information readily available from residual evaluations.

Furthermore, we can gain additional accuracy by applying (10) yet again, in between the time nodes, using the reconstructed cubic solution. In particular, for error estimation we will be computing integrals over the time interval via quadrature. For two quadrature points, we evaluate the cubic at those two points and apply (10) to obtain the slopes there. With these additional two pieces of information, we can reconstruct a quintic solution in time (i.e. the unique quintic in time that matches the two endpoint values, two endpoint slopes, and two in-between slopes). However, as the original state came from a cubic reconstruction, the order of accuracy of this quintic will only be one higher than a cubic. To obtain true quintic accuracy, we simply iterate the process, using the current quintic approximation to re-evaluate the slopes at the midpoints, which then yields a new quintic. Just one such iteration will yield the expected convergence rates, but an additional iteration will further lower the errors.

Table 1 shows the results from a numerical test of the reconstruction. The system of interest is a scalar ordinary differential equation, $\frac{du}{dt} = u^2$. Reconstruction is performed over one

[2] For reconstructing the adjoint, we would use (9).

interval, $[0, \Delta t]$, and the error is defined as $E^2 = \frac{1}{\Delta t} \int_0^{\Delta t} (u - u_{\text{exact}})^2 \, dt$. The results show the optimal fourth-order convergence for the cubic reconstruction, and sixth-order convergence for the quintic reconstruction with at least one iteration.

	Cubic		Quintic-0		Quintic-1		Quintic-2	
Δt	Error	Rate	Error	Rate	Error	Rate	Error	Rate
1/4	8.87e-05	–	6.02e-06	–	8.54e-07	–	1.56e-07	–
1/8	7.24e-06	3.62	2.58e-07	4.55	1.93e-08	5.47	3.55e-09	5.46
1/16	5.22e-07	3.79	9.55e-09	4.75	3.68e-10	5.71	6.79e-11	5.71
1/32	3.52e-08	3.89	3.26e-10	4.87	6.38e-12	5.85	1.18e-12	5.85
1/64	2.28e-09	3.94	1.07e-11	4.93	1.03e-13	5.95	1.75e-14	6.07

Table 1: Errors and orders of accuracy of time-step reconstruction of a sample scalar problem, using endpoint values and slopes only (cubic) and additional interior slopes and n iterations (quintic-n).

3.3 Error estimation

An adjoint solution can be used to estimate the numerical error in the corresponding output of interest, $\bar{J}$, through the adjoint-weighted residual [2, 6]. Let's first consider only temporal errors. Denote by $\mathbf{U}_H(t)$ the approximate primal solution obtained from a chosen time integration method and time step size. If we had the exact unsteady adjoint solution, $\mathbf{\Psi}(t)$, we could use (4) to estimate the error in $\bar{J}$,

$$\delta \bar{J} \equiv \bar{J}(\mathbf{U}_H) - \bar{J}(\mathbf{U}) \approx \bar{J}_{\mathbf{U}}(\delta \mathbf{U}) \approx - \int_0^T \mathbf{\Psi}^T \bar{\mathbf{R}}(\mathbf{U}_H) \, dt, \tag{11}$$

where $\delta \mathbf{U} \equiv \mathbf{U}_H - \mathbf{U}$ is the state error, and $\bar{\mathbf{R}}(\mathbf{U}_H) \approx \bar{\mathbf{R}}_{\mathbf{U}}(\delta \mathbf{U})$ is the generally nonzero unsteady residual obtained from the approximate primal. In practice, the exact adjoint is not available and must be approximated in a *fine space*, denoted by subscript h, which in our work will be a higher-order time integration method. Numerical error due to the spatial discretization can be measured by also making the fine space higher order in space. In the present work this is accomplished by increasing the spatial order of the DG discretization by one. The final form of the error estimate is

$$\delta \bar{J} \approx - \int_0^T \mathbf{\Psi}_h^T \bar{\mathbf{R}}_h(\mathbf{U}_h^H) \, dt, \tag{12}$$

where $\mathbf{U}_h^H$ is the injection of the primal from space H to space h. In the spatial domain, this is a pure injection to higher order: $p \to p+1$. In the temporal domain, this involves a sufficiently-accurate reconstruction over the time interval. For the time integration schemes considered in this work, we employ a cubic time reconstruction for this injection. The integral in (12) is a summation of integrals over all time intervals, and within each time interval we perform the integral using quadrature – typically between 2 and 4 points, depending on the accuracy of the time integration. For this integral we need to evaluate the adjoint inside the time intervals, and we use a reconstruction appropriate for the fine-space temporal discretization – typically quintic.

Finally, we note that the output error estimate in (12) can be separated into spatial and temporal components by selectively refining the fine space only in space or only in time. Specifically, we obtain the temporal error, $\delta \bar{J}^{\text{time}}$, by projecting the fine-space adjoint, $\mathbf{\Psi}_h$, spatially

back down to order p and recalculating the error via (12) with the projected adjoint. As non-variational time integrators do not support a straightforward projection, we do not use a similar approach for the spatial error. Instead, we simply define $\delta\bar{J}^{\text{space}} \equiv \delta\bar{J} - \delta\bar{J}^{\text{time}}$.

3.4 Space-time adaptation

The error estimate in (12) can be localized to contributions of individual time intervals and spatial elements, which could then be used to drive targeted mesh and time-step refinements. However, in the present work we confine adaptation to very simple mechanics: we assume independent but uniform spatial and temporal meshes, and we use the error estimates to guide adaptation either in space or in time (or a combination of both).

The adaptation relies on models of how the spatial and temporal errors behave with refinement. For this we use a standard a priori model,

$$\delta\bar{J}^{\text{space}} \propto (\Delta x)^{p+1}, \quad \delta\bar{J}^{\text{time}} \propto (\Delta t)^{r+1}, \tag{13}$$

where Δx is the spatial mesh size, Δt is the time step size, and r is the formal order of convergence of the time integration. The adaptation also requires a cost model, which we simply take as the total number of space-time degrees of freedom,

$$C \equiv N_{\text{elem}} N_t n_p n_r, \tag{14}$$

where N_t is the number of time intervals, n_p is the number of spatial degrees of freedom per element ($p+1$ in 1D), and n_r is the number of temporal degrees of freedom, i.e. system solves, per time step. For example, BDF2 would nave $n_r = 1$, while the DIRK schemes would have $n_r = n_{\text{stage}}$.

In our simple adaptive mechanics, at each adaptive iteration we multiply the number of spatial elements by f^{space}, and the number of time steps by f^{time}, with a constraint on the total cost rise, $C = C_0 f^{\text{tot}}$. Here, f^{tot} is prescribed by the user, e.g. a factor of 2. Using the cost formula in (14), the constraint becomes $f^{\text{space}} f^{\text{time}} = f^{\text{tot}}$. To determine f^{space} and f^{time}, we require that the marginal error to cost ratios of each refinement option are equal: $\lambda^{\text{space}} = \lambda^{\text{time}}$, where

$$\lambda^{\text{space}} = \frac{\partial(\delta\bar{J}^{\text{space}})}{\partial f^{\text{space}}} \left[\frac{\partial C}{\partial f^{\text{space}}}\right]^{-1}, \quad \lambda^{\text{time}} = \frac{\partial(\delta\bar{J}^{\text{time}})}{\partial f^{\text{time}}} \left[\frac{\partial C}{\partial f^{\text{time}}}\right]^{-1}. \tag{15}$$

Substituting into the above equations the error and cost models from (13) and (14), and using the constraint $f^{\text{space}} f^{\text{time}} = f^{\text{tot}}$, we obtain the following solution:

$$f^{\text{time}} = \left[\frac{r+1}{p+1} \frac{\delta\bar{J}^{\text{time}}}{\delta\bar{J}^{\text{space}}} (f^{\text{tot}})^{p+2}\right]^{-(r+p+4)}, \quad f^{\text{space}} = \frac{f^{\text{tot}}}{f^{\text{time}}}.$$

This equation appropriately adjusts the spatial and temporal refinement fractions to make equal the marginal error to cost ratio of the two refinement options. Depending on the ratio of the spatial and temporal errors, either the spatial or the temporal mesh will be refined more, and in some cases one of the meshes could be coarsened.

4 RESULTS

In this section we present results demonstrating the accuracy of the output error estimate, the accuracy of the spatial/temporal error breakdown, and the efficacy of output-based adaptation

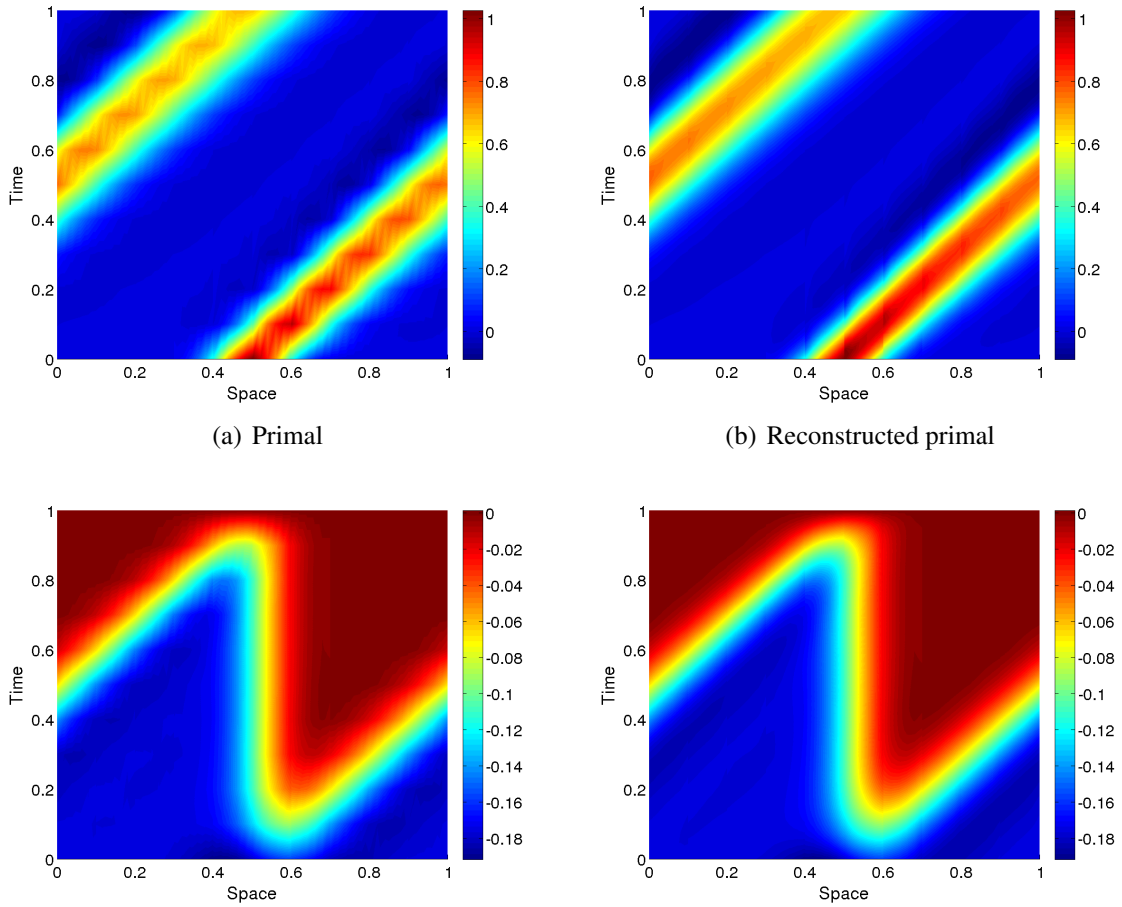
for an unsteady advection simulation. The governing equation is

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad u(x, t = 0) = u_0(x), \tag{16}$$

where $a = 1$ is the advection speed. The boundary conditions are periodic, and the initial condition is $u_0(x) = \exp\left[100(x/L - .5)^2\right]$, where $L = 1$ is the domain length. The output of interest is a weighted function of the state,

$$\bar{J} = \int_0^T \int_0^L w(x)u(x, t)\, dx\, dt, \tag{17}$$

where $w(x) = \exp\left[100(x/L - .52)^2\right]$, and the final simulation time is $T = 1$. Figure 2 shows the primal and adjoint solutions computed using 10 elements in space and 10 intervals in time, with $p = 2$ and DIRK3. The space-time fields without reconstruction are visualized by linearly interpolating the time node states, and these are visibly "choppy", masking the true high-order temporal accuracy. Reconstructing the primal and adjoint solutions using cubic functions over the time steps produces visible improvements in the space-time solutions in between the time nodes.



(a) Primal

(b) Reconstructed primal

(c) Adjoint

(d) Reconstructed adjoint

Figure 2: Primal and adjoint space-time solutions for the advection problem, with and without temporal reconstruction (cubic), shown for the case of $N_{\text{elem}} = 10$, $N_t = 10$, $p = 2$, DIRK3.

We first examine the effectivity of the total error estimate and the breakdown into spatial and temporal contributions. Table 2 lists these errors for various combinations of element and time step numbers, all for $p = 2$ and DIRK3 as the coarse space, and $p = 3$ and DG2 as the fine space. We see that the total error estimates are very accurate, with a maximum error of less than 7%. In addition, the spatial and temporal error estimates behave as expected: the spatial error diminishes when the number of elements increases, and similarly for the temporal error. The "actual" values of these errors were computed by refining the spatial and temporal discretizations independently, using concurrent uniform refinement and order increase, and the estimates agree very well with the actual values.

		δJ		δJ^{time}		δJ^{space}	
N_{elem}	N_t	Predicted	Actual	Predicted	Actual	Predicted	Actual
10	10	-5.66e-05	-6.03e-05	-8.89e-05	-1.11e-04	3.23e-05	1.56e-05
10	100	4.96e-05	5.14e-05	9.98e-07	8.70e-07	4.86e-05	5.01e-05
100	10	-7.00e-05	-7.59e-05	-7.00e-05	-7.65e-05	6.62e-09	1.80e-10
100	100	1.30e-06	1.30e-06	1.30e-06	1.30e-06	5.32e-10	5.28e-10

Table 2: Comparison of estimated and predicted output errors, and their breakdown into spatial and temporal components, for the advection simulation on several space-time meshes.

Next, several different orders and time integration schemes were tested in the context of space-time output-based refinement. In all cases, the total space-time degrees of freedom (i.e. the cost) were allowed to grow by a factor of $f^{\text{tot}} = 2$ per adaptive iteration, and 10 adaptive iterations were taken. The fine space consisted of $p + 1$ spatial approximation and time integration of one order higher (e.g. DIRK3 for BDF2). Figure 3 compares the convergence of the output error for different combinations of orders and time marching schemes. We see that the most efficient combination for this smooth problem is $p = 3$ and DIRK4 – as expected, high order in space and time.

To further examine the performance of the output-based adaptation, we compare the results for each run to uniform refinement of the starting space-time mesh. Figure 4 shows this comparison and includes an additional set of curves: the output error resulting after correction of the output with the error estimate. These curves converge the fastest, as expected given the previously observed effectivity of the error estimate. The rate of convergence is generally at least one higher than the rate of convergence of the output error without correction. Performing the worst is uniform refinement, where the convergence rate is limited by the dominant errors in space or in time.

Poor performance of uniform refinement is tied to the starting space-time mesh. For example, a suboptimal ratio of elements to time steps will not be addressed by uniform refinement. Figure 5 illustrates this point by comparing the separate spatial and temporal errors for adaptive runs and uniform refinement. In the case of $p = 2$ and BDF2, the temporal errors will dominate and must be addressed by increased allocation of resources to the temporal discretization (i.e. the number of time steps). This is what the adaptive refinement accomplishes in just a couple iterations, and for the remainder of the simulation the spatial and temporal errors roughly balance (since their marginal cost is identical). In the case of $p = 1$ and DIRK4, the spatial errors dominate and should be addressed by relatively more spatial refinement. Again, the adaptive run accomplishes this, whereas the uniform refinement run is consistently dominated by spatial errors.

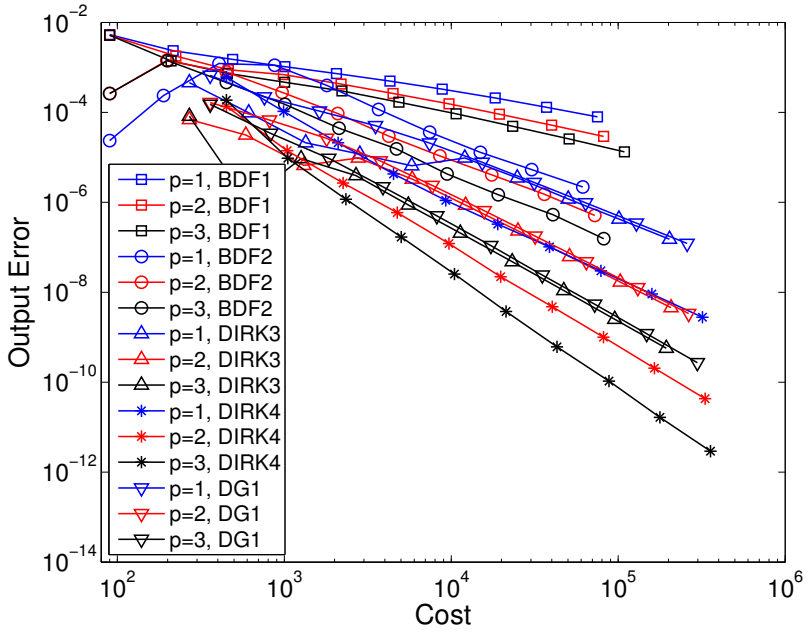


Figure 3: Convergence of the output error with cost, measured as the total space-time degrees of freedom, using output-based space-time adaptation.

5 CONCLUSIONS

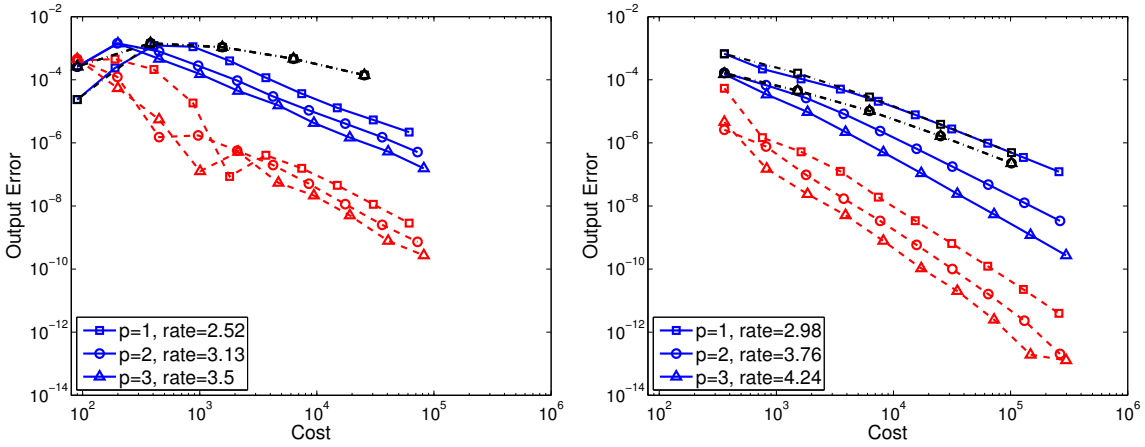
We have presented a general approach for estimating numerical errors in outputs of unsteady simulations when using non-variation time integration methods. The two key ingredients are: (1) a continuous-in-time adjoint solution that simplifies adjoint-consistency considerations and removes constraints on the choice of the adjoint time integration scheme; and (2) a scheme-agnostic temporal reconstruction procedure based on primal or adjoint residual evaluations, which provides a consistent, high-order functional representation of the solution within each time step. When used in conjunction with a spatial discretization, the method allows for separate space and time error estimates via adjoint down-projection. The resulting error estimates can then drive an space-time adaptive procedure, as demonstrated for a prototypical advection problem. Remaining work in this area includes investigation of fine-space adjoint approximations and application to more complex nonlinear simulations, such as those governed by the compressible Navier-Stokes equations.

REFERENCES

[1] Timothy J. Barth. Space-time error representation and estimation in Navier-Stokes calculations. In Stavros C. Kassinos, Carlos A. Langer, Gianluca Iaccarino, and Parviz Moin, editors, *Complex Effects in Large Eddy Simulations*, pages 29–48. Springer Berlin Heidelberg, Lecture Notes in Computational Science and Engineering Vol 26, 2007.

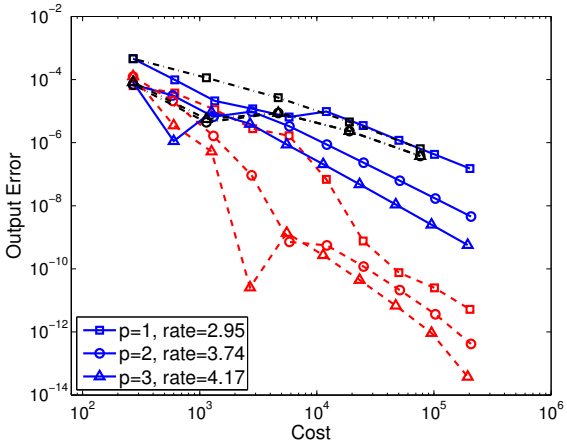
[2] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. In A. Iserles, editor, *Acta Numerica*, pages 1–102. Cambridge University Press, 2001.

[3] A. Belme, A. Dervieux, and F. Alauzet. Error estimation and adaptation for functional outputs in time-dependent flow problems. *Journal of Computational Physics*, 231:6323–

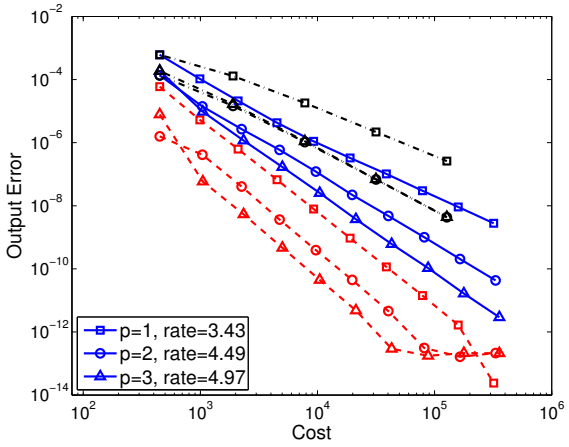


(a) BDF2

(b) DG1

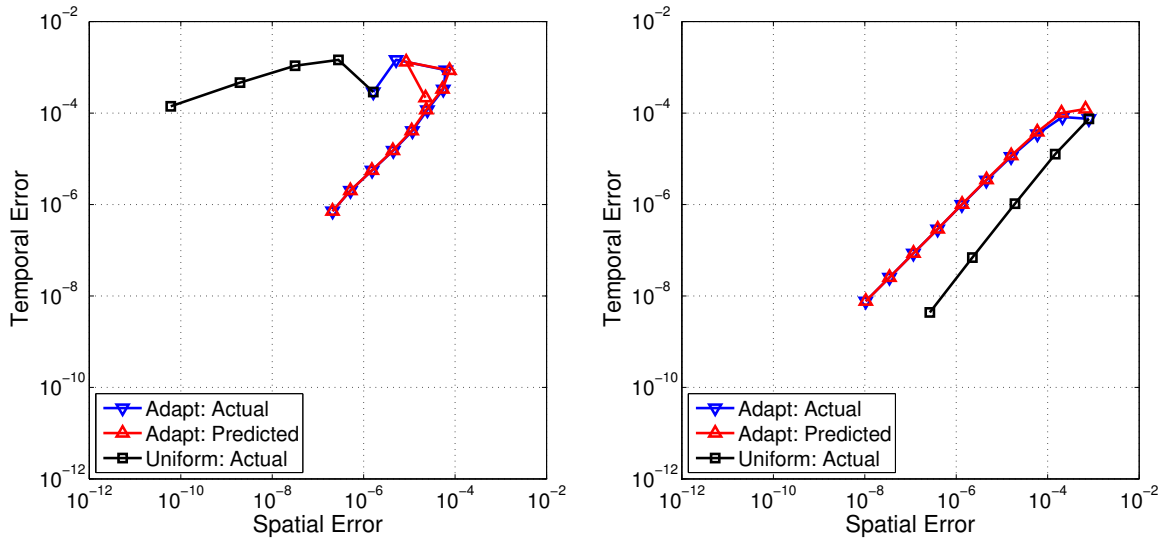


(c) DIRK3



(d) DIRK4

Figure 4: Convergence of the output error with cost, measured as the total space-time degrees of freedom, for various spatial orders and temporal discretizations. Adaptive refinement (solid blue) is compared to uniform refinement (dash-dot black), and to adaptive refinement corrected by the error estimate (dashed red).



(a) $p = 2$, BDF2 (b) $p = 1$, DIRK4

Figure 5: Spatial and temporal error distribution histories for adaptive and uniform refinement of the advection simulation, using two combinations of spatial and temporal discretizations.

6348, 2012.

[4] Michael Besier and Rolf Rannacher. Goal-oriented space-time adaptivity in the finite element galerkin method for the compution of nonstationary incompressible flow. *International Journal for Numerical Methods in Fluids*, 70:1139–1166, 2012.

[5] Krzysztof J. Fidkowski. An output-based dynamic order refinement strategy for unsteady aerodynamics. AIAA Paper 2012-77, 2012.

[6] Krzysztof J. Fidkowski and David L. Darmofal. Review of output-based error estimation and mesh adaptation in computational fluid dynamics. *American Institute of Aeronautics and Astronautics Journal*, 49(4):673–694, 2011.

[7] Krzysztof J. Fidkowski and Yuxing Luo. Output-based space-time mesh adaptation for the compressible Navier-Stokes equations. *Journal of Computational Physics*, 230:5753–5773, 2011.

[8] Bryan T. Flynt and Dimitri J. Mavriplis. Discrete adjoint based adaptive error control in unsteady flow problems. AIAA Paper 2012-0078, 2012.

[9] Ralf Hartmann and Paul Houston. Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. *Journal of Computational Physics*, 183(2):508–532, 2002.

[10] Steven M. Kast and Krzysztof J. Fidkowski. Output-based mesh adaptation for high order Navier-Stokes simulations on deformable domains. *Journal of Computational Physics*, 252(1):468–494, 2013.

[11] Karthik Mani and Dimitri J. Mavriplis. Error estimation and adaptation for functional outputs in time-dependent flow problems. *Journal of Computational Physics*, 229:415–440, 2010.

[12] Marian Nemec and Michael J. Aftosmis. Error estimation and adpative refinement for embedded-boundary Cartesian meshes. AIAA Paper 2007-4187, 2007.

[13] Niles A. Pierce and Michael B. Giles. Adjoint recovery of superconvergent functionals from PDE approximations. *SIAM Review*, 42(2):247–264, 2000.

[14] Michael Schmich and Boris Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM Journal on Scientific Computing*, 30(1):369–393, 2008.

[15] S. Sen, K. Veroy, D.B.P. Huynh, S. Deparis, N.C. Nguyen, and A.T. Patera. "Natural norm" a posteriori error estimators for reduced basis approximations. *Journal of Computational Physics*, 217:37–62, 2006.

[16] D. A. Venditti and D. L. Darmofal. Anisotropic grid adaptation for functional outputs: application to two-dimensional viscous flows. *Journal of Computational Physics*, 187(1):22–46, 2003.

[17] Michael Woopen, Aravind Balan, Georg May, and Jochen Schütz. A comparison of hybridized and standard DG methods for target-based hp-adaptive simulation of compressible flow. *Computers & Fluids*, 98:3–16, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MATRIX-FREE MODIFIED EXTENDED BACKWARD DIFFERENTIATION FORMULAE APPLIED TO THE DISCONTINUOUS GALERKIN SOLUTION OF COMPRESSIBLE UNSTEADY VISCOUS FLOWS

A. Nigro[1], C. De Bartolo[1], A. Crivellini[2], F. Bassi[3]

[1] University of Calabria-Department of Mechanical, Energy and Management Engineering-DIMEG
Ponte P. Bucci cubo 44/C, 87036 - Rende (CS), Italy
alessandra.nigro@unical.it, carmine.debartolo@unical.it

[2] Polytechnic University of Marche-Department of Industrial Engineering and Mathematical Sciences
Via Brecce Bianche, 60100 - Ancona (AN), Italy
a.crivellini@univpm.it

[3] University of Bergamo-Department of Engineering
Viale Marconi 5, 24044 - Dalmine (BG), Italy
francesco.bassi@unibg.it

Keywords: Modified Extended Backward Differentiation Formulae, Discontinuous Galerkin discretization, Matrix-free approximation, Frozen preconditioner.

Abstract. *In this work a matrix-free modified extended backward differentiation time integration method has been implemented in a high-order discontinuous Galerkin solver for the unsteady Navier-Stokes equations. The resulting non-linear systems at each time step are solved iteratively using a preconditioned inexact Newton/Krylov method. In order to speed-up the solution process a frozen preconditioner formulation and a polynomial extrapolation technique for computing a better initial guess for the Newton iterations have been considered. Numerical results for compressible inviscid and viscous test cases show the effectiveness of the proposed numerical strategies and the performance advantages of the matrix-free method compared to its matrix-explicit counterpart for this class of implicit multi-stage time schemes. Furthermore, the influence of some physical (low Mach) and space discretization (stretched grid) aspects is examined to highlight pros and cons of the proposed time integration algorithm and its potential in solving non-stiff and stiff systems with respect to the widely used explicit Runge-Kutta schemes.*

1 INTRODUCTION

Because of the availability of very fast and massively parallel computers, there is a current trend to use advanced flow models to predict the unsteady turbulent structures directly. In order of complexity from most to least complex we distinguish Direct Numerical Simulation (DNS) (no turbulence modeling at all), the Large Eddy Simulation (LES) that is based on a SubGrid Scale (SGS) model for the non-resolved turbulent scales, the Implicit Large Eddy Simulation (ILES), where the spatial discretization itself acts like a SGS model, and hybrid RANS-LES (RANS near the solid body, LES elsewhere). All of these approaches require the coupling of high space accuracy with high temporal accuracy to convect all resolved turbulent scales at the right speed with minimal dissipation and dispersion. In this respect, the use of very accurate time integration schemes in the context of high-order methods gives the possibility of efficient simulations with high accuracy. Among the high-order methods, the Discontinuous Galerkin (DG) finite element method turns out to be an ideal candidate for dealing with such complex flow computations do to its high precision together with a natural amenability to parallelization (for an overview of the DG method see [1]).

Explicit high-order Runge-Kutta schemes combined with the DG space discretization (RK-DG methods) [2, 3, 4], are commonly used to address the numerical solution of unsteady flows. Explicit methods are well suited for problems with similar spatial and temporal scales, but become inefficient for unsteady flows of low reduced frequency, as well as for stiff problems. Due to their better stability properties, implicit methods are potentially more effective than explicit methods in this respect. Several high-order semi-implicit multi-stage (Rosenbrock) and implicit multi-step multi-stage (MEBDF, TIAS) time integration schemes for DG space approximations have been recently proposed for both inviscid and laminar flows [5, 6, 7, 8]. Although these schemes allow to overcome the strong restriction to the size of the time step typical of explicit methods, their use require to compute, store and factorize large Jacobian matrices. These tasks require vast computational resources both in terms of CPU time and memory, especially for high-order spatial discretizations and large scale problems.

This problem can be addressed using the matrix-free approach. Different matrix-free methods for the numerical solution of partial differential equations have been develop in the past by Gear and Saad [9], Chan and Jackson [10] and Brown and Hindmarsh [11]. Further, Johan [12] and McHugh and Knoll [13, 14] have used these methods to solve compressible and incompressible flows. More recently, Rasetarinera and Hussani [15] and K. Hillewaert et al. [16] applied the matrix-free method in the context of the DG solution of the Euler equations. A similar approach has been adopted by Crivellini and Bassi [17] for the DG discretizion of the Navier-Stokes and RANS equations. In the above cited DG papers the authors developed implicit solvers for steady compressible flows. In this work we adopt the matrix-free approach for unsteady computations solving the DG discretized compressible Navier-Stokes equations with the high-order accurate MEBDF method. These implicit time integration formulae belong to the family of predictor-corrector methods. MEBDF schemes involve three non-linear stages: the first two are predictor stages that use a standard k-step BDF scheme, the last one is a corrector stage that uses an advanced implicit k-step formula of order k + 1. MEBDF are A-stable for k = 1, .., 3 and stiffly stable for k = 4, .., 8.

The matrix-free approach seems well suited to be coupled with this class of non-linear multi-stage time schemes [18]. In fact, the additional cost with respect to a matrix-explicit approach to approximate the matrix-vector product, at each non-linear iteration, is equal to one residual evaluation, which is much cheaper with respect to the evaluation of the Jacobian matrix required

by the matrix-explicit approach for this update. This computational advantage is expected to compensate the further residual evaluations for each linear iteration, required by the matrix-free approach and not necessary for the matrix-explicit one, as the MEBDF schemes are composed by three non-linear stages. Furthermore, at large time steps as well as for stiff problems, the update of the implicit matrix-vector product is determinant for the convergence of the non-linear Newton iterations at each stage.

The system matrix of high-order accurate DG discretizations is usually ill-conditioned, hence preconditioning is essential to accelerate the convergence of the linear solver. Here, according to [17], the preconditioner consists of an incomplete lower-upper (ILU0) factorization of the analytically computed system matrix. The matrix-free approach allows to approximate the preconditioner without affecting the quadratic convergence of the Newton loops. Furthermore, the MEBDF stages represent non-linear systems with the same system matrix. Both these features are central to the performance of the unsteady solver, avoiding the update of the preconditioning matrix within each time step by applying the same preconditioner to the linear systems of the predictor-corrector stages. Note that the efficiency of the algorithm will be further enhanced when the preconditioner can be frozen for succeeding time steps, e.g. non-stiff ODE systems, small time steps. Another important computational aspect is related to the choice of the initial guess for the Newton-type solver at each non-linear stage of the MEBDF scheme. In particular, for the second Newton loop a polynomial extrapolation technique to generate a more accurate Newton initialization is adopted using a set of previous solutions.

The proposed test cases for inviscid and viscous flows are the vortex transport by uniform flow and a laminar manufactured solution, both with an analytical solution. The shown results aim at demonstrating the capability of the proposed matrix-free approach (MF-MEBDF), in combination with the numerical strategies proposed, to improve the performance of the implicit solver with respect to its explicit counterpart (ME-MEBDF), highlighting its pros and cons and its performance in solving stiff and non-stiff systems of equations with respect to the widely used explicit Runge-Kutta schemes.

In the following of the paper the governing equations are presented in Section 2. Sections 3 and 4 are devoted to space (DG) and time (MEBDF) discretizations, respectively. Numerical results are presented and discussed in Section 5. Conclusions are finally reported in Section 6.

2 GOVERNING EQUATIONS

The compressible Navier-Stokes equations in conservative form are

$$\frac{\partial \mathbf{q}}{\partial t} + \nabla \cdot \mathbf{F}_c\left(\mathbf{q}\right) = \nabla \cdot \mathbf{F}_v\left(\mathbf{q}, \nabla \mathbf{q}\right), \tag{1}$$

where $\mathbf{q}$ is the vector of conservative variables, $\mathbf{q} = \left[\rho, \rho u, \rho v, \rho E\right]^T$, and $\mathbf{F}_c\left(\mathbf{f}_c, \mathbf{g}_c\right)$ and $\mathbf{F}_v\left(\mathbf{f}_v, \mathbf{g}_v\right)$ are the inviscid and viscous flux vectors respectively, given by

$$\mathbf{f}_c = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ \rho H u \end{pmatrix}, \qquad \mathbf{g}_c = \begin{pmatrix} \rho v \\ \rho v u \\ \rho v^2 + p \\ \rho H v \end{pmatrix}; \tag{2}$$

$$\mathbf{f}_v = \begin{pmatrix} 0 \\ \tau_{xx} \\ \tau_{yx} \\ \tau_{xx} u + \tau_{yx} v - q_x \end{pmatrix}, \qquad \mathbf{g}_v = \begin{pmatrix} 0 \\ \tau_{xy} \\ \tau_{yy} \\ \tau_{xy} u + \tau_{yy} v - q_y \end{pmatrix}. \tag{3}$$

In these equations ρ is the fluid density, u and v are the x and y velocity components, respectively, p is the pressure and E is the total internal energy for unit mass. The total enthalpy for unit mass H is given by $H = E + p/\rho$. The shear stress tensor components τ_{ij} and the heat flux vector components q_i of the viscous flux vectors can be calculated as

$$\tau_{xx} = \left(2\mu\frac{\partial u}{\partial x} + \lambda\nabla\cdot\mathbf{v}\right), \qquad \tau_{yy} = \left(2\mu\frac{\partial v}{\partial y} + \lambda\nabla\cdot\mathbf{v}\right), \tag{4}$$

$$\tau_{xy} = \tau_{yx} = \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \tag{5}$$

$$q_x = -\kappa\frac{\partial T}{\partial x}, \qquad q_y = -\kappa\frac{\partial T}{\partial y}. \tag{6}$$

In order to close the system of equations, the Navier-Stokes equations must be augmented by algebraic expressions that relate the internal energy E, the pressure p, the dynamic viscosity μ, the second viscosity coefficient λ and the conductivity coefficient κ to the thermodynamic state of the fluid. For an ideal gas, assuming that the fluid satisfies the equation of state of perfect gas, the pressure is given by $p = \rho\left(\gamma - 1\right)\left[E - \left(u^2 + v^2\right)/2\right]$, where γ is the ratio of specific heats of the fluid, given by $\gamma = C_p/C_v$. The dynamic viscosity coefficient μ can be approximated using the power-law viscosity formula

$$\frac{\mu}{\mu_0} = \left(\frac{T}{T_0}\right)^{3/4}, \tag{7}$$

where T_0, μ_0 are temperature and viscosity at the reference conditions, respectively.

3 DG DISCRETIZATION

To discretize in space Eq. (1), we consider an approximation Ω_h of the domain Ω consisting of a set of non-overlapping elements $\tau_h = \{K\}$, denoting by $\partial\Omega_h$ the boundary of the discrete approximation and by Γ_h^0 the set of internal edges. We consider piecewise polynomial functions on τ_h with no global continuity requirement. Denoting with $P_n\left(K\right)$ the space of polynomial functions of degree at most n in the element K, the solution and test functions space is defined by

$$\mathbf{V}_h = \{\mathbf{v}_h \in \left(L^2\left(\Omega_h\right)\right)^{N+2} : \mathbf{v}_h \in \left(P_n\left(K\right)\right)^{N+2} \forall K \in \tau_h\}, \tag{8}$$

where N is the number of spatial dimensions. To build a set of hierarchical and orthogonal shape functions for the space (8) the Modified Gram-Schmidt (MGS) procedure [19, 20, 21] is used, considering the starting set of monomial functions of the same degree n. By using the BR2 scheme presented in [22, 23] and theoretically analyzed in [24, 25], we obtain the DG formulation of Eq. (1), given by

$$\begin{aligned}
&\int_{\Omega_h}\mathbf{v}_h\cdot\frac{\partial\mathbf{q}_h}{\partial t}dx - \int_{\Omega_h}\nabla\mathbf{v}_h : \left(\mathbf{F}_c\left(\mathbf{q}_h\right) - \mathbf{F}_v\left(\mathbf{q}_h, \nabla\mathbf{q}_h + \mathcal{R}\left(\left[\!\left[\mathbf{q}_h\right]\!\right]_0\right)\right)\right)dx \\
&+ \int_{\Gamma_h^0}\left(\mathbf{v}_h^- - \mathbf{v}_h^+\right)\cdot\mathbf{H}\left(\mathbf{q}_h^+, \mathbf{q}_h^-, \mathbf{n}^-\right)d\sigma - \int_{\Gamma_h^0}\left[\!\left[\mathbf{v}_h\right]\!\right] : \{\mathbf{F}_v\left(\mathbf{q}_h, \nabla\mathbf{q}_h + \eta_e\mathcal{R}_e\left(\left[\!\left[\mathbf{q}_h\right]\!\right]_0\right)\right)\}d\sigma \\
&+ \int_{\partial\Omega_h}\left(\mathbf{v}_h\otimes\mathbf{n}\right) : \left(\mathbf{H}\left(\mathbf{q}_h^+, \mathbf{q}_h^b, \mathbf{n}\right) - \mathbf{F}_v\left(\mathbf{q}_h, \nabla\mathbf{q}_h + \eta_e\mathcal{R}_e\left(\left[\!\left[\mathbf{q}_h\right]\!\right]_0\right)\right)_b\right)d\sigma = 0.
\end{aligned} \tag{9}$$

In the above equation η_e is called "penalty" parameter and its lower bound is established as the number of neighbours elements of the generic element K to guarantee the stability of the method. $\mathcal{R}_e\left([\![\mathbf{q}_h]\!]_0\right)$ and $\mathcal{R}\left([\![\mathbf{q}_h]\!]_0\right)$ are the local and global lifting operators, respectively, accounting in the gradient of the diffusive fluxes for the jumps in $\mathbf{q}_h$ occurring at the element interfaces, defined as

$$[\![\mathbf{q}_h]\!]_0 = \begin{cases} [\![\mathbf{q}_h]\!] & \text{on } \Gamma_h^0 \\ \mathbf{q}_h \otimes \mathbf{n} & \text{on } \partial\Omega_h \end{cases}, \tag{10}$$

$$[\![\mathbf{q}_h]\!] = \mathbf{q}_h^+ \otimes \mathbf{n}^+ + \mathbf{q}_h^- \otimes \mathbf{n}^-. \tag{11}$$

The trace operator $\{(\cdot)\}$ denotes the average between left $(\cdot)^-$ and right state $(\cdot)^+$, see Fig.(1). Finally, $\mathbf{H}\left(\mathbf{q}_h^+, \mathbf{q}_h^-, \mathbf{n}\right)$ and $\mathbf{H}\left(\mathbf{q}_h^+, \mathbf{q}_h^b, \mathbf{n}\right)$ are the numerical flux functions at the interior and boundary faces, respectively, for which any of the numerical flux functions commonly considered in the finite volume method can be used.

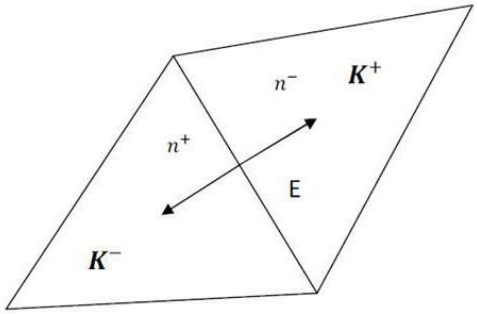


Figure 1: Two elements K^+ and K^- sharing edge E.

A detailed analysis of the Discontinuous Galerkin method, including mathematical foundations, practical implementation aspects and examples that demonstrate the advantages of this approach, can be found in the collection of articles edited by Cockburn et al. [1] and in the book written by Di Pietro et al. [21].

4 TIME INTEGRATION

Numerical integration of the DG space discretization, reported in Eq. (9), results in the following system of Ordinary Differential Equations (ODEs) in time

$$\mathbf{M}\frac{d\mathbf{Q}}{dt} + \mathbf{R}\left(\mathbf{Q}\right) = 0, \tag{12}$$

where $\mathbf{M}$ is the global block mass matrix that, for the conservative variables, reduces to the identity matrix by using orthonormal shape functions (MGS algorithm), $\mathbf{Q}$ is the global vector of unknown degrees of freedom and $\mathbf{R}$ is the vector of residuals. For the sake of simplicity in what follows we have omitted the mass matrix, being $\mathbf{M} = \mathbf{I}$.

4.1 MEBDF method

In this work the system of ODEs of Eq. (12) is advanced in time implicitly by using the MEBDF method [7, 26, 27, 28], which has been originally proposed for the numerical solution

of stiff initial value problems in an attempt to derive a class of multi-step integration formulae characterized by better stability properties and higher-order accuracy than the standard Backward Differentiation Formulae (BDF). The result is a class of k-step formulae of order $k+1$ that are A-stable for $k = 1, .., 3$ and stiffly stable for $k = 4, .., 8$. The general k-step MEBDF algorithm consists of successively solving the following three stages to advance the solution in time:

- Stage 1. Compute the first predictor $\bar{\mathbf{Q}}_{n+1}$ of order k with a k-step BDF:

$$\hat{\alpha}_1 \bar{\mathbf{Q}}_{n+1} + \sum_{j=2}^{k+1} \hat{\alpha}_j \mathbf{Q}_{n+2-j} + \Delta t \hat{\beta}_k \mathbf{R}\left(\bar{\mathbf{Q}}_{n+1}\right) = 0; \tag{13}$$

- Stage 2. Compute the second predictor $\bar{\mathbf{Q}}_{n+2}$ of order k with a k-step BDF:

$$\hat{\alpha}_1 \bar{\mathbf{Q}}_{n+2} + \hat{\alpha}_2 \bar{\mathbf{Q}}_{n+1} + \sum_{j=3}^{k+1} \hat{\alpha}_j \mathbf{Q}_{n+3-j} + \Delta t \hat{\beta}_k \mathbf{R}\left(\bar{\mathbf{Q}}_{n+2}\right) = 0; \tag{14}$$

- Stage 3. Compute the corrected solution $\mathbf{Q}_{n+1}$ of order $k+1$ as:

$$\bar{\alpha}_1 \mathbf{Q}_{n+1} + \sum_{j=2}^{k+1} \bar{\alpha}_j \mathbf{Q}_{n+2-j} + \Delta t \left[\hat{\beta}_k \mathbf{R}\left(\mathbf{Q}_{n+1}\right) + \left(\bar{\beta}_k - \hat{\beta}_k\right) \bar{\mathbf{R}}_{n+1} + \bar{\beta}_{k+1} \bar{\mathbf{R}}_{n+2}\right] = 0, \tag{15}$$

where

$$\bar{\mathbf{R}}_{n+1} = \mathbf{R}\left(\bar{\mathbf{Q}}_{n+1}\right) \quad \text{and} \quad \bar{\mathbf{R}}_{n+2} = \mathbf{R}\left(\bar{\mathbf{Q}}_{n+2}\right). \tag{16}$$

$\hat{\alpha}_j$ and $\hat{\beta}_j$ are BDF coefficients, while $\bar{\alpha}_j$ and $\bar{\beta}_j$ are MEBDF coefficients. The values of both these coefficients are reported in [26]. Note that, the better accuracy and stability properties of MEBDF with respect to BDF are obtained extending this latter by adding two further stages and a future point at t_{n+2}. Therefore, the first two stages of the method are standard k-step BDF schemes, while the last stage uses an advanced implicit k-step formula of order $k+1$.

4.2 Implementation issues

All the non-linear stages are solved with a Newton-type method. By reformulating Eqs. (13–15 in terms of a non-linear residual function $\mathcal{F}$ and a constant vector $\mathbf{b}$, a generic stage can be rewritten in a more compact form as

$$\mathcal{F}\left(\mathbf{Q}\right) = \mathbf{b}, \tag{17}$$

with

$$\mathcal{F}\left(\mathbf{Q}\right) = \mathbf{Q} + \beta \Delta t \mathbf{R}\left(\mathbf{Q}\right), \tag{18}$$

$$\mathbf{b} = -\sum_j \alpha_j \mathbf{Q}_j \qquad \text{(Predictor stages)}, \tag{19}$$

$$\mathbf{b} = -\sum_j \alpha_j \mathbf{Q}_j - \Delta t \sum_i \beta_i \mathbf{R}_i \qquad \text{(Corrector stage)}, \tag{20}$$

where $\mathbf{Q}_j$ is a previous known solution calculated at t_j time and $\mathbf{R}_i$ the residual evaluation at the end of each predictor stage. Notice that the overbars on the quantities computed at each predictor stage have been omitted, and that β, α_j and β_i are generic stage coefficients.

After applying the Newton scheme to Eq. (17) the following system of equations is obtained:

$$\mathbf{W}^l \Delta \mathbf{Q}^l = -\mathcal{F}\left(\mathbf{Q}^l\right) + \mathbf{b}, \tag{21}$$

with

$$\Delta \mathbf{Q}^l = \mathbf{Q}^{l+1} - \mathbf{Q}^l, \tag{22}$$

$$\mathbf{W}^l = \frac{\partial \mathcal{F}}{\partial \mathbf{Q}}\left(\mathbf{Q}^l\right) = \mathbf{I} - \Delta t \beta \frac{\partial \mathbf{R}}{\partial \mathbf{Q}}\left(\mathbf{Q}^l\right), \tag{23}$$

where $\mathbf{W}^l$ is the system matrix.

To successfully and efficiently solve the linear system of equations given by (21), the preconditioned GMRES algorithm available in PETSc [29] is used. Applying left-preconditioning we solved

$$\left[\mathbf{P}^l\right]^{-1} \mathbf{W}^l \Delta \mathbf{Q} = -\left[\mathbf{P}^l\right]^{-1} \left[\mathcal{F}\left(\mathbf{Q}^l\right) + \mathbf{b}\right], \tag{24}$$

where the preconditioning matrix $\mathbf{P}$ is chosen to be the effective incomplete lower-upper factorization ILU(0) of the analitically computed system matrix $\mathbf{W}$.

In order to ensure an accurate and efficient iterative solution of the Newton loop at each stage, the non-linear initialization procedure implemented in the algorithm is as follows:

- the solution of the first stage at the current time step is initialized with the solution of the second stage calculated at the previous time step;
- the solution of the second stage is initialized with a polynomial extrapolation using a set of previous solutions;
- the solution of the third stage is initialized with the solution of the first stage.

4.2.1 Matrix-Free formulation

Referring to Eq. (21), we note that the system matrix does not need to be computed and stored explicitly, but only its action on the vector $\Delta \mathbf{Q}^l$ is required. Therefore, the product of $\mathbf{W}^l$ times the vector $\Delta \mathbf{Q}^l$ has been replaced by the first-order Taylor series expansion

$$\mathbf{W}^l \Delta \mathbf{Q}^l \simeq \frac{\mathcal{F}\left(\mathbf{Q}^l + \varepsilon \Delta \mathbf{Q}^l\right) - \mathcal{F}\left(\mathbf{Q}^l\right)}{\varepsilon}, \tag{25}$$

where the ϵ value is computed as proposed by Pernice and Walker [30].

This method requires two residual evaluations per outer Newton iteration and one residual function evaluation per inner GMRES iteration. Since the residual evaluation is more expensive than a matrix-vector product, the Matrix-Free (MF) GMRES is less performing than its explicit counterpart. However, MEBDF consists of three stages and for each non-linear system it may require a system matrix construction to ensure the convergence of the Newton loop when the Matrix-Explicit (ME) approach is used. In fact, not updating the implicit matrix at the beginning of each stage or within the Newton loop reduces the convergence rate of the non-linear

iterations, especially in the case of large time steps, stiff ODEs systems and high-resolution discretizations. Since only one additional residual evaluation is required by the MF approach with respect to its explicit counterpart, which is much more cheaper than the implicit matrix evaluation, it is expected that, in the above cited cases, for this class of multi-stage schemes MF becomes competitive in cost to the ME approach.

Although the iterative matrix is not built, a preconditioning matrix $\mathbf{P}$ is still needed for the convergence of GMRES. Thus, the proposed MF solver adopts the storage of the analitically computed system matrix for preconditioning purpose. The MF approach allows to approximate the preconditioner without affecting the quadratic convergence of the Newton loop. Furthermore, the MEBDF stages are non-linear systems with the same β coefficient of the implicit matrix (see Eq.(23)). Therefore, for small time steps, non-stiff problems or low-resolution discretizations, the MF approach is expected to effectively exploit the preconditioner freeze strategy. The freezing strategy considered in this paper consists in computing the preconditioner at the beginning of stage 1 every n time steps.

4.2.2 Matrix Explicit formulation

In this version of the implicit solver the system matrix is explicitly formed. This matrix is then used as preconditioner and to avoid the matrix-vector product approximation and the related residual function evaluation. Thus, the algorithm requires the storage of both the system matrix $\mathbf{W}$ and the preconditioning matrix $\mathbf{P}$. The system matrix is computed at the beginning of each time step and it is recomputed if the convergence of the Newton method becomes too slow. In particular, the implicit matrix is re-evaluated when the ratio between the L_2 norm of two successive Newton solution variations is less than 5.

5 RESULTS

Two test cases are presented to assess the temporal accuracy and the performance of the proposed MF-MEBDF scheme for the time-dependent compressible Euler and Navier-Stokes equations: an inviscid isentropic convecting vortex and a laminar Manufactured Solution, both with a known analytical solution. Furthermore, to demonstrate the performance advantages of the MF approach with the preconditioner freeze strategy, it will be compared with its ME counterpart and with the explicit 5-stage fourth order accurate Strong Stability Preserving Runge-Kutta scheme (RK4) [31] for non-stiff and stiff systems of equations. In particular, the convection of the isentropic vortex is investigated for high and low free stream Mach numbers, while the laminar computations are performed on uniform and stretched cartesian grids.

The simulations have been performed using the 3-step fourth-order accurate MEBDF scheme (MEBDF4) and the exact Riemann solver of Gottlieb and Groth [32] to approximate the inviscid numerical fluxes. As concerns the second stage's initial guess for the Newton-type solver, Eq. (14), the following fourth-order polynomial approximation is used:

$$\bar{\mathbf{Q}}^0_{n+2} = 4\bar{\mathbf{Q}}_{n+1} - 6\mathbf{Q}_n + 4\mathbf{Q}_{n-1} - \mathbf{Q}_{n-2}. \tag{26}$$

Another issue to be addressed using the MEBDF multi-step scheme concerns the unknown initial solutions needed to start the temporal integration. For both test cases the additional starting values are obtained evaluating the exact solution at the appropriate time levels. Furthermore, as a result of preliminary computations, it has established that a linear-solver normalized-residual tolerance of 10^{-2} allows for efficient computations while the Newton tolerance of one order of magnitude smaller than the solution error allows to obtain careful estimates of the

time integration error avoiding unnecessarily restrictive and computationally more expensive tolerance levels, especially at large time steps.

All the computations are performed in serial using a 2.80 GHz Intel Xeon CPU E5-2680v2 processor with 6.4 Gb of RAM memory.

5.1 Convection of an isentropic vortex

The convection of an inviscid isentropic vortex is a standard test case to evaluate the performance of the numerical methods. As the vortex is convected without distortion by the mean flow, the exact solution is obtained by translating the initial solution at the velocity set by the freestream, providing a reference for measuring the accuracy of the numerical solution. The initial conditions are taken from [33, 34].

The setup of the problem is as follows: the computational domain is defined as $0 \leq x, y \leq 10$ and the boundary conditions are periodic. Temporal convergence studies are conducted, up to a final time T corresponding to one period of vortex revolution, to assess the accuracy and the efficiency of the implicit time integration solvers. The simulations are performed on a 50×50 uniform cartesian grid with a seventh-order accurate polynomial approximations ($P6$ elements). The analysis is carried for stiff and non-stiff ODEs systems obtained by considering the following free stream Mach numbers: $M_\infty \simeq 1.4; 0.14; 0.014$.

5.1.1 High Mach number: non-stiff problems

The results of the convergence study at the highest Mach number ($M_\infty \simeq 1.4$) are presented in Fig. 2 for the MF, freezing the preconditioner, and the ME algorithms. The L_2-norm errors of the density field as a function of the time step are presented on the left while the work-precision characteristics on the right. The left plot demonstrates that the convergence histories computed using the two algorithms are indistinguishable. Furthermore, it can be seen that the design-order of convergence of 4 is achieved over the time interval that has been investigated. The right plot highlights that MF always outperforms ME. Moreover, it appears that using the proposed MF algorithm the efficiency, in terms of CPU time, significantly improves as the time step reduces. The results of both temporal refinement study and efficiency are collected in Tab. 1. To investigate the performance improvement of MF compared to ME, the CPU ratios and the optimal updating period of the preconditioner (n_{opt}), that minimizes the overall CPU time, are also reported in the table. It appears evident the benefit of the preconditioner freeze strategy, with the CPU ratio that increases with the growing of the n_{opt} parameter as the time step reduces. For example, to reach an error level of $8.10 \cdot 10^{-3}$ the MF algorithm is about 1.34 times faster than the ME algorithm, while for an accuracy level of $\mathcal{O}\left(10^{-10}\right)$ this ratio increases to 10.10.

The influence of the time consuming Jacobian evaluation on the efficiency of the implicit solver at large ($\Delta t = T/40$) and small ($\Delta t = T/160$) time steps is described in Tab. 2, where the performance of MF and ME algorithms are compared at each stage of the time scheme.

As regards the largest time step, at stage 1, when both the algorithms evaluate the Jacobian at the beginning of the Newton iterations, the performance of the two implicit solvers are roughly the same except for the number of function evaluations. Notice that the same can be observed in all stages of both time steps. The reason is in the additional residual evaluations necessary in the GMRES and Newton methods when the MF approach is used.

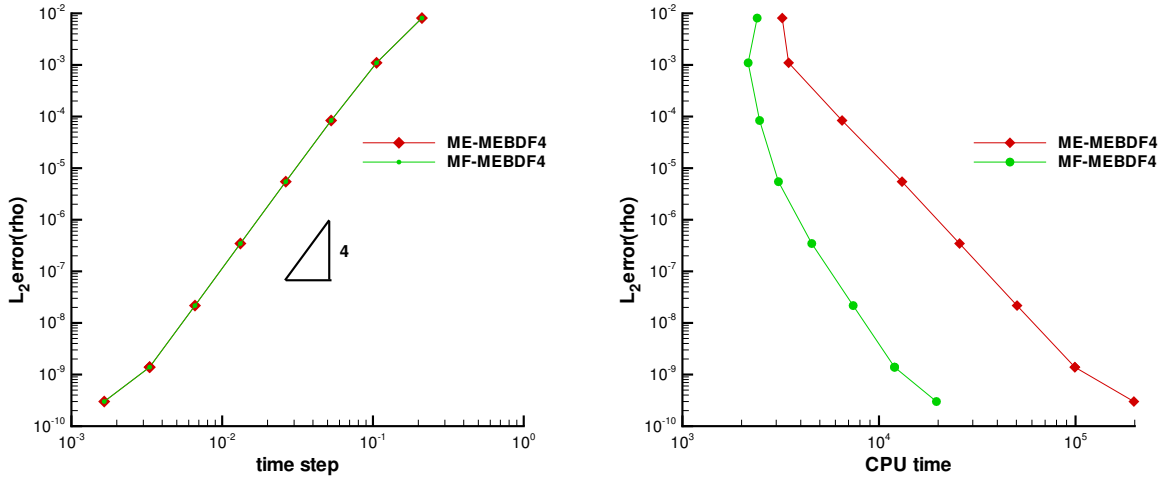


Figure 2: L_2-norm errors of the density field as a function of the time step (left) and of the CPU time (right). Comparison between MF, freezing the preconditioner, and ME algorithms at $M\infty \simeq 1.4$ using $P6$ elements.

				MF-MEBDF4		ME-MEBDF4	
$T/\Delta t$	Tol.$_{Newton}$	L_2 error(ρ)	order	n_{opt}	CPU(s)	CPU(s)	CPUs ratio
40	$8.10 \cdot 10^{-4}$	$8.10 \cdot 10^{-3}$	2.88	1	$2.40 \cdot 10^3$	$3.22 \cdot 10^3$	1.34
80	$1.10 \cdot 10^{-4}$	$1.10 \cdot 10^{-3}$	3.71	2	$2.16 \cdot 10^3$	$3.47 \cdot 10^3$	1.61
160	$8.38 \cdot 10^{-6}$	$8.38 \cdot 10^{-5}$	3.94	4	$2.47 \cdot 10^3$	$6.49 \cdot 10^3$	2.63
320	$5.47 \cdot 10^{-7}$	$5.47 \cdot 10^{-6}$	3.98	11	$3.08 \cdot 10^3$	$1.31 \cdot 10^4$	4.25
640	$3.46 \cdot 10^{-8}$	$3.46 \cdot 10^{-7}$	3.99	26	$4.55 \cdot 10^3$	$2.57 \cdot 10^4$	5.65
1280	$2.17 \cdot 10^{-9}$	$2.17 \cdot 10^{-8}$	3.96	39	$7.39 \cdot 10^3$	$5.04 \cdot 10^4$	6.82
2560	$1.39 \cdot 10^{-10}$	$1.39 \cdot 10^{-9}$	2.21	83	$1.20 \cdot 10^4$	$9.92 \cdot 10^4$	8.27
5120	$3.00 \cdot 10^{-11}$	$3.00 \cdot 10^{-10}$		115	$1.96 \cdot 10^4$	$1.98 \cdot 10^5$	10.10

Table 1: Newton tolerance, L_2-norm error of the density field, order of convergence and CPU time for different numbers of time steps. Comparison between MF, freezing the preconditioner for n_{opt} time steps, and ME algorithms at $M_\infty \simeq 1.4$ using $P6$ elements. In the last column the CPUs ratio between ME and MF is reported.

At stage 2, although no difference exists in the number of non-linear iterations, marked discrepancies appear in the number of linear iterations and Jacobian evaluations. The reason of the equal number of non-linear iterations between the two algorithms is due to the updating of the system matrix in the ME algorithm caused by the lack of/reduced convergence of this Newton loop. This is because at large time steps the system matrix is rapidly varying from stage 1 to stage 2. The significant increase in the number of linear iterations of the MF approach, about 7 times greater than that of its explicit counterpart, is due to a worse preconditioner with respect to the ME approach. In fact, using MF the preconditioner is never recomputed during a time step, unlike what happens in the ME algorithm, where the system matrix updating implies the recomputation of the preconditioner.

At stage 3 ME is less performant, in terms of number of non-linear/linear iterations, than MF because of the less effectiveness of the system matrix remained frozen from stage 2 and

not recomputed in this stage. On the contrary, the MF approach take advantages of the updated matrix-vector product at each Newton iteration and of the preconditioner evaluated at the first predictor stage, showing better convergence properties with respect to ME.

At $\Delta t = T/160$ the system matrix is slowly varying and it remains more effective even if frozen. Thus, by using the ME strategy the system matrix remains unchanged for all the time step without the need to be recomputed. The number of non-linear iterations is roughly the same for both algorithms in all stages. The significant increase in the number of linear iterations of MF with respect to ME is due to the effect of the frozen preconditioner. Note in fact that the number of total Jacobian evaluations for the MF approach is equal to 40 as the preconditioner is updated every 4 time steps. Furthermore, we can notice that the negative effect of the frozen preconditioner is more evident for stage 2 where at each time step the solution is predicted two time steps ahead with respect to the previous known solution at time t_n.

This stage analysis demonstrates that the MF implementation of the MEBDF scheme improves the robustness of the implicit solver. Furthermore, it can be argued that MF outperforms its explicit counterpart since at each time step the ME Jacobian evaluations exceed the cost of the larger number of GMRES iterations plus the additional cost of the function evaluations of the MF algorithm.

		MF-MEBDF4				ME-MEBDF4			
$T/\Delta t$	Stage	non-linear it.	linear it.	func. eval.	Jacob. eval.	non-linear it.	linear it.	func. eval.	Jacob. eval.
40	1	77	346	538	38	76	342	76	39
	2	114	3537	3803	0	114	495	114	38
	3	76	344	534	0	114	518	114	0
160	1	316	942	1732	40	316	632	316	158
	2	337	1438	2270	0	316	632	316	0
	3	316	933	1723	0	316	632	316	0

Table 2: Total number of non-linear iterations, linear iterations, function evaluations and Jacobian evaluations, at each stage and for large and small time steps. Comparison between MF, freezing the preconditioner, and ME.

However, in the above results the recomputed system/preconditioner matrix in the ME approach at $\Delta t = T/40$ and the use of the freezing strategy in the MF approach at $\Delta t = T/160$ do not allow to clarify the effect of the update of the matrix-vector product at each outer iteration on the performance of the implicit solver.

Therefore, the stage analysis is now carried-out at $\Delta t = T/160$ by comparing the performance of the algorithms updating the system/preconditioner matrix, for the ME approach, and the preconditioner matrix, for the MF approach, only at the beginning of stage 1. The data reported in Tab. 3, that refer to one time step iteration, show that at each stage the number of non-linear/linear iterations are almost the same for both algorithms, whereas the number of MF function evaluations is about $4 - 5$ times greater than that of ME. Nevertheless, the CPUs ratio indicates that the matrix free algorithm is slightly more efficient than its explicit counterpart. Thus, it can be concluded that the update of the matrix-vector product at each outer iteration allows to solve more efficiently the non-linear stages, compensating the cost of the additional function evaluations. We remark that this result mainly depends on the conditioning of the system matrix. Further studies are ongoing to investigate this aspect by extending the analysis to stiff systems and other polynomial degrees and by considering the set of primitive variables.

		MF-MEBDF4				ME-MEBDF4				
$T/\Delta t$	Stage	non-linear it.	linear it.	func. eval.	Jacob. eval.	non-linear it.	linear it.	func. eval.	Jacob. eval.	CPUs ratio
160	1	2	4	9	1	2	4	2	1	1.03
	2	2	5	10	0	2	4	2	0	
	3	2	4	9	0	2	4	2	0	

Table 3: Total number of non-linear iterations, linear iterations, function evaluations and Jacobian evaluations at each stage for one small time step. Comparison between MF and ME computing the system matrix only at the beginning of stage 1.

5.1.2 Low Mach numbers: stiff problems

The convection of the isentropic vortex is here investigated for different free stream Mach numbers to highlight pros and cons of the proposed time integration algorithm and its potential in solving non-stiff and stiff systems with respect to the explicit Runge-Kutta schemes. It is in fact well known that RK schemes become inefficient for the solution of stiff-problems, whereas the MEBDF method is well suited for ill-conditioned ODE systems.

In Tab. 4 the results obtained for high and low free stream Mach numbers by using MF-MEBDF4 and RK4 schemes are reported. As the vortex is simply convected, for the RK4 scheme a fixed time step size has been considered. To fairly compare the performance between the two time integration algorithms, at the highest Mach number the maximum cfl number ensuring that the error differs by less than 5% from the lowest achieveable error value ($2.99 \cdot 10^{-10}$) has been chosen for the explicit computations. For the lower Mach numbers the maximum cfl value allowed by the restrictive stability conditions has been used. Furthermore, a fine tuning, up to 2 significant digits, of the cfl number has been performed. For the MF-MEBDF4 scheme the fixed time step has been chosen such that, for the higher Mach numbers, very similar values of the error with respect to the ones obtained by the RK4 scheme are achieved ($\Delta\%$ error $= -2.88\%$ and 2.27% for $M_\infty \simeq 1.4$ and 0.14, respectively). For the lowest Mach number the error is even lower than that achieved by RK4, with $\Delta\%$ error$= -18.53\%$.

It is worth noting that the period of revolution of the vortex T varies according to the free stream velocity as highlighted in Tab. 4, where the number of time steps, $T/\Delta t$, and the time step size, Δt, are both shown. The last column of Tab. 4, where the ratio between the CPUs required by RK4 and MF-MEBDF4 is reported, highlights that RK4 largely outperforms MF-MEBDF4 for $M_\infty \simeq 1.4$, 0.14 (CPUs ratio = 0.03, 0.10). However, at $M_\infty \simeq 0.014$ the proposed time integration algorithm outperforms the explicit scheme with a CPUs ratio equal to 1.31, despite the lower error value computed.

	MF-MEBDF4					RK4					
M_∞	$T/\Delta t$	Δt	L_2 error(ρ)	n_{opt}	CPU (s)	$T/\Delta t$	Δt	L_2 error(ρ)	CPU (s)	$\Delta\%$ error	CPUs ratio
1.4	5000	$1.69 \cdot 10^{-3}$	$3.03 \cdot 10^{-10}$	100	$1.94 \cdot 10^{4}$	2381	$3.55 \cdot 10^{-3}$	$3.12 \cdot 10^{-10}$	$6.51 \cdot 10^{2}$	-2.88%	0.03
0.14	4000	$2.11 \cdot 10^{-2}$	$5.85 \cdot 10^{-10}$	50	$2.04 \cdot 10^{4}$	7875	$1.07 \cdot 10^{-2}$	$5.72 \cdot 10^{-10}$	$2.09 \cdot 10^{3}$	2.27%	0.10
0.014	640	1.32	$4.09 \cdot 10^{-7}$	4	$1.63 \cdot 10^{4}$	78741	$1.07 \cdot 10^{-2}$	$5.02 \cdot 10^{-7}$	$2.14 \cdot 10^{4}$	-18.53%	1.31

Table 4: Number of time steps, time step value, L2-norm error of the density field and CPU time obtained with MF-MEBDF4, freezing the preconditioner for n_{opt} time steps, and RK4 by using $P6$ elements and different M_∞ values. In the last two columns the percentage relative difference of the errors the and the CPUs ratio between RK4 and MF-MEBDF4 are reported.

5.2 Laminar manufactured solution

For viscous flows, the proposed time integration scheme has been evaluated on the following Manufactured Solution [35]

$$\rho(x,y,t) = \rho_0 \left[\sin\left(x^2+y^2+\omega t\right)+1.5\right] \quad (27)$$
$$u(x,y,t) = u_0 \left[\sin\left(x^2+y^2+\omega t\right)+\epsilon\right] \quad (28)$$
$$v(x,y,t) = v_0 \left[\cos\left(x^2+y^2+\omega t\right)+\epsilon\right] \quad (29)$$
$$e_t(x,y,t) = e_{t0} \left[\cos\left(x^2+y^2+\omega t\right)+1.5\right] \quad (30)$$

In the above equations x and y are the spatial coordinates of a generic point in the computational domain. Furthermore, the constants of Eqs. (27–30) are defined as: $\rho_0 = 1$, $u_0 = 0.1$, $v_0 = 0.001$, $e_{t0} = 1$, $\epsilon = 1/2$ and $\omega = 2\pi$.

The computational domain is a square of side one, the boundary conditions are time-dependent dirichlet boundary conditions and the simulation are performed up to a final time equal to one period T for progressively finer time step by using $P6$ elements. The numerical solution, starting from $t_0 = 0$, evolves in time according to the exact solution imposed by adding a source term to the Navier-Stokes equations. The reader is referred to the Appendix D of [35] for the vector of source terms.

The results obtained by the proposed time integration scheme are evaluated for different discretizations of the computational domain by using uniform and stretched cartesian grids. In particular, the cartesian grid used are: $20 \times 20; 5 \times 80; 2 \times 200$. These grids have the same total number of elements but are characterized by different Aspect Ratio (AR): $1; 16; 100$.

5.2.1 Cartesian elements: non-stiff problems

The results obtained by performing a temporal refinement study on the cartesian uniform grid ($AR = 1$) by using the MF approach are here presented and compared with the the results obtained with the ME approach.

In the left plot of Fig. 3 the L_2-norm error of the density field as a function of the time step is presented while in the right plot the work-precision characteristics are shown. As for the inviscid test case, the convergence histories obtained using the two algorithms are indistinguishable and achieve the design order convergence of 4. Furthermore, the performance advantages of the MF approach with respect to the ME approach are remarkable and become increasingly more significant as the time step reduces. As it can be seen form Tab. 5 , where the above results are summarized, the MF algorithm is 2.76-7.84 times faster than the ME approach.

Finally, we remark that the analysis carried out to investigate the influence of both the MF approach and the freeze strategy, on the performance of the implicit solver, is not presented since the results are similar to those shown for the inviscid test case.

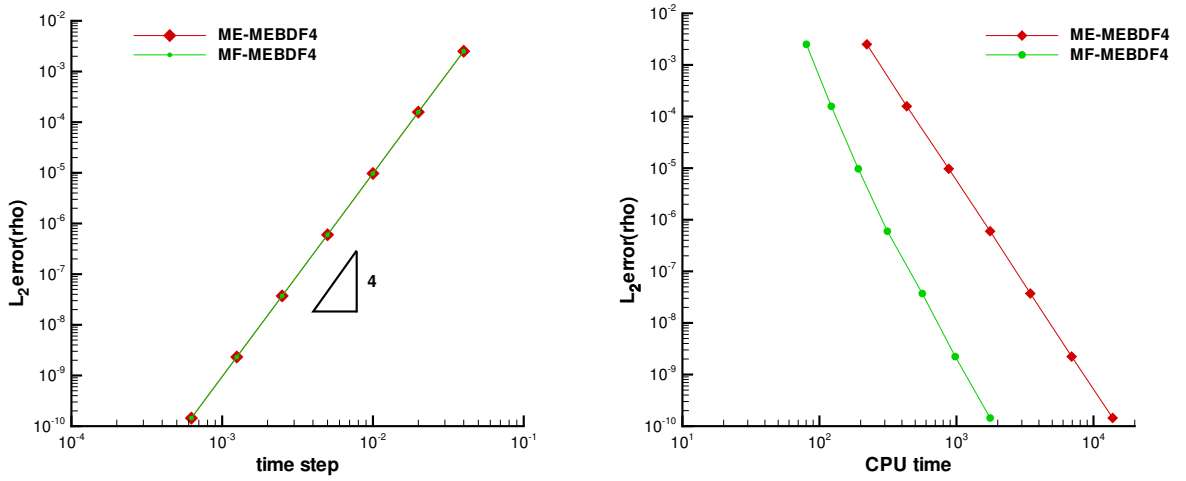


Figure 3: L_2-norm errors of the density field as a function of the time step (left) and of the CPU time (right). Comparison between MF, freezing the preconditioner, and ME algorithms on the cartesian grid with $AR = 1$ using $P6$ elements

				MF-MEBDF4		ME-MEBDF4	
$T/\Delta t$	Tol.$_{Newton}$	L_2 error(ρ)	order	n_{opt}	CPU(s)	CPU(s)	CPUs ratio
25	$2.51 \cdot 10^{-4}$	$2.51 \cdot 10^{-3}$	3.99	5	$8.03 \cdot 10^{1}$	$2.22 \cdot 10^{2}$	2.76
50	$1.58 \cdot 10^{-5}$	$1.58 \cdot 10^{-4}$	4.02	10	$1.22 \cdot 10^{2}$	$4.34 \cdot 10^{2}$	3.56
100	$9.71 \cdot 10^{-7}$	$9.71 \cdot 10^{-6}$	4.02	20	$1.92 \cdot 10^{2}$	$8.79 \cdot 10^{2}$	4.58
200	$5.99 \cdot 10^{-8}$	$5.99 \cdot 10^{-7}$	4.01	35	$3.13 \cdot 10^{2}$	$1.76 \cdot 10^{3}$	5.62
400	$3.72 \cdot 10^{-9}$	$3.72 \cdot 10^{-8}$	4.00	67	$5.63 \cdot 10^{2}$	$3.46 \cdot 10^{3}$	6.15
800	$2.32 \cdot 10^{-10}$	$2.32 \cdot 10^{-9}$	4.01	140	$9.79 \cdot 10^{2}$	$6.93 \cdot 10^{3}$	7.08
1600	$1.44 \cdot 10^{-11}$	$1.44 \cdot 10^{-10}$		200	$1.76 \cdot 10^{3}$	$1.38 \cdot 10^{4}$	7.84

Table 5: Newton tolerance, L_2-norm error of the density field, order of convergence and CPU time for different numbers of time steps. Comparison between MF, freezing the preconditioner for n_{opt} time steps, and ME algorithms on the grid with $AR = 1$ using $P6$ elements. In the last column the CPUs ratio between ME and MF is reported.

5.3 Stretched elements: stiff problems

The comparison between the performance of MF-MEBDF4 and RK4 algorithms is reported in Tab. 6. Unlike the inviscid test case, the shape of the solution in this case is varying over time. Therefore, to fairly compare the performance of the two time integration schemes, the RK4 time integration uses a variable time step.

For the uniform grid case ($AR = 1$), according to Subsection 5.1.2, the maximum cfl number ensuring that the error differs by less than 5% from the lowest achievable error value ($1.44 \cdot 10^{-10}$) has been chosen to not penalize the performance of the explicit computation. For the higher aspect ratios ($AR = 16$ and $AR = 100$) the maximum cfl numbers ensuring stability have been used. We remark that, also in this case the cfl number has been tuned up to 2 significant digits. For the computations that refer to the MF-MEBDF4 algorithm the

largest time step values that ensure almost the same error levels of the corresponding RK4 solutions have been adopted ($\Delta\%$ error $= -4.00\%$, -0.33% and 0.47% for $AR = 1$, 16 and 100, respectively). The entries of the last column of Tab. 6, where the ratio between the CPUs required by MF-MEBDF4 and RK4 is reported, show that the RK4 scheme outperforms the MF-MEBDF4 scheme on the uniform cartesian grid (CPUs ratio = 0.49). However, increasing the ill-conditioning by increasing the aspect ratio, the severe time step restriction of the explicit scheme leads to an inefficient time integration, that becomes more expensive as the stiffness increases. At the intermediate AR level of 16 the MF-MEBDF4 becomes 1.10 times faster than RK4. Increasing the aspect ratio to 100 the remarkable speed-up of 18.64 is achieved.

	MF-MEBDF4				RK4				
AR	$T/\Delta t$	L_2 error(ρ)	n_{opt}	CPU (s)	$T/\Delta t_{ave}$	L_2 error(ρ)	CPU (s)	$\Delta\%$ error	CPUs ratio
1	1600	$1.44 \cdot 10^{-10}$	200	$1.76 \cdot 10^{3}$	3273	$1.50 \cdot 10^{-10}$	$8.66 \cdot 10^{2}$	-4.00%	0.49
16	1600	$9.05 \cdot 10^{-10}$	58	$1.73 \cdot 10^{3}$	7122	$9.08 \cdot 10^{-10}$	$1.90 \cdot 10^{3}$	-0.33%	1.10
100	400	$4.30 \cdot 10^{-7}$	28	$5.58 \cdot 10^{2}$	38532	$4.28 \cdot 10^{-7}$	$1.04 \cdot 10^{4}$	0.47%	18.64

Table 6: Number of time steps, L2-norm error of the density field and CPU time obtained with MF-MEBDF4, freezing the preconditioner for n_{opt} time steps, and RK4 by using $P6$ elements and cartesian grids with different aspect ratio (AR). In the last two columns the percentage relative difference of the errors and the CPUs ratio between RK4 and MF-MEBDF4 are reported.

6 CONCLUSIONS

We have presented a matrix-free (MF) preconditioned Newton/GMRES method for the solution of the unsteady compressible Navier-Stokes equations. To ensure high accuracy and efficiency of the proposed implicit method, the three-stage fourth-order accurate MEBDF time integration scheme has been coupled with a high-order accurate DG discretization. The MF solver uses less computer memory than the corresponding matrix-explicit (ME) version since the system matrix is stored only for preconditioning purpose. Furthermore, MF-MEBDF outperforms its explicit counterpart for the problems here investigated. In particular, MF-MEBDF is more robust and faster than ME-MEBDF for large time steps. In these cases, in fact, ME-MEBDF shows a lack of/reduced convergence and requires the time consuming updating of the system matrix. At small time steps, when the system matrix is slowly varying and remains effective for all non-linear stages, MF-MEBDF slightly outperforms ME-MEBDF, although for the matrix-free computations a larger number of function evaluations with respect to ME are required. These findings allow to conclude that the MF-MEBDF algorithm benefits from the inherent update of the matrix-vector product at each Newton iteration. This mainly depends on the multi-stage structure of the time scheme and on the conditioning of the system matrix. We remark that a seventh-order accurate DG discretization has been adopted in our computations, leading to ill-conditioned stage matrices. Thus, a better performance is expected to be achieved when additional sources of ill-conditioning are considered, such as low speed flows, RANS equations and stretched grids. The proposed MF algorithm, equipped with a preconditioner freeze strategy, shows a noteworthy computational time reduction with respect to ME for non-stiff ODE systems, with an increasing CPUs ratio, ranging from 1.34 to 10.10, as the time step reduces. Furthermore, MF-MEBDF outperforms the five-stage fourth-order accurate Runge-Kutta scheme in solving stiff systems, with an increasing saving of computational time as the stiffness increases. For the inviscid vortex transport test-case, the CPUs ratio of 1.31 is achieved at $M \simeq 0.014$. For the laminar test case the speed up ratio significantly improves with

the grid aspect ratio from 1.10 ($AR = 16$) to 18.64 ($AR = 100$).

Further work is ongoing to analyze the performance of the solver for different polynomial degrees and by considering the set of primitive variables.

REFERENCES

[1] Cockburn B., Karniadakis G.E., Shu C.W. Eds. (2000): Discontinuous Galerkin Methods. Theory, Computation and Applications. New York: Springer Verlag, vol. 11, Lect. Notes Comput. Sci. Eng.

[2] Bassi F., Rebay S. (1997): A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. J. Comput. Phys. 131, 267–279.

[3] Luo H., Baum J.D., Lohner R. (2006): A p-multigrid discontinuous Galerkin method for the Euler equations on unstructured grids. J. Comput. Phys. 211, 767–783.

[4] Cockburn B. (2001): Devising discontinuous Galerkin methods for non-linear hyperbolic conservation laws. J. Comput. Appl. Math. 128, 187–204.

[5] Nigro A., Renda S., De Bartolo C., Hartmann R., Bassi F. (2013): A high-order accurate discontinuous Galerkin finite element method for laminar low Mach number flows. Internat. J. Numer. Methods Fluids 72 (1), 43–68.

[6] Bassi F., Botti L., Colombo A., Ghidoni A., Massa F. (2015): Linearly implicit Rosenbrock-type Runge-Kutta schemes applied to the Discontinuous Galerkin solution of compressible and incompressible unsteady flows. Computer & Fluids 118, 305–320.

[7] Nigro A., Ghidoni A., Bassi F., Rebay S. (2014): Modified extended BDF scheme for the Discontinuous Galerkin solution of unsteady compressible flows. Int. J. Numer. Methods Fluids 76, 549–574.

[8] Nigro A., De Bartolo C., Bassi F., Ghidoni A. (2014): Up to sixth-order accurate A-stable implicit schemes applied to the Discontinuous Galerkin discretized Navier-Stokes equations. J. Comput. Phys. 276, 136–162.

[9] Gear C.W., Saad Y. (1983): Iterative solution of linear equations in ODE codes. SIAM Journal on Scientific and Statistical Computing 4, 583–601.

[10] Chan T.F., Jackson K.R. (1984): Nonlinearly preconditioned Krylov subspace methods for discrete Newton algorithms. SIAM Journal on Scientific and Statistical Computing. 5, 533–542.

[11] Brown P.N., Hindmarsh A.C. (1986): Matrix-free methods for stiff systems of ODEs. SIAM Journal on Numerical Analysis. 23, 610–638.

[12] Johan J. (1992): Data parallel finite element techniques for large scale computational fluid dynamics. PhD thesis, Department of Mechanical Engineering, Stanford University, U.S.A, (Chapter 2).

[13] McHugh P.R., Knoll D.A. (1994): Comparison of standard and matrix-free implementations of several Newton-Krylov solvers. AIAA Journal. 32, 2394–2400.

[14] McHugh P.R., Knoll D.A. (1993): Inexact Newton's method solutions to the incompressible Navier-Stokes and energy equations using standard and matrix-free implementations. AIAA Paper, 93–3332.

[15] Rasetarinera P., Hussaini M.Y. (2001): An efficient implicit discontinuous spectral Galerkin method. J Comput Phys. 172(2), 718–738.

[16] Hillewaert K., Remacle J.F., Chevaugeon N., Bernard P.E., Geuzaine P. (2006): Analysis of a hybrid P-multigrid method for the discontinuous Galerkin discretisation of the Euler equation. In: Wesseling P., Onate E., Periaux J. (Eds.), ECCOMAS CFD Conference2006, TU Delft, The Netherlands.

[17] Crivellini A., Bassi F. (2011): An implicit matrix-free Discontinuous Galerkin solver for viscous and turbulent aerodynamic simulations. Computer & Fluids. 50, 81–93

[18] Hosseini S.M., Hojjati G. (1999): Matrix free MEBDF method for the solution of stiff systems of ODEs. Mathematical and Computer Modelling 29(4), 67–77

[19] Tesini P. (2008): An h-Multigrid Approach for High-Order Discontinuous Galerkin Methods. Ph.D. thesis, University of Bergamo, Bergamo, Italy.

[20] Bassi F., Botti L., Colombo A., Di Pietro D.A., Tesini P. (2012): On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. J. Comput. Phys. 231 (1), 45–65.

[21] Di Pietro D.A., Ern A. (2012): Mathematical Aspects of Discontinuous Galerkin Methods. Vol. 69 of Mathematiques et Applications, Springer-Verlag.

[22] Bassi F., Rebay S., Mariotti G., Pedinotti S., Savini M. (1997): A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In: R. Decuypere and G. Dibelius (Eds.), Proceeding of the 2nd European Conference on turbomachinery Fluid Dynamics and Thermodynamics, pp. 99-108, Antwerpen, Belgium, Technologisch Instituut

[23] Bassi F., Rebay S. (2000): A high order discontinuous Galerkin method for compressible turbulent flows. In: Discontinuous Galerkin Methods: Theory, Computation and Applications, B. Cockburn, G. Karniadakis, and C.-W. Shu (Eds.), Springer, 11: 77–88.

[24] Brezzi F., Manzini M., Marini D., Pietra P., Russo A. (2000): Discontinuous Galerkin approximations for elliptic problems. Numer. Methods Partial Differential Equations, 16: 365–378.

[25] Arnold D.N., Brezzi F., Cockburn B., Marini D.(2002): Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal., 39(5): 1749–1779.

[26] Nigro A., Ghidoni A., Rebay S., Bassi F. (2012): High-order discontinuous Galerkin solution of unsteady problems using Modified Extended Backward Differentiation Formulae. In: 6th European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS 2012, pp. 6672–6690.

[27] Cash J.R. (1983): The integration of stiff initial value problems in ODEs using Modified Extended Backward Differentiation Formulae. Computers & Mathematics with Applications, 5 (9) pp. 645–657.

[28] Cash J.R. (2000): Modified extended backward differentiation formulae for the numerical solution of stiff initial value problems in ODEs and DAEs. Journal of Computational and Applied Mathematics, 125(1-2), pp. 117–130.

[29] Balay S., Abhyankar S., Adams M.F., Brown J., Brune P., Buschelman K., Dalcin L., Eijkhout V., Gropp W.D., Kaushik D., Knepley M.G., McInnes L.C., Rupp K., Smith B.F., Zampini S., Zhang H.(2016): Petsc-web-page: http://www.mcs.anl.gov/petsc

[30] Pernice M., Walker H.F. (1998): NITSOL: a Newton iterative solver for nonlinear systems. SIAM J. Sci . Comput. 19, 302–318

[31] Spiteri R.J., Ruuth S.J. (2002): A new class of optimal high-order strong-stability-preserving time discretization methods. SIAM J. Numer. Anal. 40(2), 469–491.

[32] Gottlieb J.J., Groth C.P.T. (1988): Assessment of Riemann solvers for unsteady one-dimensional inviscid flows of perfect gases. J. Comput. Phys. 78, 437–458.

[33] Nigro A., De Bartolo C., Bassi F., Ghidoni A. (2014): High-order discontinuous Galerkin solution of unsteady flows by using an advanced implicit method. Lecture Notes in Computational Science and Engineering 99, 135–149.

[34] Bassi F., Botti L., Colombo A., Crivellini A., Ghidoni A., Nigro A., Rebay S. (2015): Time integration in the Discontinuous Galerkin code MIGALE - Unsteady Problems. In: Kroll N., Hirsch C., Bassi F., Johnston C., Hillewaert K.(Eds.), IDIHOM: Industrialization of high-order methods - A top-down approach. Vol. 128 of Notes on numerical fluid mechanics and multidisciplinary design, Springer International Publishing, pp. 205–230.

[35] Salari K., Knupp. P. (2000): Code Verification by the Method of Manufactured Solutions. SAND 2000-1444, Sandia National Laboratories, Albuquerque, NM.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

PRECONDITIONERS FOR ISOGEOMETRIC ANALYSIS BASED ON SOLVERS FOR THE SYLVESTER EQUATION

Giancarlo Sangalli[1] **and Mattia Tani**[1]

[1] Università di Pavia, Dipartimento di Matematica "F. Casorati"
Via A. Ferrata 1, 27100 Pavia, Italy
e-mail: {sangia05,mattia.tani}@unipv.it

Keywords: Isogeometric analysis, preconditioning, Kronecker product, Sylvester equation.

Abstract. *In this work, which represents a condensed version of [1], we consider large linear systems arising from the isogeometric discretization of the Poisson problem on a single-patch domain. We consider a preconditioning strategy which is based on the solution of a Sylvester-like equation at each step of an iterative solver. This strategy, which fully exploits the tensor structure that underlies isogeometric problems, is robust with respect to both mesh size and spline degree. The application of the preconditioner is performed by a popular direct solver for the Sylvester equation, whose implementation details are given in the 2D and 3D case, with particular emphasis on the latter. We show numerical experiments for 3D problems, which demonstrate the potential of this approach.*

1 Introduction

The present manuscript represents a condensed version of [1]. We are concerned with the problem of solving large linear systems arising from isogeometric discretizations [2]. An extension of the classical finite element method, the isogeometric method based on the idea of using splines or other functions constructed from splines (e.g., non-uniform rational B-splines, NURBS) both for the parametrization of the computational domain, as it is typically done by computer aided design software, and for the representation of the unknown solution fields of the PDE of interest. Unlike standard finite element methods, the isogeometric method makes possible to use high-regularity functions. The so-called isogeometric k-method, based on splines of degree p and global C^{p-1} regularity, has shown significant advantages in term of higher accuracy per degree-of-freedom in comparison to C^0 finite elements of degree p [3, 4].

The study of the computational efficiency of linear solvers for isogeometric discretizations has been initiated in the papers [5] and [6], where it has been shown that the algorithms used with the finite element method suffer of performance degradation when used used in the context of the isogeometric k-method. Since then, a number of papers have been published that analyse and develop preconditioners for isogeometric linear systems, see e.g. [7, 8, 9, 10]. While containing important advances, these papers also confirms the difficulty in achieving both *robustness* and computational *efficiency* for the high-degree k-method.

Robust multigrid preconditioners have been proposed in the recent papers [11] and [12]. We emphasize that, in both works, the authors exploit the tensor-product structure of isogeometric spaces in order to achieve efficiency in multidimensional problems.

In this paper we also exploit the tensor-product structure of multivariate spline space, on a different basis. We consider the Bartels-Stewart direct solver, a method that has been developed for the so-called Sylvester equation, and discuss its application as a preconditioner for isogeometric systems. We describe this method in both 2D and 3D cases. We then focus on 3D problems, discussing its computational cost and showing numerical experiments which assess the potential of this approach.

In 3D, application of the Bartels-Stewart solver requires $O(N^{4/3})$ FLOPs. However, in all our benchmarks that uses a conjugate gradient (CG) iterative solver, the computational time spent in the Bartels-Stewart preconditioner application is even lower than the residual computation (multiplication matrix $\mathcal{A}$ times a vector). This surprising performance is due to the fact that the Bartels-Stewart solver requires dense matrix-matrix operations that takes advantages of modern computer architecture. Furthermore, the Bartels-Stewart method is especially suited to parallelisation which may significantly speed up the execution time, though this is not considered in our analysis.

We emphasize that the proposed approach is robust, in the sense that the condition number of the preconditioned system is uniformly bounded with respect to the degree p and mesh size h when the parametrization is regular. In all cases, it is important to have strategies to further improve the condition number, and this will be the topic of further researches.

2 The problem

We consider, as a model problem, the Poisson equation with Dirichlet boundary conditions:

$$\begin{cases} -\mathrm{div}(K(\mathbf{x})\nabla \mathrm{u}(\mathbf{x})) = f(\mathbf{x}) \quad \text{on} \quad \Omega \subseteq \mathbb{R}^d \\ \qquad\qquad\qquad\qquad \mathrm{u} = 0 \quad \text{on} \quad \partial\Omega \end{cases} \tag{1}$$

where $K(\mathbf{x})$ is a symmetric positive definite matrix for each $\mathbf{x} \in \Omega$. In isogeometric methods, Ω is given by a spline or NURBS parametrization. For the sake of simplicity, we consider a single-patch spline parametrization $\boldsymbol{F}$ which satisfies $\Omega = \boldsymbol{F}\left([0,1]^d\right)$.

Following the isogeometric approach, we consider the Galerkin discretization of (1) relative to a space of multivariate B-splines on Ω. We recall that multivariate B-splines in dimension d ($d = 2, 3$ are the interesting cases) are defined from univariate B-splines by tensorization. We assume for simplicity that the spline degree p and the number n of univariate basis functions which vanish at the boundary is the same in all directions. We are primarily interested in the so called k-refinement or isogeometric k-method [2], hence we consider C^{p-1} continuous splines, i.e. splines with maximum regularity.

Let $N = n^3$ and let $B_1, \dots, B_N$, denote the basis function of the isogeometric space, which incorporate the homogeneous Dirichlet boundary conditions. Then the Galerkin stiffness matrix reads

$$\begin{aligned} \mathcal{A}_{ij} &= \int_\Omega \left(\nabla B_i(\boldsymbol{x})\right)^T K(\boldsymbol{x}) \nabla B_j(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{[0,1]^d} \left(\nabla \hat{B}_i(\boldsymbol{\xi})\right)^T Q(\boldsymbol{\xi})\, \nabla \hat{B}_j\left(\boldsymbol{\xi}\right)\, d\boldsymbol{\xi}, \qquad i,j = 1, \dots, N \end{aligned} \tag{2}$$

where

$$Q = \det\left(J_{\boldsymbol{F}}\right) J_{\boldsymbol{F}}^{-T} K J_{\boldsymbol{F}}^{-1} \tag{3}$$

and $J_{\boldsymbol{F}}$ denotes the Jacobian of $\boldsymbol{F}$.

The support of each basis spline that does not touch $\partial\Omega$ intersects the support of $(2p+1)^d$ basis splines (including itself), while the support of a basis spline intersecting $\partial\Omega$ overlaps at least $(p+1)^d$ and up to $(2p+1)^d$ basis splines supports (including itself). Thus, the number of nonzeros of $\mathcal{A}$ is about $(2p+1)^d N$.

3 The preconditioner

Consider the matrix

$$\mathcal{P}_{ij} = \int_{[0,1]^d} \left(\nabla \hat{B}_i\right)^T \nabla \hat{B}_j \, d\xi, \qquad i,j = 1, \dots, N \tag{4}$$

Observe that $\mathcal{P} = \mathcal{A}$ in the special case when K is the identity matrix and $\boldsymbol{F}$ is the identity function, which means that $\Omega = [0,1]^d$.

By exploiting the tensor product structure of the basis functions, the matrix $\mathcal{P}$ can be expressed as the sum of d Kronecker products. For example, when $d = 2$ we have

$$\mathcal{P} = K_1 \otimes M_2 + M_1 \otimes K_2$$

where M_1, M_2 represent the mass, and K_1, K_2 the stiffness univariate matrices. Such matrices are all symmetric positive definite and banded with bandwidth p (we say that a matrix B has bandwidth p if $B_{ij} = 0$ for $|i-j| > p$). These matrices have the same order n. Similarly, when $d = 3$

$$\mathcal{P} = K_1 \otimes M_2 \otimes M_3 + M_1 \otimes K_2 \otimes M_3 + M_1 \otimes M_2 \otimes K_3.$$

By comparing (2) and (4), observe that $\mathcal{P}_{ij} \neq 0$ if and only if $\mathcal{A}_{ij} \neq 0$. Thus, despite having different entries in general, $\mathcal{A}$ and $\mathcal{P}$ have the same sparsity pattern.

We propose $\mathcal{P}$, defined in (4), as a preconditioner for the isogeometric matrix $\mathcal{A}$. In other words, we want to precondition a problem with arbitrary geometry and coefficients with a solver for the same operator on the parameter domain, with constant coefficients. This is a common approach, see e.g. [13], [11] and [12].

It is known (see e.g. [1] for a proof) that

$$\kappa\left(\mathcal{P}^{-1}\mathcal{A}\right) \leq \frac{\sup_{\Omega} \lambda_{\max}\left(Q\right)}{\inf_{\Omega} \lambda_{\min}\left(Q\right)} \tag{5}$$

where the matrix Q is given in (3).

As long as the considered problem doesn't depart much from the model problem on the square with constant coefficients, the right-hand side of (5) will be small and the preconditioner is expected to perform well. On the other hand, if the eigenvalues of Q vary widely, due to the presence of complicated geometry or coefficients, the preconditioner performance decreases. In these cases, it is useful to have strategies to improve the spectral conditioning of $\mathcal{P}^{-1}\mathcal{A}$: this is a topic that we will address in a forthcoming paper. We emphasize that bound (5) does not depend neither on the mesh size nor on the spline degree, but only on $\boldsymbol{F}$ and K.

4 The Bartels-Stewart method

If the system $\mathcal{A}u = b$ is solved using the Preconditioned Conjugate Gradient (PCG) method, at each iteration we need to solve a system of the form

$$\mathcal{P}s = r \tag{6}$$

where r is the current residual. The computational effort required for solving this system (exactly or approximately) is of paramount importance in assessing the overall performance of PCG.

When $d = 2$, equation (6) can be reformulated as

$$K_1 S M_2 + M_1 S K_2 = R \tag{7}$$

where $r = \text{vec}(R)$ and $s = \text{vec}(S)$, with the vec operator stacking the rows of a matrix $X \in \mathbb{R}^{n\times n}$ into a single column vector $x \in \mathbb{R}^{n^2}$. Equation (7) takes the name of (generalized) Sylvester equation. Due to its many applications, the literature dealing with Sylvester equation (and its variants) is vast, and a number of methods have been proposed for its numerical solution. We refer to [14] for a recent survey on this subject.

In this paper we consider the Bartels-Stewart method [15], which represents the state-of-the-art direct solver for (7). In fact, the method considered here departs slightly from the standard Bartels-Stewart solvers in that it exploits the symmetry of the coefficient matrices M_1, M_2, K_1, K_2, and was first presented in [16]. However, due to its popularity in the context of the Sylvester equation, we still refer to this approach as the Bartels-Stewart method. We remark that in [1] an iterative approach, namely the Alternating Direction Implicit (ADI) method is also considered. However, it yields a worst performance, especially for 3D problems, than the direct method, and hence we do not discuss it here.

We first describe the Bartels-Stewart method in the 2D case, where the system (6) reads

$$\left(K_1 \otimes M_2 + M_1 \otimes K_2\right) s = r \tag{8}$$

We consider the generalized eigendecomposition of the matrix pencils (K_1, M_1) and (K_2, M_2), namely

$$K_1 U_1 = M_1 U_1 D_1 \qquad K_2 U_2 = M_2 U_2 D_2 \tag{9}$$

where D_1 and D_2 are diagonal matrices whose entries are the eigenvalues of $M_1^{-1} K_1$ and $M_2^{-1} K_2$, respectively, while U_1 and U_2 satisfy

$$U_1^T M_1 U_1 = I, \qquad U_2^T M_2 U_2 = I,$$

which implies in particular $U_1^{-T} U_1^{-1} = M_1$ and $U_2^{-T} U_2^{-1} = M_2$, and also, from (9), $U_1^{-T} D_1 U_1^{-1} = K_1$ and $U_2^{-T} D_2 U_2^{-1} = K_2$. Therefore we factorize $\mathcal{P}$ in (8) as follows:

$$(U_1 \otimes U_2)^{-T} (D_1 \otimes I + I \otimes D_2) (U_1 \otimes U_2)^{-1} s = r,$$

and adopt the following strategy:

Algorithm 1 Bartels-Stewart direct method (symmetric case)

1: Compute the generalized eigendecompositions (9)
2: Compute $\tilde{r} = (U_1 \otimes U_2)^T r$
3: Compute $\tilde{s} = (D_1 \otimes I + I \otimes D_2)^{-1} \tilde{r}$
4: Compute $s = (U_1 \otimes U_2) \tilde{s}$

This approach can be generalized to the 3D case, where (6) takes the form

$$(K_1 \otimes M_2 \otimes M_3 + M_1 \otimes K_2 \otimes M_3 + M_1 \otimes M_2 \otimes K_3)\, s = r \tag{10}$$

in a straightforward way. We consider the generalized eigendecompositions

$$K_1 U_1 = M_1 U_1 D_1, \qquad K_2 U_2 = M_2 U_2 D_2, \qquad K_3 U_3 = M_3 U_3 D_3 \tag{11}$$

with D_1, D_2, D_3 diagonal matrices and

$$U_1^T M_1 U_1 = I, \qquad U_2^T M_2 U_2 = I, \qquad U_3^T M_3 U_3 = I.$$

Then, (10) can be factorized as

$$(U_1 \otimes U_2 \otimes U_3)^{-1} (D_1 \otimes I \otimes I + I \otimes D_2 \otimes I + I \otimes I \otimes D_3) (U_1 \otimes U_2 \otimes U_3)^{-T} s = r,$$

which suggests the following algorithm.

Algorithm 2 3D Bartels-Stewart method (symmetric case)

1: Compute the generalized eigendecompositions (11)
2: Compute $\tilde{r} = (U_1 \otimes U_2 \otimes U_3) r$
3: Compute $\tilde{s} = (D_1 \otimes I \otimes I + I \otimes D_2 \otimes I + I \otimes I \otimes D_3)^{-1} \tilde{r}$
4: Compute $s = (U_1 \otimes U_2 \otimes U_3)^T \tilde{s}$

4.1 Computational cost for the 3D approach

We discuss in detail the computational cost of the Bartels-Stewart method only in the 3D case, i.e. Algorithm 2, as this represents the focus of the present manuscript. The exact cost of the eigendecompositions in line 1 depends on the algorithm employed, however it is safe to assume that this step costs $O(n^3)$ FLOPs. We further mention that when a divide-and-conquer approach is used [17, Chapter 8], the eigendecomposition can be naturally performed in parallel computer architectures. Moreover, when Algorithm 2 is applied as a preconditioner for CG, then this step may be performed only once, since the matrices involved do not change throughout the CG iteration. Line 3 is just a diagonal scaling, and it also requires $O(n^3)$ FLOPs. Lines 2 and 4, as it can be seen by exploiting the properties of Kronecker products [18], are equivalent to performing a total of 6 products between dense matrices of size $n \times n$ and $n \times n^2$. Thus, neglecting lower order terms the overall computational cost of Algorithm 2 is $12n^4$ FLOPs. We emphasize that this cost is independent of p.

We point out that the main computational effort of the method consists in a few (dense) matrix-matrix products, which are level 3 BLAS operations and typically yield high efficiency thanks to a dedicated implementation on modern computers by optimised usage of the memory cache hierarchy [17, Chapter 1]. Matrix-matrix products are also naturally suited for parallelisation. Second, in a preconditioned CG iteration the cost for applying the preconditioner has to be compared with the cost of the residual computation (a matrix-vector product with $\mathcal{A}$) which can be quantified in approximately $2(2p+1)^3n^3$ for 3D problems, resulting in a FLOPs ratio of the preconditioner application to residual computation (matrix-vector product) of $(3n)/(4p^3) \approx n/p^3$. For example, if $N = 256^3$ and $p = 4$, the preconditioner requires only 3 times more FLOPs than the residual computation, while for degree $p = 6$ the matrix-vector product is even more costly than the preconditioner itself. However in numerical tests we will see that, for all cases of practical interest in 3D, the computational time used by the preconditioner application is far lower that the residual computation itself. This is because the computational time depends not only on the FLOPs count but also on the memory usage and, as mentioned above, dense matrix-matrix multiplications greatly benefit of modern computer architecture.

5 3D numerical experiments

We now numerically show the potential of the Bartels-Stewart method in solving isogeometric systems. We only show 3D experiments here, and refer to [1] for experiments in the 2D case.

Algorithm 2 was implemented in MATLAB Version 8.5.0.197613 (R2015a), with the toolbox GEOPDES [19], on a 12x Intel Xeon i7-5820K, 3.30GHz, 64 GB of RAM. We used the MATLAB function `eig` to compute the generalized eigendecomposition (11), and the products involving Kronecker matrices (lines 2 and 4 of Algorithm 2) were performed using the function from the free MATLAB toolbox Tensorlab [20]. We recall MATLAB allows implicit parallelism for some of its operations, e.g. dense matrix multiplication and `eig`.

We first consider a domain with trivial geometry, namely the unit cube $[0,1]^3$, and then turn to more complicated domains. The first is a thick quarter of ring (Figure 1); note this solid has a trivial geometry on the third direction. The second complicated domain is the solid of revolution obtained by a 2D quarter of ring (Figure 2). Specifically, we performed a $\pi/2$ revolution around the axis having direction $(0,1,0)$ and passing through $(-1,-1,-1)$. We emphasize that here the geometry is nontrivial along all directions. In the cube case, we set $b =$`randn`$(n^3,1)$

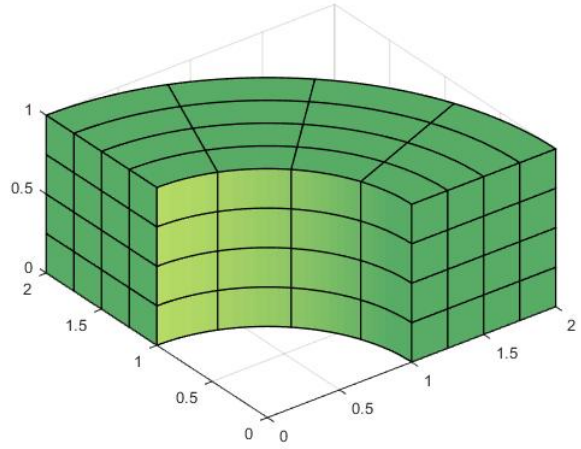


Figure 1: Thick ring domain

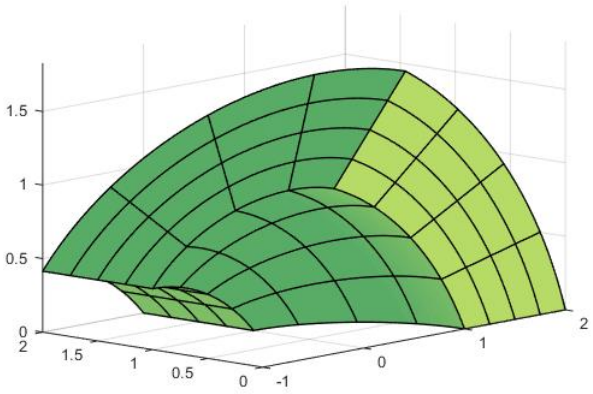


Figure 2: Revolved ring domain

for computational ease, while in the other two cases b is the vector representing the function $f(x, y, z) = 2(x^2 - x) + 2(y^2 - y) + 2(z^2 - z)$. In all problems we set K as the identity matrix, since according to (5) the presence of coefficients and of a nontrivial geometry have an analogous impact on the difficulty of the problem.

We start by considering the problem on the cube domain, which is simply $[0, 1]^3$. As already said, in this case $\mathcal{P} = \mathcal{A}$ and we can directly apply the considered method to the system $\mathcal{A}u = b$. This is not a realistic case but serves as a preliminary check on the proposed theory and implementation. Results for different values of h and p are shown in Table 1. We can see that the computational time required by the direct solver is independent of the degree p. In fact, the timings look impressive and show the great efficiency of this approach. We emphasize that, on the finer discretization level, problems of more than one billion variables is solved in slightly more than one minute (regardless of p).

We now consider the problems with nontrivial geometries, where the two methods are used

	3D Direct Solver Time (seconds)					
h^{-1}	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
128	0.07	0.06	0.06	0.07	0.08	0.08
256	0.53	0.61	0.64	0.65	0.65	0.56
512	5.58	8.01	6.80	5.83	8.00	5.76
1024	66.79	68.46	67.30	66.38	66.60	63.86

Table 1: Performance of the Bartels-Stewart method, cube domain.

	CG + $\mathcal{P}$ Iterations / Time (seconds)				
h^{-1}	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
32	26 / 0.21	26 / 0.42	26 / 0.83	26 / 1.63	26 / 2.84
64	27 / 1.46	27 / 3.36	27 / 7.41	27 / 13.44	27 / 22.95
128	28 / 12.74	28 / 30.30	28 / 60.14	*	*

	CG + IC Iterations / Time (seconds)				
h^{-1}	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
32	21 / 0.36	15 / 1.22	12 / 3.60	10 / 10.22	9 / 27.30
64	37 / 4.65	28 / 13.63	22 / 36.48	18 / 94.42	16 / 242.22
128	73 / 65.91	51 / 168.31	41 / 386.84	*	*

Table 2: Thick quarter of ring domain. Performance of CG preconditioned by the Bartels-Stewart method (upper table) and by Incomplete Cholesky (lower table).

as preconditioners for CG. In both cases, we set 10^{-8} as tolerance on the relative residual for CG. To better judge the efficiency of the Bartels-Stewart approach, we compare the results with those obtained when using a preconditioner based on the Incomplete Cholesky (IC) factorization (implemented by the MATLAB function `ichol`). We remark that incomplete factorizations have been considered as preconditioners for IGA problems in [6], where the authors observed that this approach is quite robust w.r.t. p.

In Table 2 we report the results for the thick quarter ring while in Table 3 we report the results for the revolved ring. The symbol "*" denotes the cases in which even assembling the system matrix $\mathcal{A}$ was unfeasible due to memory limitations. We emphasize that the reported computation time always includes the time needed to setup the preconditioner. Below we report some comments on the numerical results for these two problems.

	CG + $\mathcal{P}$ Iterations / Time (seconds)				
h^{-1}	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
32	40 / 0.30	41 / 0.62	41 / 1.25	42 / 2.61	42 / 4.47
64	44 / 2.16	44 / 5.21	45 / 11.15	45 / 21.83	45 / 36.60
128	47 / 21.99	47 / 56.58	47 / 113.42	*	*

	CG + IC Iterations / Time (seconds)				
h^{-1}	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
32	24 / 0.40	18 / 1.31	15 / 3.76	12 / 10.36	11 / 27.01
64	47 / 5.53	35 / 15.70	28 / 40.60	24 / 99.76	20 / 249.54
128	94 / 108.53	71 / 241.43	57 / 889.05	*	*

Table 3: Revolved quarter of ring domain. Performance of CG preconditioned by the Bartels-Stewart method (upper table) and by Incomplete Cholesky (lower table).

h^{-1}	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
32	24.43	12.40	6.52	4.15	2.66
64	19.60	9.10	3.78	2.21	1.24
128	15.58	7.00	3.73	*	*

Table 4: Percentage of time spent in the application of the 3D Bartels-Stewart preconditioner with respect to the overall CG time. Revolved ring domain.

- For the Bartels-Stewart preconditioner, the number of CG iterations is practically independent on p and slightly increases as the mesh is refined, but stays uniformly bounded according to (5).
- Interestingly, in the IC approach the number of CG iterations decreases for higher p. On the other hand, the CPU time still increases due to the greater computational cost of forming and applying the preconditioner. In particular, the performance of the IC approach shows a stronger dependence on p than the Bartels-Stewart approach.
- The Bartels-Stewart preconditioner yields better performance, in terms of CPU time, than the IC preconditioner for all considered values of h and p.

It is fundamental to observe that the computation times reported in Tables 2 and 3 include the time spent in the residual computation of the outer CG iteration (a sparse matrix-vector product, costing $O(p^3n^3)$ FLOPs). Somewhat surprisingly, in the case of the Bartels-Stewart preconditioner this step actually represents a significant effort in the overall CG performance. In fact, our numerical experience shows that the 3D direct method is so efficient that its cost is often negligible w.r.t. to the cost of the residual computation. This effect is clearly shown in Table 4, where we we report the percentage of time spent in the application of the preconditioner when compared with the overall time of CG, in the case of the revolved ring domain.

6 Conclusions

In this work we have analysed and tested the use of a fast solver for Sylvester-like equation, namely the Bartels-Stewart method, as preconditioner for isogeometric discretizations. We considered here a Poisson problem on a single-patch domain, and we focused on the k-method, i.e., splines with maximal smoothness. The considered preconditioner $\mathcal{P}$ is robust w.r.t. h and p, and we have compared two popular methods for its application. We found that the considered approach is very effective and easily outperform a simple-minded incomplete Cholesky preconditioner.

Our conclusion is that the Bartels-Stewart method represents a very promising preconditioning approach in an iterative solver for isogeometric discretizations. In a forthcoming paper we will further study the role of the geometry parametrization on the performance of the approach based on Sylvester equation solvers, and propose possible strategies to improve it.

REFERENCES

[1] G. Sangalli, and M. Tani, Isogeometric preconditioners based on fast solvers for the Sylvester equation, arXiv preprint arXiv:1602.01636, 2016.

[2] J. A. Cottrell, T. J. R. Hughes, Y. Bazilevs, *Isogeometric analysis: toward integration of CAD and FEA*, John Wiley & Sons, 2009

[3] J. A. Cottrell, T. J. R. Hughes, A. Reali, Studies of refinement and continuity in isogeometric structural analysis, *Computer methods in applied mechanics and engineering*, **196**, 4160–4183, 2007.

[4] L. Beirão da Veiga, A. Buffa, G. Sangalli, R. Vázquez, Mathematical analysis of variational isogeometric methods, *Acta Numerica*, **23**, 157–287, 2014.

[5] N. Collier, D. Pardo, L. Dalcin, M. Paszynski, V. M. Calo, The cost of continuity: a study of the performance of isogeometric finite elements using direct solvers, *Computer Methods in Applied Mechanics and Engineering*, **213**, 353–361, 2012.

[6] N. Collier, L. Dalcin, D. Pardo, V. M. Calo, The cost of continuity: performance of iterative solvers on isogeometric finite elements, *SIAM Journal on Scientific Computing*, **35**, A767–A784, 2013.

[7] A. Buffa, and H. Harbrecht, and A. Kunoth, and G. Sangalli, *BPX-preconditioning for isogeometric analysis*, Computer Methods in Applied Mechanics and Engineering, **265**, 63–70, 2013.

[8] K. P. S. Gahalaut, J. K. Kraus, S. K. Tomar, Multigrid methods for isogeometric discretization, *Computer methods in applied mechanics and engineering*, **253**, 413–425, 2013.

[9] L. Beirão da Veiga, D. Cho, L. F. Pavarino, S. Scacchi, Overlapping Schwarz methods for isogeometric analysis, *SIAM Journal on Numerical Analysis*, **50**, 1394–1416, 2012.

[10] L. Beirão da Veiga, D. Cho, L. F. Pavarino, S. Scacchi, BDDC preconditioners for isogeometric analysis, *Mathematical Models and Methods in Applied Sciences*, **23**, 1099–1142, 2013.

[11] M. Donatelli, and C. Garoni, and C. Manni, and S. Serra-Capizzano, and H. Speleers, Robust and optimal multi-iterative techniques for IgA Galerkin linear systems, *Computer Methods in Applied Mechanics and Engineering*, **284**, 230–264, 2015.

[12] C. Hofreither, and S. Takacs, and W. Zulehner, A Robust Multigrid Method for Isogeometric Analysis using Boundary Correction, *NFN Technical Report 33*, 2015.

[13] L. Gao, Kronecker Products on Preconditioning, PhD Thesis, King Abdullah University of Science and Technology, 2013.

[14] V. Simoncini, Computational methods for linear matrix equations, to appear on *SIAM Review*, 2013.

[15] R. H. Bartels, G. W. Stewart, Solution of the matrix equation AX+ XB= C, *Communications of the ACM*, **15**, 820–826, 1972.

[16] R. E. Lynch, J. R. Rice, T. H. Donald, Direct solution of partial difference equations by tensor product methods, *Numerische Mathematik*, **6**, 185–199, 1964.

[17] G. H. Golub, C. F. Van Loan, *Matrix computations*, JHU Press, 2012.

[18] C. F. Van Loan, The ubiquitous Kronecker product, *Journal of computational and applied mathematics*, **123**, 85–100, 2000.

[19] C. De Falco, A. Reali, and R. Vázquez, GeoPDEs: a research tool for isogeometric analysis of PDEs, *Advances in Engineering Software*, **42**, 1020–1034, 2011.

[20] L. Sorber, M. Van Barel, L. De Lathauwer, Tensorlab v2. 0, Available online, URL: www.tensorlab.net, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EFFECTIVE 3D REGULARIZED XFEM FOR PULL-OUT OF STEEL BARS IN CONCRETE, BENDING AND SHEAR TESTS ON FRP-REINFORCED CONCRETE BEAMS

Elena Benvenuti[1*], **Nicola Orlando**[1]

[1] Università di Ferrara, Dipartimento di Ingegneria
address Via Saragat 1, I-44122 Ferrara, Italy
e-mail: elena.benvenuti@unife.it, nicola.orlando@unife.it
ph. nr. +39 0532 974935

Keywords: Delamination, debonding, 3-Dimensional Finite element analysis, XFEM

Abstract. *We present an effective three-dimensional tool for modeling the influence on the structural behavior of the mechanical behavior of the interface between steel bars and FRP plates glued to concrete beams. The proposed tool reproduces the pull out of the bars and the detachment of the FRP plates. Both opening and sliding of the interface between reinforcements and beam are accounted for. The tool is based on the regularized extended finite element method (XFEM) developed by the authors: the discontinuities are regularized and a mechanically consistent variational formulation is adopted. The considered cases range from the simulation of pull-out tests of steel rebars, to the simulation of single-lap tests in shear and three point bending tests in FRP-reinforced concrete beams.*

1 INTRODUCTION

This contribution deals with the modeling of the influence of steel bars and FRP external reinforcements on the structural behavior of concrete beams. We present an original three-dimensional effective tool for modeling the detachment of steel and FRP plates from concrete beams. The proposed extended finite element model is capable of reproducing the pull out of steel bars and the detachment of the FRP-concrete layer taking into account both opening and sliding at the interface between the reinforcements and the concrete. The approach adopted in the present analysis is original, and is based on a recently developed 3D procedure [1] extending the previous 2D formulation for cohesive-like interfaces [2, 3]. The approach can be regarded as an original variant of the widely known eXtended Finite Element Method (XFEM) [4]. In the present XFEM, the discontinuities are regularized, so to introduce an additional internal length in the description. Moreover, a suitable variational formulation is adopted, based on the concept of the equivalent strain proposed in [5]. A wide range of loading cases and types of reinforcement is considered to prove the generality of our approach. In particular, we show some of the results obtained for the three-dimensional simulation of:

- the pull-out of steel bars embedded in concrete blocks;
- single-lap shear tests where the detachment of FRP plates glued to the concrete blocks occurs;
- three point bending tests on concrete beams externally-reinforced by FRP-plates.

2 BASIC RELATIONSHIPS OF THE REGULARIZED XFEM

The proposed approach has several advantages that make it suitable for modeling mechanical problems such as pull-out and delamination. Firstly, the model is enriched with a regularization length that governs the width of the process zone. This feature avoids the possible dependency on the mesh-size adopted within the zone where strain localization occurs. For instance, this drawback occurs when softening constitutive laws are adopted for the concrete. Secondly, a smooth continuous-discontinuous transition is obtained [6]. Thirdly, the regularized XFEM is based on a damage law that can correctly reproduce both the opening and the shearing modes characterizing the debonding of the FRP layer from the concrete. A thorough numerical investigation on single-lap shear tests has been presented in a recent paper [1].

The basic relationships of the regularized XFEM approach are shown hereinafter. Let the displacement field $\mathbf{u}$ be discontinuous across the failure surface defined according the mechanical problem $\mathcal{S} \in \mathbb{R}^3$ of normal $\mathbf{n}_S$. Within a single element of nodal degrees of freedom $\mathbf{U}^e$ and $\mathbf{A}^e$ interpolated by the usual FE interpolation functions $\mathbf{N}^e$, the regularized XFEM displacement $\mathbf{u}^e$ of element e is [4]

$$\mathbf{u}^e = \mathbf{N}^e\mathbf{U}^e + \mathcal{H}_\rho\mathbf{N}^e\mathbf{A}^e\,, \tag{1}$$

where $\mathcal{H}_\rho$ is a regularized Heaviside function that approximates the Heaviside function for vanishing regularization length ρ. The vector $\mathbf{A}^e$ collects the jump components along x, y and z for the finite element e. $\mathcal{H}_\rho$ is assumed a function of the distance from the level set surface. By compatibility, the strain field is

$$\boldsymbol{\varepsilon}^e = \mathbf{B}^e\mathbf{U} + \mathcal{H}_\rho\mathbf{B}^e\mathbf{A}^e + \|\nabla\mathcal{H}_\rho\|(\mathbf{N}^e\mathbf{A}^e \otimes \mathbf{n}^e)\,, \tag{2}$$

where $\mathbf{B}^e = \nabla\mathbf{N}^e$ is the standard FE compatibility matrix.

FRP and adhesive are modeled as linear elastic materials. At each Gauss point, we introduce the damage variables D and D_c for the concrete and the zone where debonding occurs, respectively, and compute the associated stresses thorough

$$\boldsymbol{\sigma}^e = (1-D)\mathbf{E}\,\mathbf{B}^e\mathbf{U} + \mathcal{H}_\rho\,(1-D)\mathbf{E}\,\mathbf{B}^e\mathbf{A}^e\,, \tag{3a}$$

$$\boldsymbol{\sigma}_c^e = (1-D_c)\mathbf{E}_c\|\nabla\mathcal{H}_\rho\|(\mathbf{N}^e\mathbf{A}^e\otimes\mathbf{n}^e)\,. \tag{3b}$$

In particular, the concrete damage D is governed by an exponential Rankine elasto-damaging

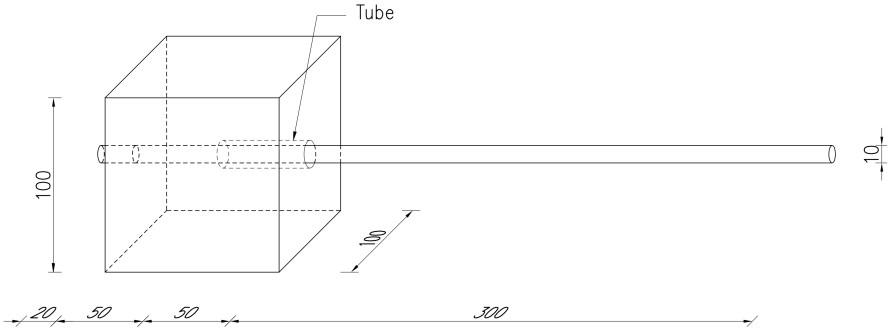


Figure 1: Pull-out: Geometry of the pullout test [7]

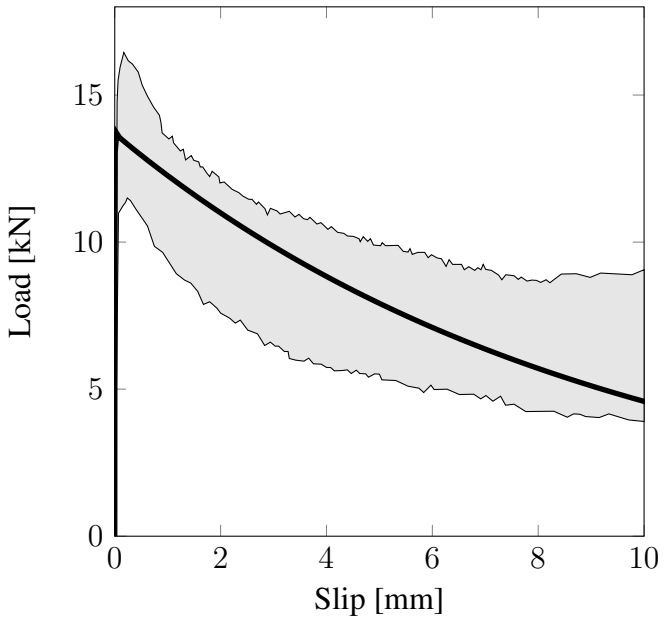


Figure 2: Pull-out: Numerical and experimental load vs slip curves with plain rebars

law until it has reached the critical value D^{cr} [6]. As soon as $D \geq D^{cr}$, the evolution of D is dropped, i.e. the concrete can only elastically unload: a regularization zone replaces the discontinuity. In particular, the damage evolution can be D_c affecting the stress $\boldsymbol{\sigma}_c^e$ is

$$D = \min\{D^{cr}, f(r)\}\,, \quad D_c = \max\{D^{cr}, f(r)\}\,, \quad f(r) = 1 - \frac{r_0}{r_c}\exp\big(-2\,H\frac{r_c - r_0}{r_0}\big)\,, \tag{4}$$

where $r_c \geq r_0$, with $r_0 = f_t$ for tensile damage, and $r_0 = f_c$ for compressive damage.

The stiffness matrix associated with the variational formulation is

$$\mathbf{K}^e_{reg} = \begin{bmatrix} \mathbf{B}^{et}\,(1-D)\mathbf{E}\,B_u & \mathbf{B}^{et}\,\mathcal{H}_\rho(1-D)\mathbf{E}\,\mathbf{B}^e \\ \mathbf{B}^{et}\,\mathcal{H}_\rho\mathbf{E}\,\mathbf{B}^e & \mathbf{B}^{et}\,\mathcal{H}^2_\rho(1-D)\mathbf{E}\mathbf{B}^e + \|\nabla\mathcal{H}_\rho\|\bar{\mathbf{N}}^{e\,t}(1-D_c)\mathbf{E}_c\bar{\mathbf{N}}^e \end{bmatrix}. \quad (5)$$

In Eq. (5), the operator $\bar{\mathbf{N}}^e$ is such that $\nabla\mathcal{H}_\rho \otimes \mathbf{N}^e A \approx \|\nabla\mathcal{H}_\rho\|\bar{\mathbf{N}}^e\mathbf{A}_e$, and $\nabla(\mathcal{H}_\rho\mathbf{N}^e) = \mathcal{H}_\rho\,\mathbf{B}^e + \nabla\mathcal{H}_\rho \otimes \mathbf{N}^e$. The adopted variational formulation has been thoroughly described in [5].

3 PULL-OUT

In this section, the regularized XFEM approach is employed to investigate the behavior of the bond between concrete and steel rebars. For a comparison with experimental results, we have considered the experimental campaign of Xiao et al. [7], where several pullout tests with plain and deformed bars were carried out. Fig. 1 shows the pullout test specimens. According to the experimental test, the load was monotonically increased. The applied load versus the slip measured at the free end of the steel rebar have been detected. The numerical and the experimental results are compared in Fig. 2, where the gray zone indicates the experimental results. The agreement between the results is very satisfying.

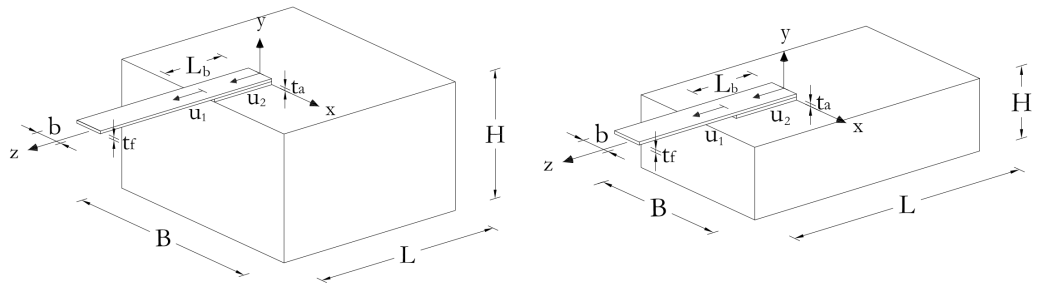
4 SINGLE-LAP SHEAR TESTS

Two single-lap-shear tests have been simulated, where an axial pulling force is applied to an FRP-plate that is glued to a concrete block restrained by a suitable supporting system. Consistently with the experiments, a layer comprehensive of both a concrete layer attached to the FRP plate is modeled.

For a comparison with the experimental results, the experimental test performed by Chajes et al. [8] and the tests carried out by Carrara et al. [9] have been simulated. It is remarkable that, in the former test, the structural response has been recorded up to the maximum load. On the contrary, in the latter test, the entire load-displacement curves with softening post-peak branches have been detected. The geometry and the material parameters of the experimental tests are specified in Tab. (4). The specimens are displayed in Fig. 3. In the geometry adopted by Chajes et al. [8], the glue layer starts at the front of the specimen, close to the loaded end of the FRP plate. On the contrary, in the specimen of Carrara et al. [9], the FRP plate is bonded at a certain distance from the front of the concrete block to avoid the detachment of a concrete wedge when pulling the FRP sheet. The maximum transferable loads computed are shown for variable bonding length in Fig. 4 for both the specimens. Fig. 5 shows a comparison between

	Chajes et al.[8]			Carrara et al. [9]		
	FRP	glue	concrete	FRP	glue	concrete
E [MPa]	108380	1585	33640	168500	3517.3	28700
ν	0.248	0.315	0.2	0.248	0.315	0.2
f_t[MPa]	–	–	3.21	–	–	2.85
thickness [mm]	1.0	1.0	–	1.3	1.3	–
H [mm]	152.4	152.4	152.4	90	90	90
B [mm]	150	150	150	152.4	152.4	152.4
L [mm]	300	300	300	228.6	228.6	228.6
b [mm]	25.4	25.4	25.4	30	30	30

Table 1: Shear tests: Material and geometry parameters [2]



(a) (b)

Figure 3: Shear tests: Geometries of the tests of Chajes et al. [8] and Carrara et al. [9]

the numerical displacement contour plot (b) and displacement vectors determined by Czaderski et al. [10] through Image Correlation System (ICS). As evident, the agreement of the computed results with the experimental measured is very satisfying. In particular, this comparison puts into evidence the crucial role played by the bending of the FRP plate during delamination. More results on these single-lap shear tests can be found in our study [1].

5 FLEXURAL STRENGTHENING: THREE-POINT BENDING TEST

In this section, the flexural strengthening carried out by FRP plates bonded to concrete beams has been investigated by means of the regularized XFEM approach. To this purpose, the three-point bending test carried out by Yin et al. [11] was modelled. The geometry of the specimen is reported in Fig. 6. Noteworthy, a steel fiber reinforced concrete was used with short steel-fibers 30 mm long with a diameter of 0.5 mm. An FRP-plate was glued to the bottom of the beam. No steel bars were employed. Concrete with compressive strength $f_{cm} = 26$ MPa, Young's modulus $E_c = 25$ GPa and Poisson ratio $\nu = 0.2$ were employed.

Different volumes of short steel-fibers were mixed with the concrete matrix. In particular, the experimental results obtained using four volume-fractions of short steel-fibers. As known, the mechanical properties of the steel-fiber reinforced concrete change with V_f. Among the available models predicting the improvement of the strength of the concrete with V_f, we adopted the Job et. al. [12] formula

$$f_{spc} = 0.63(f_{cm})^{0.5} + 0.288(f_{cm})^{0.5} R_I + 0.052 R_I \,, \tag{6}$$

where the strength is a function of the fiber reinforced index ($R_I = V_f L_f / \phi_f$), L_f is the length of fiber, and ϕ_f is the diameter of fiber.

In these experiments, the FRP-debonding is due to the concentration of the interfacial shear stress after the occurrence of diffused flexural cracks in the concrete. In Fig. 7, the experimental crack pattern [11] obtained by setting $f_t = 2.89$ MPa, and $H = 0.008$ for $V_f = 0\%$is compared with that obtained with the present regularized XFEM. In particular, the damage variable in the concrete has been plotted, with the red color denoting value of the damage variable close to 1. The evolution of the process of detachment of the FRP plate from the concrete beam is shown in Fig. 8.

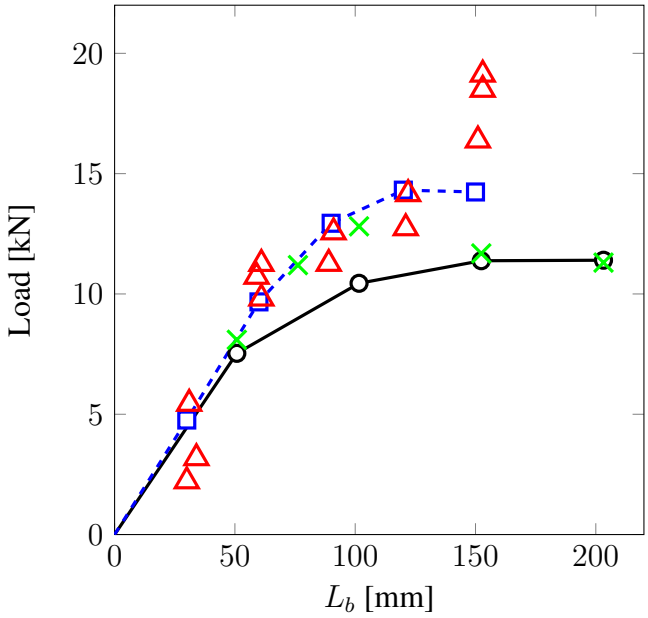
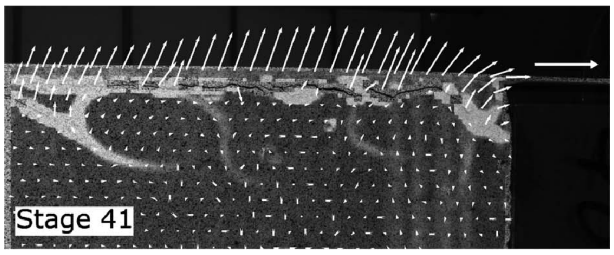
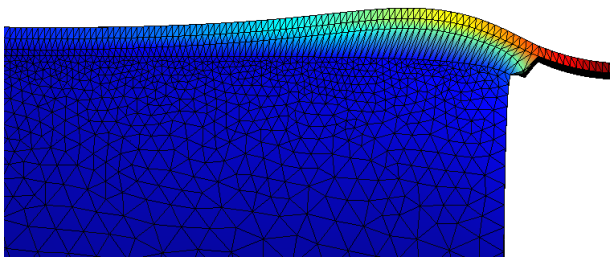


Figure 4: Shear tests: Numerical and experimental maximum lads versus bonding length L_b for the tests of Carrara et al. [9] (numerical data: dashed line and rectangle marker, experimental data: triangle marker) and Chajes et al. [8] (numerical data: continuous line and circle markers [2], experimental data: cross markers).



(a)



(b)

Figure 5: Shear tests: Qualitative comparison between the numerical displacement contour plot (b) and the displacement vectors (a) determined by Czaderski et al. [10] through a Image Correlation System (ICS).

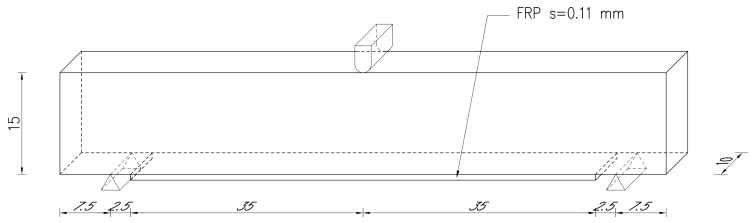
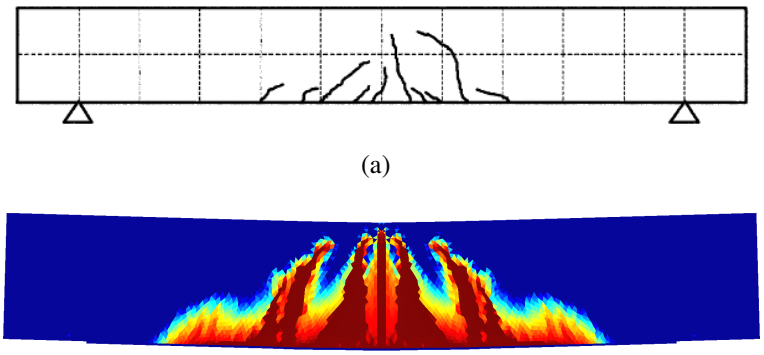


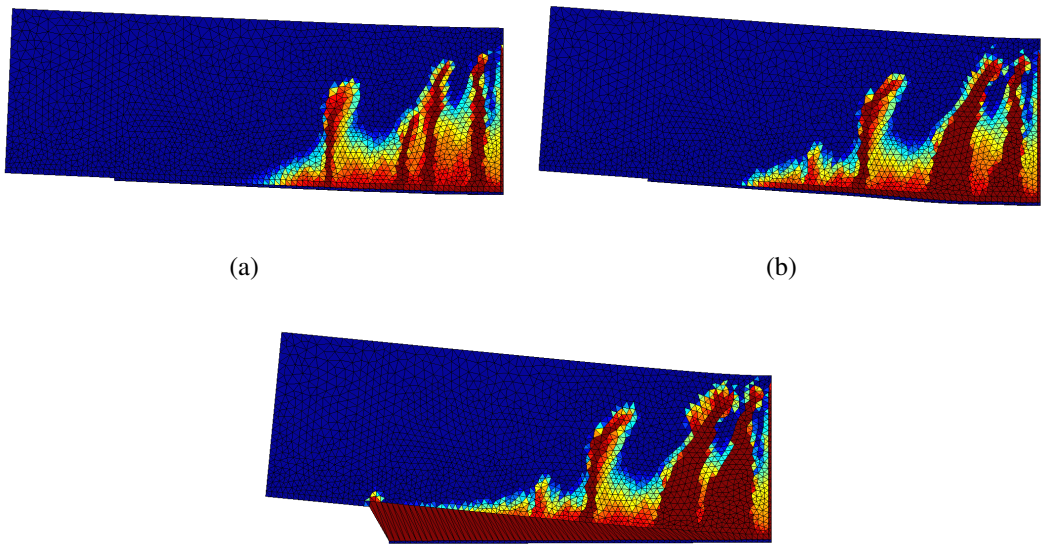
Figure 6: Three point bending test: Geometry of the test carried out by Yin et al.[11]



(a)

(b)

Figure 7: Three point bending test: Experimental (a) and numerical (b) crack pattern of the test carried out by Yin et al. [11]



(a) (b)

(c)

Figure 8: Three point bending test: damage evolution up to final delamination

6 CONCLUSIONS

Accounting for both the shearing and the opening of the interface between the reinforcements and the concrete, the regularized three-dimensional XFEM approach has been successfully applied to simulate the debonding of steel rebars and FRP-plates in pull out, shear and bending tests. The proposed approach is able to catch the main features of the structural behavior. The agreement with the reference results is highly satisfying. Hence, the proposed approach can be used together with and in alternative to experimental tests.

REFERENCES

[1] E. Benvenuti, N. Orlando, D. Ferretti, A. Tralli, *A new 3D experimentally consistent XFEM to simulate delamination in FRP-reinforced concrete*, **91**, 346–360, 2016.

[2] E. Benvenuti, O. Vitarelli, A. Tralli, Delamination of FRP-reinforced concrete by means of an extended finite element formulation. *Composites Part B: Engineering*, **43.8**, 3258–3269, 202.

[3] E. Benvenuti, G. Ventura, N. Ponara, A. Tralli, Variationally consistent eXtended FE model for 3D planar and curved imperfect interfaces. *Computer Methods in Applied Mechanics and Engineering*, **267**, 434–457, 2013.

[4] T. Belytschko, R. Gracie, G. Ventura, A review of extended/generalized finite element methods for material modeling, *Modelling and Simulation in Material Science and Engineering*, **17.4**, 043001, 2009.

[5] E. Benvenuti, XFEM with equivalent eigenstrain for matrix–inclusion interfaces. *Computational Mechanics*, **53.5**, 893–908, 2014.

[6] E. Benvenuti, A. Tralli, Simulation of finite-width process zone for concrete-like materials, *Computational Mechanics*, **50**, 479–497, 2012.

[7] J. Xiao, H. Falkne, Bond behaviour between recycled aggregate concrete and steel rebars. *Construction and Building Materials*, **21**, 395–401, 2007.

[8] MJ. Chajes, WW. Finch Jr, TF. Januszka, TA. Thomson Jr, Bond and force transfer of composite material plates bonded to concrete, *ACI Structural Journal*, **93**, 209–217, 1996.

[9] P. Carrara, D. Ferretti, F. Freddi, G. Rosati, Shear tests of carbon fiber plates bonded to concrete with control of snap-back. *Engineering Fracture Mechanics*, **15**, 2663–2678, 2011.

[10] C. Czaderski, K. Soudki, M. Motavalli, Front and Side View Image Correlation Measurements on FRP to Concrete Pull-Off Bond Tests *Journal of Composites for Costruction*, **47**, 1326–1336, 2011.

[11] J. Yin, Z.S. Wu, Structural performances of short steel-fiber reinforced concrete beams with externally bonded {FRP} sheets. *Construction and Building Materials*, **17**, 463–470, 2007.

[12] T. Job, R. Ananth, Mechanical properties of steel fiber-reinforced concrete. *Journal of materials in civil engineering*, **19**, 385–392, 2007.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EXPERIMENTAL VALIDATION OF COMPOSITE STRUCTURES IN EXPLICIT IMPACT ANALYSIS

Konstantinos T. Fotopoulos[1], **George N. Lampeas**[2]

Laboratory of Technology and Strength of Materials
Department of Mechanical Engineering and Aeronautics
University of Patras, 26500 Rion, Greece
[1] fotopk@mech.upatras.gr
[2] labeas@mech.upatras.gr

Keywords: Model Validation, Composite laminate, Low velocity impact, Finite element analysis, Delamination, Stacked shell modeling.

Abstract. *The prediction of composite laminate behavior when subjected to low velocity impact loading is a challenging task due to the complexity of the fully three dimensional stress state that develops inside the structure. Several methods have been employed in order to analyze the intralaminar and interlaminar failure modes that evolve during impact. In the present work a stacked shell methodology is proposed for the simulation of composite laminates. The approach is experimentally validated through the comparison of its results to the respective of a low velocity impact on laminated specimens test. The numerical predictions correlate well to the experimental results, showing the accuracy of the stacked-shell approach in explicit impact analysis of composite structures.*

1 INTRODUCTION

Lightweight structural design has become very important in the automotive, marine and aircraft engineering fields. The increasing demand for materials that combine low weight with high mechanical strength has generated high interest in the analysis of the behavior of composite materials. One of the most significant issues of this analysis is the effect of impact loading on composite laminated structures, especially because of their susceptibility to out-of-plane failure. Although a wide range of scenarios exist, impact on composites can generally be categorized in low and high velocity impact. The former usually includes cases of projectiles with significant mass and volume that are dropped on the structure, while the latter mostly comprises instances of small projectiles impacting composites at high speeds. The effects of such impact events on a laminate include fiber failure, matrix cracks as well as separation of plies (delamination); hence, the outcome consists of a combination of failure modes making the analysis of the impact response a demanding task.

In order to accurately simulate the different phenomena arising during an impact event, several predictive methods has been utilized. The main approaches for the numerical simulation of impact induced effects are based on Continuum Damage Mechanics (CDM) and Fracture Mechanics. In CDM material models [1-3] both the plies and the interfaces of a composite are modeled using appropriate damage theories. On the other hand, fracture mechanics techniques – e.g. the Virtual Crack Closure Technique (VCCT) or the Cohesive Zone Method (CZM) – treat in-plane and out-of-plane behavior separately, concentrating individually on the progression of interlaminar damage [4]. The VCCT [5] has been applied extensively in the analysis of metals and has been extended to composite materials using the energy release rate concepts to directly predict the crack growth. However, the location and size of the crack are prerequisites for the application of the method, making it difficult to be employed in more complex structures. The CZM – based on the works of Dugdale [6] and Barenblatt [7] – assumes that a cohesive zone is developed in front of the crack tip, with the cohesive surfaces associated with the crack being connected by cohesive laws. This technique is capable of predicting separation growth without the assumptions of the VCCT, although it presents a certain degree of mesh dependency.

Considering laminated composites as a conjunction of fiber reinforced plies and matrix interfaces, the analysis of a composite's behavior can be approached in the Finite Element (FE) environment using CDM techniques to address intralaminar failure and Fracture Mechanics techniques to simulate interlaminar failure. In this frame, the representation of out-of-plane normal and shear deformation through-thickness can be achieved by stacking an appropriate number of 2D damage mechanics elements through the laminate thickness, held together by cohesive traction. This approach is mentioned in the literature as 'stacked shell' or 2.5D approach. It is based on the CDM approach proposed by Allix and Ladeveze [1], replacing the CDM matrix representation with cohesive zone interfaces. The main advantage of this modeling technique is the revocation of aspect ratio restrictions concerning the in-plane and through thickness element dimensions, presenting obvious advantages in regard to simulation accuracy and computational efficiency, compared to traditional modeling techniques. In the studies [8-14] of the literature the use of the stacked shell method has been reported for the simulation of composite laminates.

The aim of the present work is the study of the appropriateness of the stacked shell approach capabilities in explicit analysis of low velocity impact events, as well as in the prediction of interlaminar damage initiation and growth. The accuracy of the stacked shell modeling technique is assessed by an experimental procedure, concerning low velocity impact on com-

posite laminates. Numerical predictions are compared to published experimental data, derived from Ultrasound A-scan, Ultrasound C-scan and Thermography non-destructive tests.

2 DESCRIPTION OF THE MODELLING APPROACH

The stress distribution produced by impact on a laminated composite is three-dimensional. Complex phenomena that include propagation of compressive, shear and surface stress waves occur during the impact event [15]. The simulation of the stress state that develops under these conditions is a challenging task, which requires a modeling strategy which is capable to predict the three-dimensional stress state of the laminate.

The idea behind the stacked shell FE model development is based on the hypothesis that a composite laminate can be simulated by a number of discrete sublaminates which are kinematically tied by cohesive interfaces simulating the matrix. In this manner, the sublaminates are mainly burdened with the axial and bending loads which are developed in the laminate, while the cohesive interfaces take into account the out-of-plane normal and shear loading effects. In this way the displacement and strain continuity between adjacent sublaminates is provided, since forces are transferred from one sublaminate to another through the cohesive interfaces. Accordingly, the FE modeling strategy can be divided in:

a) In-plane damage modeling

b) Interface modeling

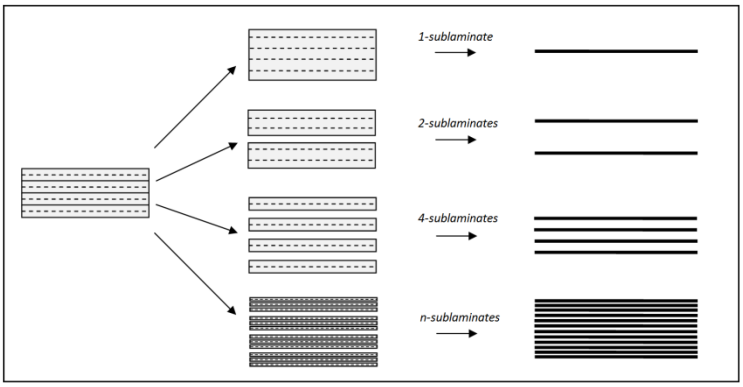


Figure 1: Representation of the stacked shell approach (taken from [14])

2.1 In-plane damage modeling

The low velocity impact induced in-plane damage mainly consists of fiber failure and matrix cracking. The simulation of these failure modes is performed through the use of a CDM elastic-plastic material model which is based on the concept of failure surfaces. The longitudinal, transverse and shear strengths are taken for the yield function, resulting to a faceted surface. The failure surfaces for this material model are:

$$\begin{aligned} \sigma_{1t} &= a\mathrm{X}_t \\ \sigma_{1c} &= -\mathrm{X}_c \\ \sigma_{2t} &= aY_t \\ \sigma_{2c} &= -Y_c \\ \sigma_{12} = \sigma_{23} &= \sigma_{13} = \pm S_c \end{aligned} \tag{1}$$

where σ_{1t}, σ_{1c}, σ_{2t}, σ_{2c}, σ_{12}, σ_{23}, σ_{13} are the stresses expressed in the material principal axes (1 for the axial fiber direction, 2 for the transverse matrix direction and 3 for the through-thickness direction) X_t, X_c are the longitudinal tensile and compressive strengths, respectively, Y_t, Y_c are the transverse tensile and compressive strengths, respectively and S_c is the in-plane shear strength. The parameter a can be used to limit the tensile failure, when the respective strengths have been reached. For a = 1 the material model results in a fully elastic-plastic model that retains the initially defined strength values for the failure surface. For a = 0, the tensile stresses σ_{1t} and σ_{2t} are reduced to zero after strengths X_t, Y_t have been reached; subsequently, only compressive and shear loads can be carried.

According to a work performed by Maimi et al [16], in order to ensure the correct calculation of dissipated energy independently of the element size, the elastic energy of an element during damage initiation must be lower than or equal to the fracture energy. Therefore, the material strengths used in the material model for each damage law have to be modified, using Equation 2.

$$X_M = \sqrt{\frac{2E_M G_M}{l^*}} \tag{2}$$

where X, E are the strengths and moduli of the material respectively, in the direction denoted by the subscript M (longitudinal, transverse or shear), l^* is the element characteristic length and G_M is the fracture toughness. The characteristic length of square elements for unknown direction of crack propagation can be approximated as proposed by Bazant and Oh [17] by

$$l^* = \frac{\sqrt{A_{IP}}}{\cos(\gamma)} \tag{3}$$

where γ is the angle that the mesh lines and the crack direction form and A_{IP} is the area corresponding to an integration point. If the direction of crack propagation is unknown, the average of Equation 3 can be used:

$$l^* = 1.12\sqrt{A_{IP}} \tag{4}$$

Consequently, the effect of element size to the material model for in-plane damage is minimized.

2.2 Interface modeling – cohesive zone method

The damage that the interfaces of a laminate undergo during impact events constitutes of de-bonding between adjacent plies, i.e. delamination. Modeling of interface failure is realized through the implementation of a fracture mechanics based approach, which is a combination of a cohesive zone model and a linear elastic adhesive penalty formulation as proposed by Borg [18].

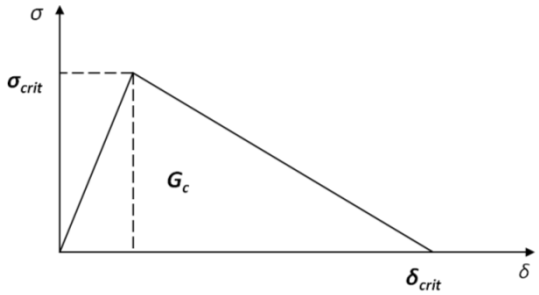


Figure 2: Cohesive law - Traction-separation curve

Regions between sublaminates are introduced through which damage initiation and growth is possible. In particular, the constitutive behavior of the cohesive surfaces is described by a bilinear law, as shown in Figure 2. The interface damage initiates when the failure criterion of Equation 5 is met.

$$\left(\frac{|\sigma_n|}{\sigma_{n,crit}}\right)^2+\left(\frac{|\sigma_s|}{\sigma_{s,crit}}\right)^2\geq 1 \tag{5}$$

In the above equation σ_n, σ_s are the normal and shear stresses respectively and $\sigma_{n,crit}$, $\sigma_{s,crit}$ are the failure stresses that correspond to each component. After the failure initiation point, damage between cohesive surfaces is a linear function of the distance between them. When the distance δ_{crit} is reached, the cohesive surfaces are released and complete ply separation occurs.

The length of the cohesive zone created at the tip of a crack is the distance from the crack tip to the point of maximum cohesive traction. According to Turon et al [19] the constitutive parameters for the simulation of delamination can be determined in order to adapt the length of the cohesive zone to the element size so that the energy dissipation is correct. This is performed using Equation 6.

$$\sigma_{crit}=\sqrt{\frac{9\pi EG_c}{32N_e l_e}} \tag{6}$$

where E is the transverse modulus of elasticity, G_c is the critical energy release rate, N_e is the number of elements in the cohesive zone and l_e is the element length in the crack propagation direction. The interface failure stress is then selected as the minimum between the value calculated by Equation 6 and the nominal failure stress. The regularization of the stress by Equation 6 ensures that the interface model will remain practically unaffected by mesh size produced artificial phenomena.

3 EXPERIMENTAL VALIDATION OF THE IMPACT SIMULATIONS

The capabilities of the stacked shell approach with respect to the prediction of the behavior of composite laminates are assessed through comparison of simulation results to the experimental results of a low velocity impact drop test.

3.1 Description of the experimental procedure

The experimental procedure has been performed by Riccio et al [20] according to the European association of Aerospace Manufacturers standard prEN 6038. The experiment involves the impact of a hemispherical impactor on a laminated plate with rectangular geometry, fabricated by the G1157/RTM-6 material system. The plate consists of 8 plies of 0.3 mm thickness each, resulting to a $[0/90/45/-45]_s$ layup. The dimensions of the flat specimen are 150 mm × 100 mm × 2.4 mm (Length × Width × thickness). The impactor's radius is 8 mm, with a mass of 3.64 kg. The specimen is supported by a fixture frame with a 125 mm × 75 mm rectangular window and is constrained by 4 rigid clamps with 10 mm diameter and 3 mm height. Impact on the plate is performed for three energy levels, namely 6 J, 10 J and 13J. Results are acquired by Ultrasound A-scan, Ultrasound C-scan and Thermography non-destructive tests.

3.2 FE model development

The FE model was developed in the explicit code of LS-Dyna, according to the simulation approach described in Section 2. It comprises eight sublaminates, each one representing a layer of the laminated plate, as shown in Figure 3. The sublaminates are modeled by fully integrated Reissner-Mindlin shell elements with six degrees of freedom per node, i.e. three translations (u*x*, u*y*, uz) and three rotations (rot*x*, rot*y*, rotz). This element type presents a Bathe-Dvorkin [21] transverse shear treatment in order to eliminate w-hourglass modes, while other hourglass modes are eliminated by selectively reduced integration. In the LS-Dyna environment this behavior is included in shell element formulation 16 [22].

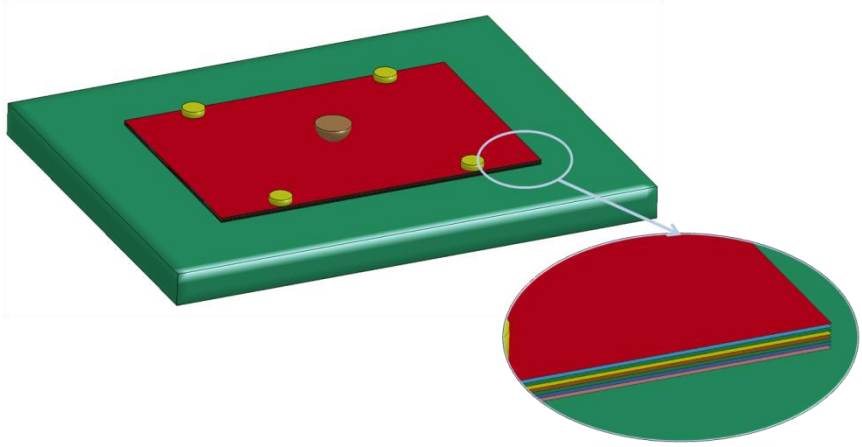


Figure 3: Finite element model of the low velocity impact test

Following a mesh convergence parametric investigation, each sublaminate is discretized by a finite element mesh consisting of 2 mm x 2 mm square elements and four integration points through-thickness. Equation 2 is used to modify material strengths in order to correspond to the element characteristic length, calculated by Equation 4. The G1157/RTM-6 lamina material properties as well as the intralaminar damage parameters (Table 1) that are used for the simulation of sublaminates are introduced in the 'Mat 59 Composite Failure' type of the LS-Dyna code, which is formulated according to the attributes of the elasto-plastic material model described in section 2.1.

G1157/RTM-6 Material Properties	ρ [kg/m^3]	E_a [GPa]	$E_b=E_c$ [GPa]	$G_{ab}=G_{ca}=G_{bc}$ [GPa]	$\nu_{ab}=\nu_{ac}=\nu_{bc}$
	1400	130.05	11.55	6.0	0.312
G1157/RTM-6 In-plane fracture energies	GF_{ftc} [kJ/m^2]	GF_{fcc} [kJ/m^2]	GF_{mtc} [kJ/m^2]	GF_{mcc} [kJ/m^2]	
	16.4	5.9	0.5	4.62	
G1157/RTM-6 Ply Damage Parameters	**Longitudinal Tensile Strength** [MPa]	**Longitudinal Compressive Strength** [MPa]	**Transverse Tensile Strength** [MPa]	**Transverse Compressive Strength** [MPa]	**In-plane Shear Strength** [MPa]
	1380	827.7	71.8	218.3	90

Table 1: G1157/RTM-6 Material properties and damage parameters

Regarding the FE implementation of interfaces, seven inter-layers were introduced between sublaminates. A tiebreak contact formulation is selected for the simulation of interlaminar behaviour. When the defined critical distance between initially tied nodes is reached, the tiebreak contact is released and converted to a typical surface-to-surface contact. According to Dobyns [23], delamination growth occurs because of transverse shear stresses developed at the area of impact. For this reason, Equation 6 is used to calculate the shear failure stress, with element length l_e equal to 2 mm and the desired number of elements in the cohesive zone N_e is set equal to 5, which was shown to produce accurate results [19]. Subsequently, the critical release distance δ_{crit} can be calculated using known values of the shear energy release rate which is approximated by Equation 7. Finally, the calculated value of the critical distance is substituted in Equation 8 that approximates the normal energy release rate in order to determine the normal failure stress. The interlaminar damage parameters, shown in Table 2, are defined in Option 8 of the 'Contact automatic one way surface to surface tiebreak' type, which is based on the CZM model presented in section 2.2.

$$G_{IIC} = \frac{1}{2}\sigma_{s,crit}\delta_{crit} \tag{7}$$

$$G_{IC} = \frac{1}{2}\sigma_{n,crit}\delta_{crit} \tag{8}$$

The impactor, fixture base and clamps are discretized with a relatively coarse mesh using three-dimensional solid elements, and their geometry is assumed to remain undeformed during the impact event. For this reason they are modeled using 'Mat 020 Rigid' of the LS-Dyna code. All degrees of freedom of the fixture base and clamps are constrained, while for the rigid impactor, only movement in the direction normal to the target surface is allowed. A typical penalty based surface to surface contact algorithm is used between the impactor, laminated plate and fixture base. To simulate the interaction between rigid clamps and the specimen, a penalty-based tied contact is used ('Contact tied surface to surface offset'), to secure the position of the composite plate. The three different energy levels used in the experimental procedure were achieved by altering the initial velocity of the rigid impactor.

G1157/RTM-6 Interface Energy Release Rates	**Normal Energy Release Rate** G_{Ic} [kJ/m^2]		**Shear Energy Release Rate** G_{IIc} [kJ/m^2]
	0.18		0.5
G1157/RTM-6 Interface Damage Parameters	**Normal Failure Stress** $\sigma_{n,crit}$ [MPa]	**Shear Failure Stress** $\sigma_{s,crit}$ [MPa]	**Critical Distance** δ_{crit} [mm]
	8.13	22.59	0.04427

Table 2: Interface properties and damage parameters

3.3 Comparison of numerical and experimental results

The numerical results of the stacked shell finite element model for the case of the laminated plate subjected to low velocity impact are presented in comparison with the experimental results of [20]. The force versus time and force versus displacement curves calculated by the

stacked shell simulation for the three different energy levels are presented in Figure 4. It can be observed that the maximum load values predicted by the FE model are in good correlation to the experimental results, as the deviation is within 1% for the 6 J energy level, within 1.5% for the 10 J case, and only for the 13 J energy level deviation is increased to about 15%.

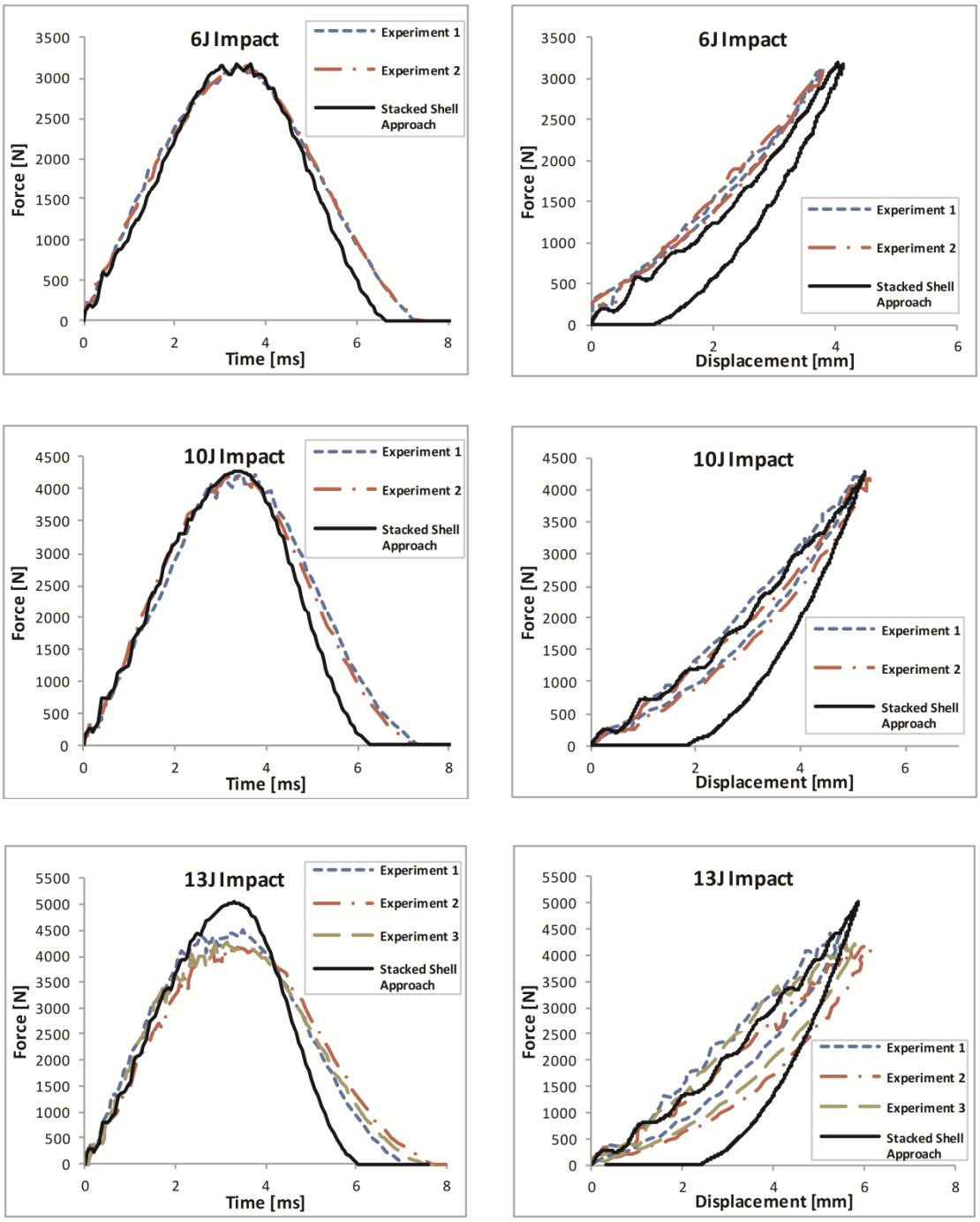


Figure 4: Comparison of simulation force-time and force-displacement curves with experimental results for 6 J, 10 J and 13 J energy levels

As far as the force versus displacement diagrams are concerned, excellent correlation to the experimental data can be observed for the loading phase, for all cases. The maximum displacement prediction is in excellent agreement with the experimental results. The deviation is ~8% for the 6J case, ~2% for the 10 J case and ~4% for the 13 J case. During the unloading

phase, the load capability tends to be underpredicted, a fact that can be attributed to a slight overprediction of the damage induced in plies.

Energy Level	Dimension [mm]	Thermography	Ultrasonic A-scan	Ultrasonic C-scan	Stacked Shell Approach
6 J	**Length**	14	20	20	**25**
	Width	7	15	14	**18**
10 J	**Length**	29	30	33	**32**
	Width	11	10	17	**19**
13 J	**Length**	41	40	42	**35**
	Width	18	22	20	**21**

Table 3: FE results for the delamination envelope dimensions compared to the experimental results of [20]

Stacked shell model predictions for the dimensions of the delamination envelope under the impacted area compared to results derived from NDT techniques are shown in Table 3. FE results correlate well with the experimental results, although for the lowest energy level, the delamination is slightly overestimated. Delamination plots of the seven contact interfaces are demonstrated in Figure 5, where the red color indicates delamination. As has also been reported by Abrate [15], it can be observed that at interfaces where adjacent plies have the same fiber orientation, almost no separation of plies occurs while, at points where damage is more intense delamination areas present an elongated 'peanut' shape, with the large dimension of the area oriented in the direction of the fibers of the lower ply that participates to the interface.

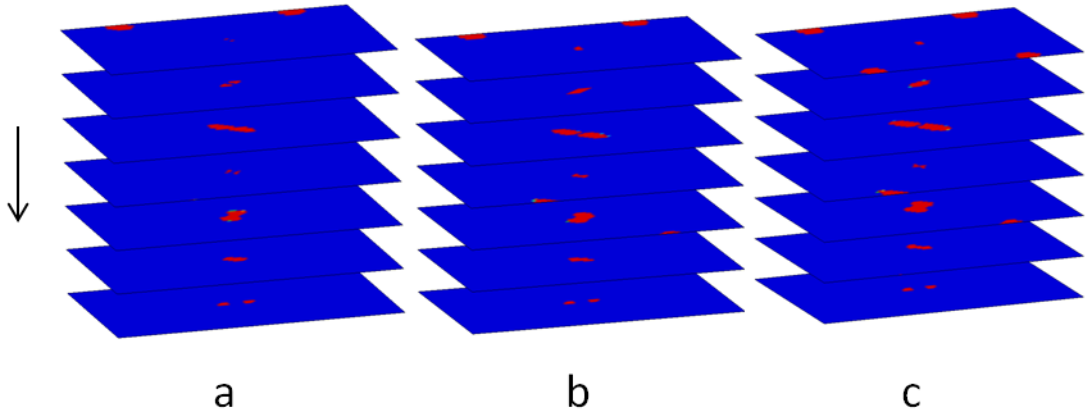


Figure 5: Delamination failure at interfaces. (a) 6 J, (b) 10 J, (c) 13 J

4 CONCLUSIONS

The stacked shell modeling methodology has been investigated in terms of explicit dynamics simulation of low velocity impact on fiber reinforced structures. The capabilities of the approach have been examined, with emphasis on the prediction of interlaminar damage. The investigation of the method's accuracy involves comparison with published experimental results concerning low speed impact on G1157/RTM-6 laminated plates. The stacked shell simulation results provide good correlation to the experimental results, as far as force and displacement predictions are concerned. The method's predictions of interlaminar failure (de-

lamination) present good agreement to the NDT derived experimental results. The good delamination prediction capabilities complimented by the method computational efficiency indicate that the stacked shell approach is a highly capable tool for the prediction of composite laminate behavior under low velocity impact conditions.

REFERENCES

[1] O. Allix and P. Ladevèze, Interlaminar interface modelling for the prediction of delamination. *Composite Structures*, **22**, 235-242, 1992.

[2] L. Daudeville and P. Ladevèze, A damage mechanics tool for laminate delamination. *Composite Structures*, **25**, 547-555, 1993.

[3] S. P. Rajbhandari, *Parametric finite element modeling of monolithic composite structures for impact damage*. Master of Engineering thesis, Royal Melbourne Institute of Technology, Australia, 2003.

[4] S. Sridharan, *Delamination behavior of composites*. Woodhead Publishing Limited, 2008.

[5] E. Rybicki and M. Kanninen, A finite element calculation of stress intensity factors by a modified crack closure integral. *Engineering Fracture Mechanics*, **09**, 931-938, 1977.

[6] D. Dugdale, Yielding of steel sheets containing slits. *Journal of the Mechanics and Physics of Solids*, **08**, 100-104, 1960.

[7] G. I. Barenblatt, The mathematical theory of equilibrium in brittle fracture. *Advances in Applied Mechanics*, **07**, 55–129, 1962.

[8] L. Greve and A. Pickett, Delamination testing and modelling for composite crash simulation. *Composites Science and Technology*, **66**, 816-826, 2006.

[9] S. Heimbs, S. Heller, P. Middendorf, F. Hähnel and J. Weiße, Low velocity impact on CFRP plates with compressive preload: Test and modeling. *International Journal of Impact Engineering*, **36**, 1182-1193, 2009.

[10] A. Johnson and M. David, Failure Mechanisms and Energy Absorption in Composite Elements under Axial Crush. *KEM*, **488-489**, 638-641, 2011.

[11] G. Pearce, A. Johnson, A. Hellier and R. Thomson, A Stacked-Shell Finite Element Approach for Modelling a Dynamically Loaded Composite Bolted Joint Under in-Plane Bearing Loads. *Applied Composite Materials*, **20**, 1025-1039, 2013.

[12] G. Pearce, A. Johnson, A. Hellier and R. Thomson, A study of dynamic pull-through failure of composite bolted joints using the stacked-shell finite element approach. *Composite Structures*, **118**, 86-93, 2014.

[13] B. Bussadori, K. Schuffenhauer and A. Scattina, Modelling of CFRP crushing structures in explicit crash analysis. *Composites Part B: Engineering*, **60**, 725-735, 2014.

[14] G. Lampeas and K. Fotopoulos, Interlaminar Stresses Calculation Using a Stacked-Shell Finite Element Modeling Approach. *International Journal of Applied Mechanics*, **07**, 1550067, 2015.

[15] S. Abrate, *Impact on composite structures*. Cambridge University Press, 1998.

[16] P. Maimi, J. Mayugo and P. Camanho, A Three-dimensional Damage Model for Transversely Isotropic Composite Laminates. *Journal of Composite Materials*, **42**, 2717-2745, 2008.

[17] Z. P. Bazant, B. H. Oh, Crack band theory for fracture of concrete. *Materials and structures*, **16**, 155-177, 1983.

[18] R. Borg, *Simulation of delamination initiation and growth in fiber composite laminates*. PhD Thesis, Linköpings universitet, Sweden, 2002.

[19] A. Turon, C. Dávila, P. Camanho and J. Costa, An engineering solution for mesh size effects in the simulation of delamination using cohesive zone models. *Engineering Fracture Mechanics*, **74**, 1665-1682, 2007.

[20] A. Riccio, F. Caputo, G. Di Felice, S. Saputo, C. Toscano and V. Lopresto, A Joint Numerical-Experimental Study on Impact Induced Intra-laminar and Inter-laminar Damage in Laminated Composites, *Applied Composite Materials*, 1-19, 2015, doi:10.1007/s10443-015-9457-0.

[21] E. Dvorkin and K. Bathe, A continuum mechanics based four-node shell element for general non-linear analysis. *Engineering Computations*, **01**, 77-88, 1984.

[22] J. O. Hallquist, *LS-DYNA Theory Manual*. Livermore Software Technology Corporation, 2015.

[23] A. Dobyns, Analysis of Simply-Supported Orthotropic Plates Subject to Static and Dynamic Loads. *AIAA Journal*, **19,** 642-650, 1981.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CODE VERIFICATION IN SHELL ANALYSIS BY THE METHOD OF MANUFACTURED SOLUTIONS

Michael Gfrerer, Martin Schanz

Graz University of Technology, Institute of Applied Mechanics
Technikerstr. 4, 8010 Graz, Austria
e-mail: m.gfrerer@tugraz.at

Keywords: verification, elasticity, shell analysis, manufactured solutions

Abstract.

The simulation of the mechanical response of shell structures is of great interest in engineering. Nevertheless, creating simulation software is an error prone task, which needs reliable tests to identify code faults. The aim of this paper is to propose a code verification procedure for shell analysis using the method of manufactured solutions.

We apply order of accurecy tests, where the reference solutions are constructed with the method of manufactured solutions. A set of five verification examples is presented. We have chosen the first example such that the numerical error arises only because of round-off errors. The second and the third example consider either a general stress state or a general geometry. The fourth example considers both a general geometry and a general stress state. We show convergence orders for a Matlab code which implements high order finite elements for a shell model based on the Reissner-Mindlin kinematics. Furthermore, the process of manufactured solution generation is discussed.

1 INTRODUCTION

In order to ensure the reliability of a numerical simulation software, verification and validation ($V\&V$) are unavoidable tasks [1, 2]. Figure 1 illustrates the relation between these activities. An observed real world phenomenon can be modeled by mathematical means, lead-

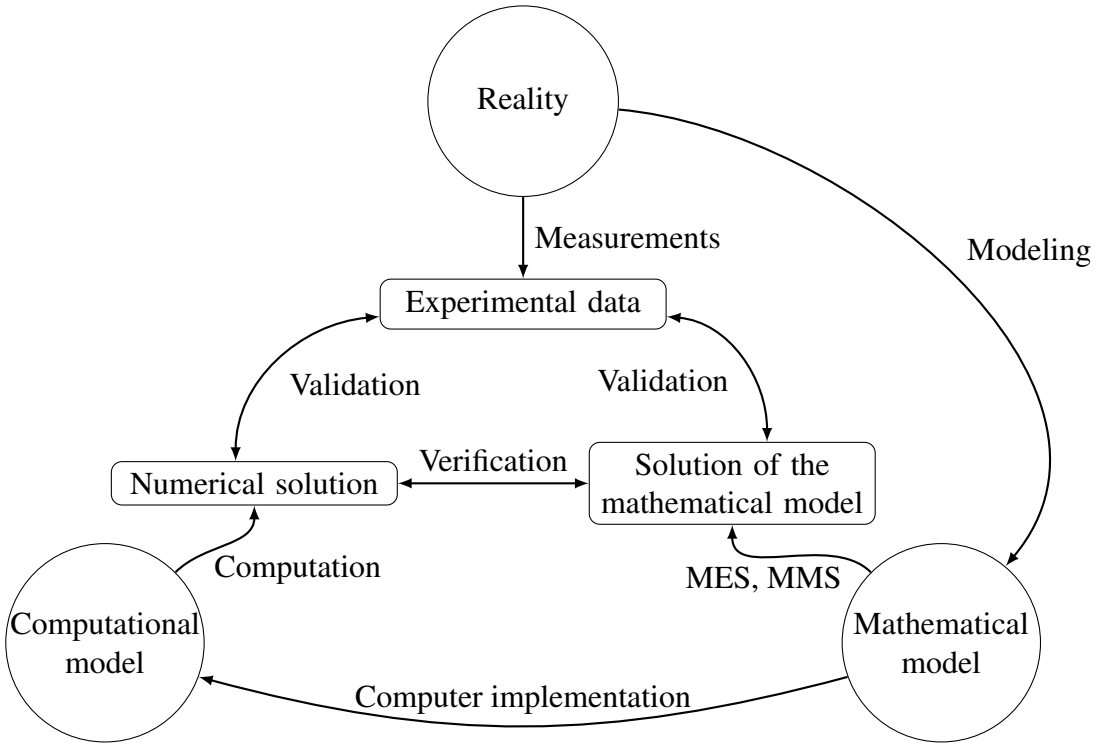


Figure 1: The role of $V\&V$ in numerical simulation, based on [3]

ing to a explicit mathematical model. Since the applicability of the Method of Exact Solutions (MES) to mathematical models describing real world problems is limited, numerical solutions are inevitable. The computer implementation of the mathematical model leads to the computational model. Each of these steps introduces errors, which have to be assessed by $V\&V$. The process of validation determines how accurate the mathematical model represents the real (physical) problem. The aim of verification is to show that the numerical method is able to produce a solution which approximates the exact solution of the mathematical model. Thus, one is interested in the numerical error, which is defined as the difference between the numerical solution and the exact solution. The comparison of numerical results with measurement data is only feasible if the results are not significantly influenced by the numerical error. Therefore, the process of verification has to precede the validation process. Verification activities can be split into code verification and solution verification. Code verification represents the process of demonstrating that the governing equations, as implemented in the code, are solved consistently. Solution verification is the assessment (estimation) of the numerical error in situations where no exact solution is known [4]. According to [5], the most rigorous tests for code verification are the order-of-accuracy tests. For any discretization method, we expect that the discretization error decreases as the mesh is refined. Within an order-of-accuracy test, the observed rate of decrease in the discretization error is compared with the theoretical rate. In order to evaluate the discretization error exact solutions are needed. These exact solutions can be constructed by the Method of Manufactured Solutions (MMS) [6, 7, 8, 5, 1].

Its central idea is to prescribe a solution and to determine a artificial source term which is added to the governing equations, such that the modified equations are fulfilled for the prescribed solution. The MMS can be applied to a wide range of problems. It has been applied to Reynolds-Averaged Navier Stokes solvers [9], nonlinear membrane elements [10] within fluid structure interaction [11, 12], conjugate heat transfer solvers [13], Cahn–Hilliard equation [14], and others.

Within shell analysis, the task is to predict the mechanical response of a thin curved structure. A structure is characterized as thin if one space dimension has a much smaller extension than the other two. Within this geometric setting, a direct solution of the three dimensional problem is difficult. Therefore, many different shell models exist (see [15] for an overview).

In shell analysis, it is a common practice to verify the code against benchmark solutions. One well known set is the shell obstacle course [16], which consists of the Scordelis-Lo problem, the pinched cylinder with a diaphragm, and the hemispherical shell problem. This problem set was introduced in order to compare the performance of different finite elements with respect to locking and accurate representation of rigid body motions. One shortcoming of this test set when used for code verification is the lack of generality in geometry, since only cylindrical and spherical geometries are taken into consideration. Another drawback is that there are only reference displacement values available, which are obtained by numerical simulation. Therefore, no order-of-accuracy tests are possible.

In the present paper, we investigate code verification of finite element codes for shell analysis based on the MMS. Thus, an exact solution is available and an order of accuracy test can be applied. We propose a series of verification examples with increasing complexity. We apply each test to a Matlab [17] research code, which implements high order finite elements for the Reissner-Mindlin theory.

2 SHELL ANALYSIS

The starting point of the mathematical modeling of a shell structure can either be a two dimensional surface (direct approach), or a three dimensional body, both equipped with mechanical properties. The direct approach is based on a deformable Cosserat surface, and has been reviewed in [18]. Its computational aspects have been addressed in a series of papers [19, 20, 21, 22]. Since the strong form of the governing equations is available, the application of the MMS poses no difficulties, other than the necessity of being general.

In case of a three dimensional body which occupies the domain Ω in space, one is interested in solving an elasticity boundary value problem of the form: find the displacement field $\tilde{\mathbf{u}}$ such that

$$\operatorname{div}(\boldsymbol{\sigma}) + \mathbf{b} = 0 \qquad \text{in } \Omega, \tag{1a}$$

$$\boldsymbol{\sigma} = W(\mathbf{E}(\tilde{\mathbf{u}})), \tag{1b}$$

$$\mathbf{u} = \mathbf{u}_D \qquad \text{on } \Gamma_D, \tag{1c}$$

$$\mathbf{t} = \mathbf{t}_N \qquad \text{on } \Gamma_N \tag{1d}$$

are fulfilled. The equilibrium equations are given in (1a), where $\boldsymbol{\sigma}$ stands for the stress tensor and $\mathbf{b}$ for the volume load. In (1b) the material law W assigns a stress tensor to a strain tensor $\mathbf{E}$, which is a function of the displacement field $\tilde{\mathbf{u}}$. The boundary of Ω is denoted by $\Gamma = \Gamma_D \cup \Gamma_N$. On the Dirichlet boundary Γ_D, the displacement vector has a given value $\mathbf{u}_D$, whereas on the Neumann boundary Γ_N, the traction vector $\mathbf{t} = \boldsymbol{\sigma}\mathbf{n}$ has a given value $\mathbf{t}_N$. For a detailed presentation and mathematical foundation of the theory of elasticity, we refer to [23] and [24].

For a given problem, the modeling steps involve the specification of the domain Ω, the respective parts of the boundary Γ, the material law and the strain measure. Finite element solution strategies to (1) can be broadly classified in 3D solid elements, degenerated shell elements, and shell elements based on a shell theory. Within the 3D solid elements approach, the unmodified problem (1) is solved. Thus, the application of the MMS is straightforward. Degenerated shell elements have been introduced in [25] and are based on a continuum discretization of Ω. Further simplifications of the kinematics are introduced in a second step. A shell theory introduces the kinematical assumptions on Ω on the continuous level. Although the degeneration concept and the shell theory approach seem quite different, the formulations are quite similar when based on the same assumptions [26].

In the present paper, we concentrate on linearized elasticity with the following geometric setup. We consider a shell as a three dimensional body which can be described by a two dimensional surface and a thickness parameter. We assume that the surface can be parametrized by a C^2 mapping from a parameter space $\bar{U}$ into the physical space

$$\begin{aligned} \overline{g}:\ \bar{U} \subset \mathbb{R}^2 &\to \bar{\Omega} \subset \mathbb{R}^3 \\ (\theta^1,\theta^2) &\mapsto \overline{g}(\theta^1,\theta^2). \end{aligned} \tag{2}$$

This makes it possible to define two covariant base vectors of the surface

$$\overline{\mathbf{G}}_\alpha = \frac{\partial}{\partial\theta^\alpha}\overline{g} \qquad \alpha = 1,2, \tag{3}$$

and the normal vector to the surface

$$\mathbf{n} = \frac{\tilde{\mathbf{n}}}{\|\tilde{\mathbf{n}}\|} \quad , \quad \tilde{\mathbf{n}} = \overline{\mathbf{G}}_1 \times \overline{\mathbf{G}}_2. \tag{4}$$

The three dimensional body is parametrized by

$$\begin{aligned} g:\ (\bar{U}\times T) \subset \mathbb{R}^3 &\to \Omega \subset \mathbb{R}^3 \\ (\theta^1,\theta^2)\times\theta^3 &\mapsto g(\theta^1,\theta^2,\theta^3) = \overline{g}(\theta^1,\theta^2) + \theta^3\ \mathbf{n}. \end{aligned} \tag{5}$$

We assume $T = [-t/2, t/2]$, with constant thickness t. Instead of searching for $\tilde{\mathbf{u}}$ defined on Ω, we refer to $\mathbf{u}$ defined on $\bar{U}\times T$. Then the components of the linearized strain tensor $\boldsymbol{\epsilon} = \epsilon_{ij}\mathbf{G}^i \otimes \mathbf{G}^j$ with respect to curvilinear coordinates are given by

$$\begin{aligned} \epsilon_{\alpha\beta} &= \frac{1}{2}\left(\mu_\alpha^\xi \overline{\mathbf{G}}_\xi \cdot \frac{\partial \mathbf{u}(\theta^j)}{\partial\theta^\beta} + \mu_\beta^\xi \overline{\mathbf{G}}_\xi \cdot \frac{\partial \mathbf{u}(\theta^j)}{\partial\theta^\alpha}\right), \\ \epsilon_{\alpha 3} &= \frac{1}{2}\left(\mu_\alpha^\xi \overline{\mathbf{G}}_\xi \cdot \frac{\partial \mathbf{u}(\theta^j)}{\partial\theta^3} + \mathbf{n} \cdot \frac{\partial \mathbf{u}(\theta^j)}{\partial\theta^\alpha}\right), \\ \epsilon_{33} &= \frac{\partial \mathbf{u}(\theta^j)}{\partial\theta^3}\cdot\mathbf{n}. \end{aligned} \tag{6}$$

The components of the elasticity tensor $\mathbb{C} = \mathbb{C}^{ijkl}\mathbf{G}_i \otimes \mathbf{G}_j \otimes \mathbf{G}_k \otimes \mathbf{G}_l$ for a linear isotropic material behavior are

$$\begin{aligned} \mathbb{C}^{\alpha\beta\gamma\varphi} &= \lambda \mathrm{G}^{\alpha\beta}\mathrm{G}^{\gamma\varphi} + \mu\left(\mathrm{G}^{\alpha\gamma}\mathrm{G}^{\beta\varphi} + \mathrm{G}^{\alpha\varphi}\mathrm{G}^{\beta\gamma}\right), \\ \mathbb{C}^{\alpha\beta 33} &= \mathbb{C}^{33\alpha\beta} = \lambda \mathrm{G}^{\alpha\beta}, \\ \mathbb{C}^{3\alpha 3\beta} &= \mathbb{C}^{3\alpha\beta 3} = \mathbb{C}^{\alpha 33\beta} = \mathbb{C}^{\alpha 3\beta 3} = \mu \mathrm{G}^{\alpha\beta}, \\ \mathbb{C}^{3\alpha\beta\gamma} &= \mathbb{C}^{\alpha 3\beta\gamma} = \mathbb{C}^{\alpha\beta 3\gamma} = \mathbb{C}^{\alpha\beta\gamma 3} = 0, \\ \mathbb{C}^{333\alpha} &= \mathbb{C}^{33\alpha 3} = \mathbb{C}^{3\alpha 33} = \mathbb{C}^{\alpha 333} = 0, \\ \mathbb{C}^{3333} &= \lambda + 2\mu. \end{aligned} \tag{7}$$

In equations (6) and (7), $\mu_\alpha^\xi = \delta_\alpha^\xi - \theta^3 \overline{G}^{\xi\gamma}(\overline{\mathbf{G}}_{\alpha,\gamma} \cdot \mathbf{n})$ are the components of the shifter tensor, $\overline{G}^{\alpha\beta}$ are the components of the inverse of the metric tensor $\overline{G}_{\alpha\beta} = \overline{\mathbf{G}}_\alpha \cdot \overline{\mathbf{G}}_\beta$, and $\mathrm{G}^{\alpha\beta}$ are the components of the inverse of $\mathrm{G}_{\alpha\beta} = \mu_\alpha^\gamma \mu_\beta^\varphi \overline{G}_{\gamma\varphi}$. λ, μ are the Lamé constants. For the sake of brevity, we used the notation $(\cdot)_{,\alpha} = \frac{\partial(\cdot)}{\partial\theta^\alpha}$.

In the present paper, we consider the classical Reissner-Mindlin model, where the thickness expansion for the displacement field reads

$$\mathbf{u}(\theta^1, \theta^2, \theta^3) = u^i(\theta^1, \theta^2)\mathbf{e}_i + \theta^3 \sum_{\alpha=1}^{2} v^\alpha \overline{\mathbf{G}}_\alpha. \tag{8}$$

In (8), u^i are the displacement components of the mid surface and v^α are the rotations. In total, five parameters are used. Furthermore, as the result of the plane stress assumption, λ has to be replaced with $\frac{2\mu\lambda}{2\mu+\lambda}$.

The finite element formulation is based on the virtual work

$$\int_\Omega \hat{\boldsymbol{\epsilon}} : \mathbb{C} : \boldsymbol{\epsilon} \, \mathrm{d}x = \int_\Omega \hat{\mathbf{u}} \cdot \mathbf{b} \, \mathrm{d}x + \int_\Gamma \hat{\mathbf{u}} \cdot \mathbf{t} \, \mathrm{d}s_x, \tag{9}$$

where $\hat{\mathbf{u}}$ is the virtual displacement field and $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}(\hat{\mathbf{u}})$. In order to derive finite elements, the parameter space is meshed with quadrilateral elements. The parameters of the shell model are discretized with integrated Legendre polynomials of order p. The discretized virtual work statement (9) results in the linear equation system

$$K\, u = f, \tag{10}$$

where K is the stiffness matrix resulting from the integral on the left side of (9), u the unknown degrees of freedom and f the load vector resulting from the integrals on the right side of (9). All integrals are computed using Gauss-Legandre quadrature for in-plane as well as through the thickness. The system of linear equations (10) is solved with a direct solver.

3 CODE VERIFICATION

In the present paper, we apply code verification based on order-of-accuracy tests and the MMS to a code for solving the shell formulations given in section 2. The necessary prerequisite to apply it to a numerical schema is the knowledge of a formal order of convergence and exact solutions. Thus, an estimate of the type

$$||u^{exact} - u^{numerical}|| \leq C\, h^q \, ||u^{exact}||, \tag{11}$$

where C is a constant and h is a characteristic element size, has to be known. Then q is called the formal order of convergence with respect to the norm $||\cdot||$. For two meshes with characteristic element sizes h_1 and h_2, the experimental order of convergence (eoc) is defined as

$$eoc = \frac{\log(e_1) - \log(e_2)}{\log(h_1) - \log(h_2)}, \tag{12}$$

where

$$e_i = ||u^{exact} - u_{h_i}^{numerical}|| \tag{13}$$

is the numerical error corresponding to the discretization h_i. The code is verified, if the eoc matches the formal order of convergence within the asymptotic range. For the finite element

method applied in this paper, we expect $q = p + 1$ for the error in the L_2 norm for smooth solutions.

In order to evaluate (13), an exact solution has to be available. Within the MMS, such a solution is prescribed. The source term is determined such that the chosen solution fulfills the governing equations. In particular, we perform the following steps:

1. Choose the form of the problem domain, $i.e.$ specify the surface parametrization (2)
2. Choose the form of the manufactured solution $\mathbf{u}^M$, $i.e$ specify the parameters in (8)
3. Derive the modified governing equations, $i.e.$ compute an artificial source term $\mathbf{b}^M = -\mathrm{div}(\mathbb{C} : \epsilon(\mathbf{u}^M))$
4. Solve the discrete form of the modified governing equations on multiple meshes, $i.e.$ solve (9) with $\mathbf{b} = \mathbf{b}^M$
5. Evaluate the numerical error (13) and the eoc (12)
6. Apply the order of convergence test

Additionally to [27], we added the first point. The finite element code needs the components b^i of the source term $\mathbf{b} = b^i \mathbf{e}_i$ with respect to the global Cartesian frame. They are given by

$$b^i = -\left(\sigma^{jl}_{,j} + \sigma^{kl}\Gamma_{kj}{}^{j} + \sigma^{jk}\Gamma_{kj}{}^{l}\right) J^i_l \tag{14}$$

with the Christoffel symbols of second kind $\Gamma_{kj}{}^{l} = \mathbf{G}_{k,j}\mathbf{G}^l$ and $J^i_l = \mathbf{G}_l \cdot \mathbf{e}^i$. As noted in [28], an arbitrary choice of manufactured solution leads easily to a very long computer code for the source term. We have observed this for general complex curved surfaces too. We use the computer algebra system Mathematica™ [29] to derive the source terms in an automated way. Nevertheless, a general verification example assessing all features of the code results in a source code with more than 10^6 characters, where simplifications are prohibitively time consuming. Therefore, we compute only the derivatives analytically and the algebraic operations are performed numerically. Instead of exporting a function for b^i, we export functions taking the spatial position as input argument for

- strain tensor ϵ_{ij},
- derivatives of the strain tensor $\epsilon_{ij,k}$,
- covariant base vectors $\mathbf{G}_i$,
- derivatives of the covariant base vectors $\mathbf{G}_{i,j}$.

We numerize these quantities and compute

$$\begin{aligned}
G_{ij} &= \mathbf{G}_i \cdot \mathbf{G}_j, \\
[G^{ij}] &= [G_{ij}]^{-1}, \\
J_l^i &= \mathbf{G}_l \cdot \mathbf{e}^i, \\
\Gamma_{kj}{}^{l} &= \mathbf{G}_{k,j} \cdot \mathbf{G}_i G^{il}, \\
G^{ij}_{,k} &= -\Gamma_{kl}{}^{j} G^{li} - \Gamma_{kl}{}^{i} G^{lj}, \\
\sigma^{jl} &= \lambda \left(G^{ij} G^{kl} \epsilon_{kl} + 2\mu G^{ik} G^{jl} \epsilon_{kl} \right), \\
\sigma^{jl}_{,j} &= \lambda \left(G^{ij}_{,j} G^{kl} \epsilon_{kl} + G^{ij} G^{kl}_{,j} \epsilon_{kl} + G^{ij} G^{kl} \epsilon_{kl,j} \right) \\
&\quad + \mu \left(G^{ik}_{,j} G^{jl} \epsilon_{kl} + G^{ik} G^{jl}_{,j} \epsilon_{kl} + G^{ik} G^{jl} \epsilon_{kl,j} \right).
\end{aligned} \tag{15}$$

Then the source term is obtained with (14).

Remark: In the present implementation, we keep the exact surface parametrization during the whole computation. Therefore, we are able to evaluate the error in a continuous norm. In case of an approximation of the surface, one has to use the discrete norm [12]

$$e = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(u^{exact}(x_n) - u^{numerical}(x_n) \right)^2}. \tag{16}$$

Furthermore, the influence of geometry approximation on the formal order of convergence has to be considered. For a discussion of this 'variational crime', we refer to [30].

4 VERIFICATION EXAMPLES

In the following, we provide a set of verification examples with increasing complexity. In principle, it would be sufficient to consider only the most general case in order to verify the code, since the special cases are included in the general case [31]. However, in order to have confirmation exercises, we suggest special cases where only parts of the code should be exercised.

In all examples, we specify a surface parametrization (2) and the displacement parameters for the shell model (8). In all examples, we use the parameter space $(\theta^1, \theta^2) \in [0, 0.56] \times [0, 0.65]$ and the thickness $t = 0.07$. All quantities are defined in the International System of Units (SI) and their derived expressions. Therefore, we omit units for the input numbers. We use the material parameters $\lambda = \mu = 4000$, which corresponds to Young's modulus $E = 1000$ and Poisson's ratio of $\nu = 0.25$.

In the verification examples, we use two types of displacement fields. Displacement field A is chosen such that it can be represented exactly with each discretization. Displacement field B is chosen to be general to assess all features of the code. The displacement fields for the shell model are given in Table 1.

Table 1: Prescribed displacement parameters for (8)

	displacement field A	displacement field B
$u^1 =$	θ^1	$\sin(\pi\theta^1)\cos(\pi\theta^2)$
$u^2 =$	θ^2	$\cos(\pi\theta^1)\sin(\pi\theta^2)$
$u^3 =$	$\theta^1\theta^2$	$\sin(\pi\theta^1\theta^2)$
$v^1 =$	$\theta^1\theta^2$	$\sin(\pi\theta^1\theta^2)$
$v^2 =$	$\theta^1\theta^2$	$\sin(\pi\theta^1\theta^2)$

4.1 Example 1: plane geometry, exact representation of solution

The goal of the first example is to determine the influence of round-off errors. To this end, we have chosen the plane geometry

$$\begin{aligned} x &= \theta^1, \\ y &= \theta^2, \\ z &= 0, \end{aligned} \tag{17}$$

and displacement field A, which can be exactly represented in the discrete system. The remaining error is due to round-off errors. We have executed the code for ansatz orders up to order three for single precision, as well as for double precision. In Figure 2, the errors for different discretization levels are shown. The error levels are in accordance with single and double precision.

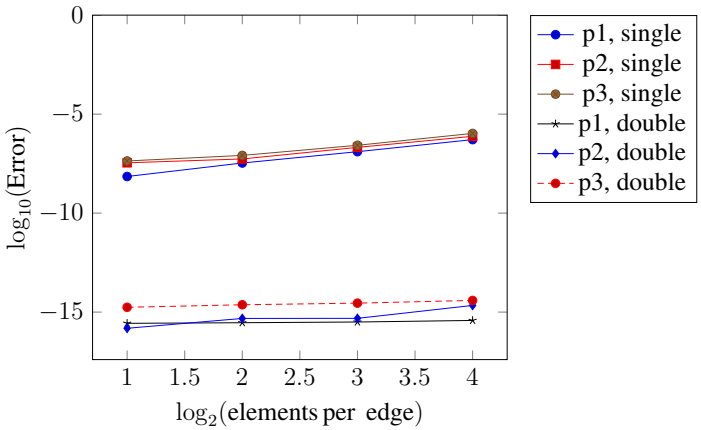


Figure 2: Numerical error for example 1

4.2 Example 2: plane geometry, general solution

In the second example, the discretization error at a plane geometry is assessed. We therefore use (17) with displacement field B. The development of the errors and the eoc for ansatz orders up to order six are shown in Figure 3. For the ansatz orders one to four the eoc tends to the respective formal order of convergence. However, for ansatz orders five and six, the eoc does not agree with the formal order of convergence. In these cases, the numerical error is dominated

by the error introduced due to round-off and not by the discretization error. This effect is also visible in the subsequent examples. The results verify the tested capabilities of the code.

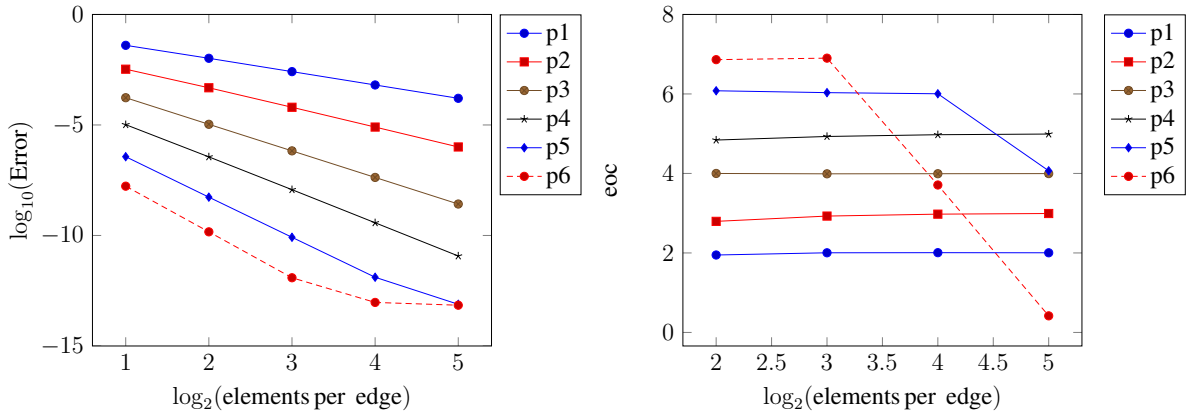


Figure 3: Development of the numerical error and the *eoc* for example 2

4.3 Example 3: general geometry, exact representation of solution

In the third example, a general geometry with displacement field A is considered. The surface parametrization reads

$$\begin{aligned} x &= \theta^1, \\ y &= \theta^2, \\ z &= (\theta^1)^2 - (\theta^2)^2. \end{aligned} \tag{18}$$

The development of the errors and the *eoc* for ansatz orders up to order six are shown in Figure 4. The numerical error present in the example stems from the numerical integration of the integrals in (9) and round-off, since displacement field A can be exactly discretized. We have set the number of quadrature points in both in-plane directions to the ansatz order. However, as long as the numerical error is dominated by the integration error, the *eoc* is higher or equal as the formal order of convergence regarding the discretization of the displacement field. Therefore, the integration error will not hamper the overall convergence rate. We remark that, with more quadrature points the *eoc* can be increased in this example.

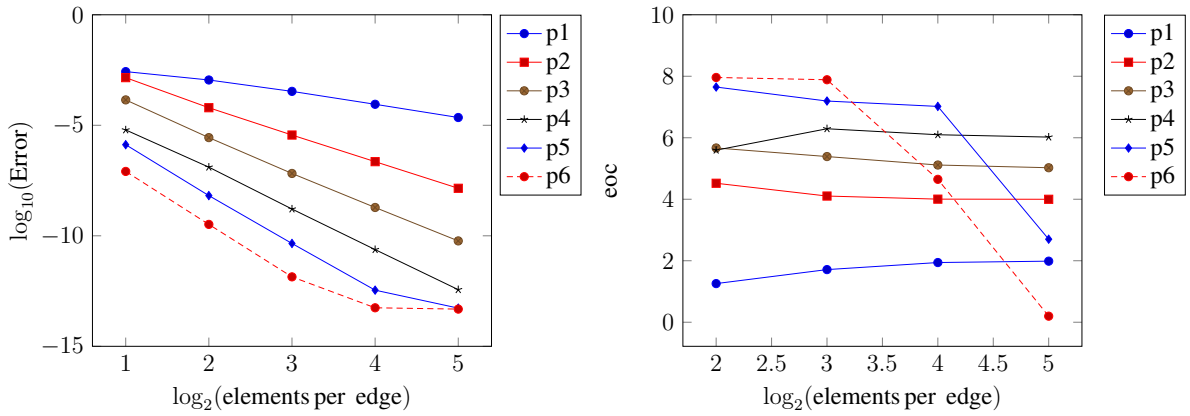


Figure 4: Development of the numerical error and the *eoc* for example 3

4.4 Example 4: general geometry, general solution

The fourth example consists of the general surface parametrization

$$\begin{aligned} x &= \theta^1 + \frac{(\theta^2)^2}{2}, \\ y &= \theta^2 - (\theta^1)^2, \\ z &= (\theta^1)^2 - (\theta^2)^2, \end{aligned} \tag{19}$$

and the general displacement field B. Thus, this example assesses all features of the code. The development of the errors and the eoc for ansatz orders up to order six are shown in Figure 5. Again, the error corresponding to ansatz orders five and six is limited by the round-off error. Thus, the eoc drops. For all other ansatz orders the eoc tend to the formal order of convergence. Thus, the code passes this verification example.

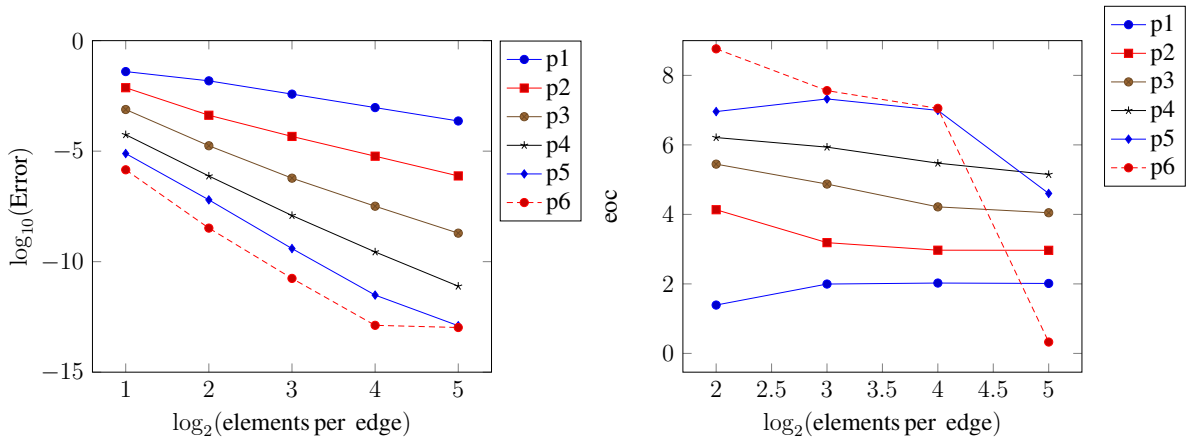


Figure 5: Development of the numerical error and the eoc for example 4

5 CONCLUSIONS

This paper addresses code verification of a finite element code for shell analysis based on the MMS. The MMS provides a powerful procedure to construct exact solutions to the governing equations. Because of the complexity of the resulting source term, we decided to partly numerized its computation. The derivatives are computed analytically and the remaining algebraic operations are performed on numerized quantities. We successfully apply the verification examples to a finite element code for Reissner-Mindlin shells. The approach based on automatic differentiation presented in [28] could further reduce the source term complexity. Nevertheless, the verification of the manufactured solution itself remains an open question.

REFERENCES

[1] W. L. Oberkampf, C. J. Roy, *Verification and validation in scientific computing*. Cambridge University Press, 2010.

[2] J. T. Oden, T. Belytschko, J. Fish, T. Hughes, C. Johnson, D. Keyes, A. Laub, L. Petzold, D. Srolovitz, S. Yip, Revolutionizing engineering science through simulation. *National Science Foundation Blue Ribbon Panel Report*, **65**, 2006.

[3] S. Schlesinger, R. E. Crosbie, R. E. Gagné, G. S. Innis, C. S. Lalwani, J. Loch, R. J. Sylvester, R. D. Wright, N. Kheir, D. Bartos, Terminology for model credibility. *Simulation*, **32**, 103–104, 1979.

[4] W. L. Oberkampf, T. G. Trucano, C. Hirsch, Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, **57**, 345–384, 2004.

[5] C. J. Roy, C. Nelson, T. Smith, C. Ober, Verification of Euler/Navier–Stokes codes using the method of manufactured solutions. *International Journal for Numerical Methods in Fluids*, **44**, 599–620, 2004.

[6] S. Steinberg, P. J. Roache, Symbolic manipulation and computational fluid dynamics. *Journal of Computational Physics*, **57**, 251–284, 1985.

[7] T. Shih, A procedure to debug computer programs. *International Journal for Numerical Methods in Engineering*, **21**, 1027–1037, 1985.

[8] P. J. Roache, *Verification and validation in computational science and engineering*. Hermosa, 1998.

[9] L. Eça, M. Hoekstra, A. Hay, D. Pelletier, Verification of RANS solvers with manufactured solutions. *Engineering with computers*, **23**, 253–270, 2007.

[10] R. Fisch, J. Franke, R. Wüchner, K. Bletzinger, Code verification examples of a fully geometrical nonlinear membrane element using the method of manufactured solutions. K.-U.Bletzinger, B. Kröplin, E. Oñate, eds., *Proceedings Structural Membranes 2013*, Munich, Germany, October 9-11, 2013.

[11] R. Fisch, J. Franke, R. Wüchner, K. Bletzinger, Code Verification of a Partitioned FSI Environment for Wind Engineering Applications using the Method of Manufactured Solutions. E. Oñate, J. Oliver, A. Huerta, eds., *5th European Conference on Computational Mechanics (ECCM V)*, Barcelona, Spain, July 20-15, 2014.

[12] R. Fisch, *Code Verification of Partitioned FSI Environments for Lightweight Structures*, Vol. 23 of *Schriftenreihe des Lehrstuhls für Statik der Technischen Universität München*. Lehrstuhl für Statik, Technische Universität München, 2014.

[13] A. Veeraragavan, J. Beri, R. Gollan, Use of the method of manufactured solutions for the verification of conjugate heat transfer solvers. *Journal of Computational Physics*, **307**, 308–320, 2016.

[14] M. Kästner, P. Metsch, R. De Borst, Isogeometric analysis of the Cahn–Hilliard equation–a convergence study. *Journal of Computational Physics*, **305**, 360–371, 2016.

[15] M. Bischoff, K.-U. Bletzinger, W. Wall, E. Ramm, Models and Finite Elements for Thin-Walled Structures. E. Stein, R. de Borst, T. Hughes, eds., *Encyclopedia of computational mechanics*, Vol. 2, 59–137, Wiley Online Library, 2004.

[16] T. Belytschko, H. Stolarski, W. K. Liu, N. Carpenter, J. S. Ong, Stress projection for membrane and shear locking in shell finite elements. *Computer Methods in Applied Mechanics and Engineering*, **51**, 221–258, 1985.

[17] MATLAB, *Version 8.5.0.197613 (R2015a)*. The MathWorks Inc., Natick, Massachusetts, 2015.

[18] J. Altenbach, H. Altenbach, V. A. Eremeyev, On generalized Cosserat-type theories of plates and shells: a short review and bibliography. *Archive of Applied Mechanics*, **80**, 73–92, 2010.

[19] J. C. Simo, D. D. Fox, On a stress resultant geometrically exact shell model. Part I: Formulation and optimal parametrization. *Computer Methods in Applied Mechanics and Engineering*, **72**, 267–304, 1989.

[20] J. C. Simo, D. D. Fox, M. Rifai, On a stress resultant geometrically exact shell model. Part II: The linear theory; computational aspects. *Computer Methods in Applied Mechanics and Engineering*, **73**, 53–92, 1989.

[21] J. C. Simo, D. D. Fox, M. S. Rifai, On a stress resultant geometrically exact shell model. Part III: Computational aspects of the nonlinear theory. *Computer Methods in Applied Mechanics and Engineering*, **79**, 21–70, 1990.

[22] J. C. Simo, M. Rifai, D. Fox, On a stress resultant geometrically exact shell model. Part IV: Variable thickness shells with through-the-thickness stretching. *Computer methods in applied mechanics and engineering*, **81**, 91–126, 1990.

[23] J. E. Marsden, T. J. Hughes, *Mathematical foundations of elasticity*. Dover Publications, New York, 1983.

[24] G. A. Holzapfel, *Nonlinear solid mechanics*, Vol. 24. Wiley Chichester, 2000.

[25] S. Ahmad, B. M. Irons, O. Zienkiewicz, Analysis of thick and thin shell structures by curved finite elements. *International Journal for Numerical Methods in Engineering*, **2**, 419–451, 1970.

[26] N. Buechter, E. Ramm, Shell theory versus degenerationa comparison in large rotation finite element analysis. *International Journal for Numerical Methods in Engineering*, **34**, 39–59, 1992.

[27] C. J. Roy, Review of code and solution verification procedures for computational simulation. *Journal of Computational Physics*, **205**, 131–156, 2005.

[28] N. Malaya, K. C. Estacio-Hiroms, R. H. Stogner, K. W. Schulz, P. T. Bauman, G. F. Carey, MASA: a library for verification using manufactured and analytical solutions. *Engineering with Computers*, **29**, 487–496, 2013.

[29] Wolfram Research, Inc., *Mathematica, Version 10.0*. Champaign, Illinios, 2014.

[30] G. Strang, G. J. Fix, *An analysis of the finite element method*, Vol. 212. Prentice-Hall Englewood Cliffs, N.J., 1973.

[31] P. J. Roache, Code verification by the method of manufactured solutions. *Journal of Fluids Engineering*, **124**, 4–10, 2002.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EXPERIMENTAL VALIDATION OF A MODEL UPDATE PROCEDURE FOCUSING ON SMALL GEOMETRIC DEVIATIONS

Thomas Maywald[1], **Arnold Kühhorn**[1] **and Sven Schrape**[2]

[1] Brandenburg University of Technology (BTU)
Siemens-Halske-Ring 14
D-03046 Cottbus, Germany
e-mail: {thomas.maywald,kuehhorn}@b-tu.de

[2] Rolls-Royce Deutschland Ltd & Co KG
Eschenweg 11
D-15827 Blankenfelde-Mahlow, Germany
e-mail: sven.schrape@rolls-royce.com

Keywords: Model Update, Mesh Morphing, Model Validation, Vibration Measurements, Geometric Mistuning, Blisk Vibrations.

Abstract. *This contribution presents a model update procedure and its experimental validation using the example of a blade integrated disk rotor. This so called blisk is discretized using the finite element method. It is well known that numerical blisk models based on the ideal tuned design show major differences in structural dynamic behavior compared to the real rotor. In this context a modification of the mechanical simulation model should lead to a better accordance of numerical results and the real blisk characteristics. The described model update procedure utilizes data of an optical 3D measurement system. Using this data enables to identify geometric deviations between the ideal design and its real counterpart. Within the update procedure the originally tuned finite element mesh is modified in order to match the measured geometry of the real part. This is done by defining several morph regions. The outer surface nodes of these morph regions change their position along the surface normal vector until they meet the defined deviation constraint. Based on eigenvalue calculations employing free boundary conditions the sensitivity of structural dynamic behavior is shown with respect to small geometric changes. Finally computed eigenvalues and eigenvectors of the updated simulation model are compared with vibration measurement data. A laser Doppler vibrometer is used to detect the vibration responses of the impact excited structure. All experiments are carried out under technical vacuum conditions in order to minimize ambient air damping. In the context of an experimental modal analysis this low damping condition helps to identify more natural frequencies of the investigated structure. This leads to a much more efficient model validation.*

1 INTRODUCTION

One of the main objectives in engineering practice is to predict structural behavior due to different load conditions. There are several possibilities reaching this target. One of the most popular ways to simulate mechanical systems behavior is to discretize a structure using the Finite Element Method (FEM). The transformation of a continuum into finite elements is accompanied by several challenges. On the one hand this discretization necessarily leads to differences between the CAD geometry of a part and its simulation model. On the other hand the manufactured part is affected by errors. That means that there are deviations between the digitally designed CAD model and its real counterpart. Commonly deviations between a simulation model and the real mechanical system cause an inaccurate prediction of structural behavior. Therefore several model update strategies are available in order to harmonize numerical prediction with experimental experiences.

This contribution presents a model update procedure which focuses on small manufacturing driven geometry changes of a rotating aircraft engine part. Object of investigation is a blade integrated disk rotor (blisk). Here, the milling process causes small geometric deviations between real part and the designed rotor. These deviations from the ideal design intention, known as 'mistuning', come along with a number of particularities in structural dynamics. Thus, mistuning can cause localized mode shapes which lead to a concentration of vibration energy in several or even a single blade [1-5]. By this the regular character of the mode shapes gets lost. In presence of a dynamic load case these blades are subject to higher magnitudes of vibration induced stresses compared to the ideal design.

Traditionally these mistuning effects are taken into account by a stiffness proportional model update approach [6-9]. The central point of this strategy is to modify the Young's modulus or respectively the stiffness matrix of each rotor blade until the finite element model is able to predict measured natural frequencies of the real rotor blade. It has been shown by a number of researchers that this is a sufficient way to simulate mistuning effects [6-9]. The advantage of this method is that there is only one model update parameter per blade. Therefore this model modification is easy to handle and no extensive computational resources are needed. Unfortunately, comprehensive measurements are necessary to identify the mistuned natural frequencies of the real system which must be known for model update. Furthermore the reliability of this stiffness proportional approach is limited to a comparatively small frequency range. This means the model update has to be done separately for each frequency range of interest.

Especially these disadvantages motivate the community to look for alternative ways of model update. A promising approach of the recent past is to consider the geometric deviations due to manufacturing within the simulation model. Existing publications dealing with this topic clearly indicate the relationship between geometry deviations and simulation model dynamics [10-12]. But finally a validation of the methodology turned out to be difficult due to the small manufacturing tolerances which have to be modeled. For this reason the present paper discusses an updated model of an industrial test blisk which could be validated for a comparatively large frequency range.

2 IDEALIZED STRUCTURAL DYNAMICS

To get a first impression of structural dynamics behavior the modal parameters of the initial rotor design has to be determined. The homogeneous equation of motion reads:

$$M\ddot{x}(\mathrm{t}) + D\dot{x}(\mathrm{t}) + Kx(\mathrm{t}) = 0. \tag{1}$$

Herein M represents the mass matrix, D the structural damping matrix and K the stiffness matrix of the blade integrated disk. Further on an undamped system is assumed ($D = 0$). The assumption results from knowledge of modal structural damping ratios of those structures which are considerably lower than $\zeta = 0.3\%$ [13, 14]. The influence of those structural damping values is negligibly in case of free blisk vibrations. Consequently the modal parameters sought for could be identified by formulating and solving the undamped eigenvalue problem

$$(M\lambda_k^2 + K)\varphi_k e^{\lambda_k t} = 0\,, \tag{3}$$

where λ_k is the k-th eigenvalue and φ_k is the k-th eigenvector of the system. To compute the needed system matrices M and K a finite element mesh has to be created. Figure 1 shows the finite element mesh for one sector with about 53.500 continuum elements of type C3D10.

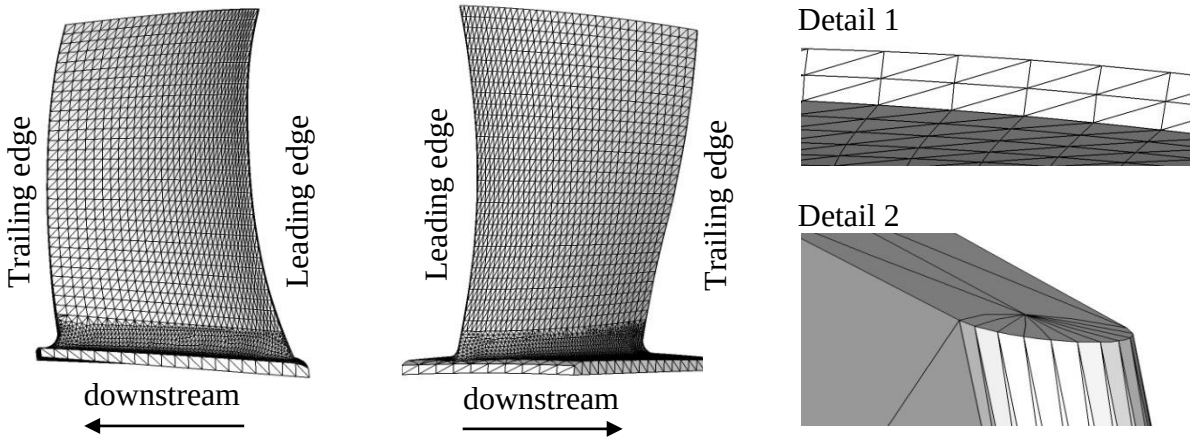
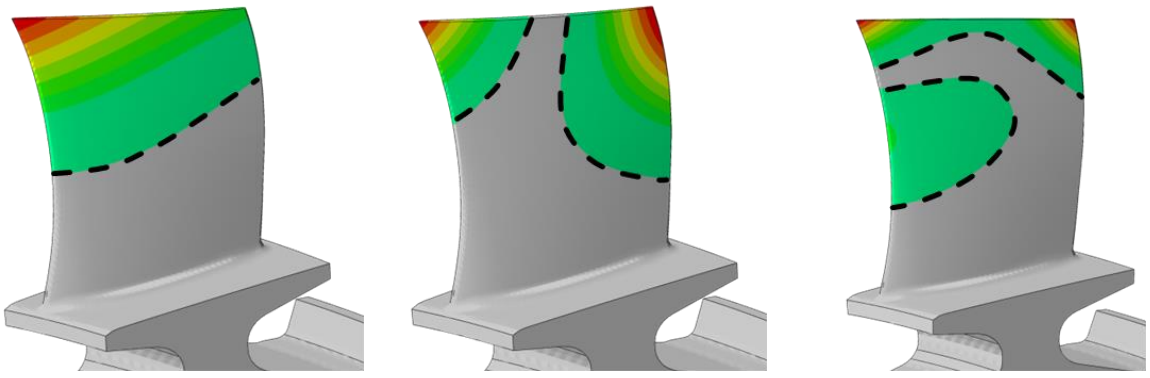


Figure 1: Finite element sector mesh of a blade integrated disk rotor (disk mesh not shown)

Two elements were used in blade thickness direction (Figure 1, Detail 1). The Leading edge curve is discretized through 12 nodes and the trailing edge curve through 21 nodes (Figure 1, Detail 2). Because all sectors are identical for the tuned design intention the modal parameters of the whole rotor were calculated by using a single sector model with cyclic symmetry boundary conditions. Figure 2 illustrates the first three computed mode shapes whose displacements are located at the blade section of the rotor.



Blade mode 1 Blade mode 2 Blade mode 3

Figure 2: Computed blade mode shapes of the tuned simulation model with cyclic symmetry boundary conditions

Furthermore the vibration modes of the tuned model are categorized according to three major criteria. The cyclic symmetry mode index (CSM), the distribution of strain energy as well as the mode shape similarity are considered in order to arrange all resonances up to $f_{norm} = 4.5$ in a nodal diameter map shown in Figure 3. The maximum cyclic symmetry mode index, that can occur, is limited by the number of blades N. In case of an even number of blades it is given by

$$\mathrm{CSM}_{\max} = \frac{\mathrm{N}}{2}, \tag{4}$$

for an odd number of blades follows

$$\mathrm{CSM}_{\max} = \frac{\mathrm{N}-1}{2}. \tag{5}$$

Secondly a mode classification according to blade-dominated, disk-dominated or mixed mode is done by the assessment of the strain energy distribution as introduced in [15]. A strain energy localization of more than 86% within the blade region indicates a blade dominated mode. Accordingly a disk dominated mode is characterized by 75 % strain energy localization inside the disk. The remaining modes are summarized as mixed modes. These limits are problem specific values and may vary for different blisk applications.

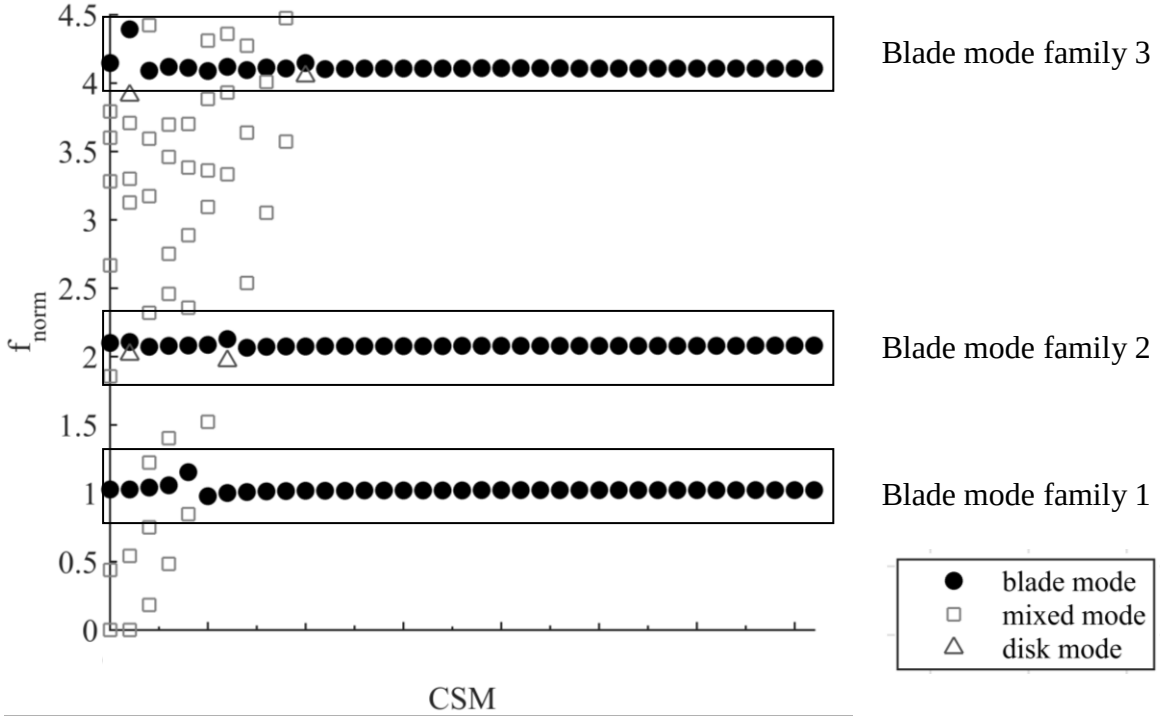


Figure 3: Nodal diameter plot

In a last step groups of classified modes with almost the same blade mode shape are identified by using the Modal Assurance Criterion (MAC) as explained in [16]. At this point several cyclic symmetry modes with similar or nearly the same blade mode shapes (MAC > 0.9) establish a family of modes. Inside Figure 3 the identified blade mode families are highlighted by a rectangle.

3 MODEL UPDATE PROCEDURE

As described in the introduction the manufacturing process always causes deviations between the ideal design intention and the real part. Considering these differences is of major importance in order to compute structural dynamics characteristics as accurate as possible. A two step procedure is needed to improve the ideal simulation model of the blade integrated disk described in chapter 2.

First of all the real geometry of the part has to be digitalized. To get the requested high resoluted surface information an optical 3D geometry measurement system is used as shown in Figure 4.

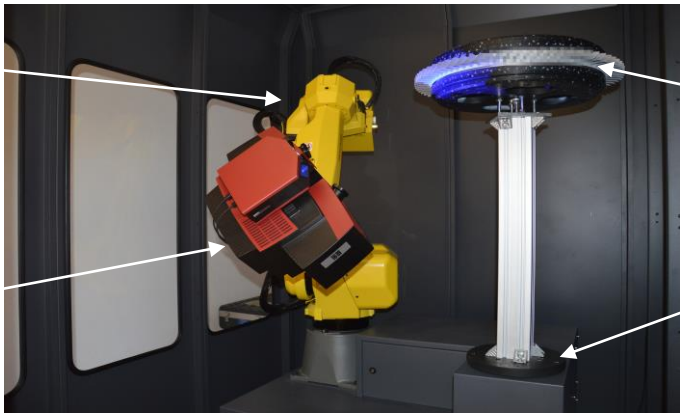


Figure 4: Non contact geometric measurement system of TU Dresden - Chair of Turbomachinery and Flight Propulsion

These high resolution measuring cameras are well proven in engineering practice and represent a state of the art measurement solution for reversed engineering. Precision of such a system has been proven to be within 2.5 µm [17]. One major advantage of such a non contact solution is the possibility to get a large number of coordinate informations very fast. Furthermore this technology allows capturing coordinates from areas which are difficult to access and could not be taken into account using conventional coordinate measurement machines. More details, advantages and disadvantages concerning optical coordinate measurements are published in [17-19]. The experimental setup which was used to get the results presented in the following includes an optical scan head, a robot arm for moving the scan head automatically and a rotatable table that carries the test blisk.

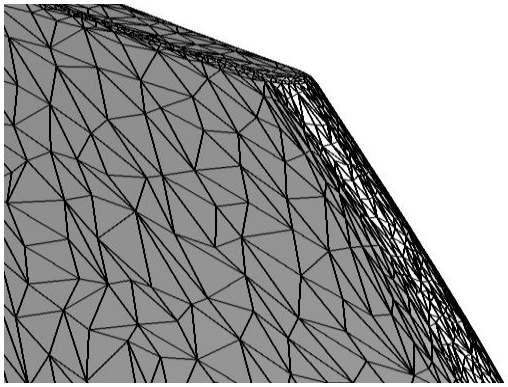


Figure 5: Triangulated surface based on optical measurements in the tip area of an exemplary trailing edge

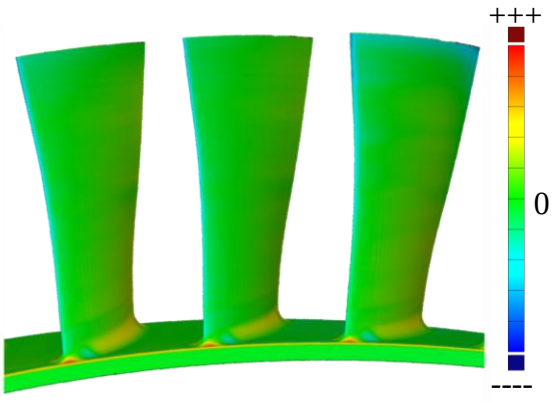


Figure 6: Deviation plot for an exemplary blade ring section - initial finite element mesh versus measured part geometry

After digitalizing the blisk the resulting point cloud data were processed to a triangulated surface mesh of the blisk shown in Figure 5. Due to experimental setup restrictions it is not possible to scan the disk of the part as well. Because of that the following data evaluation focuses on the blade ring (Figure 6). The deviations between tuned design and the real part are quantified by calculating the normal to surface distances for every node of the surface mesh. As shown in Figure 6 the deviations remain in lower triple digit micro meter range (triple + or triple −). Positive or respectively negative values indicate additional or missing material. Large deviations are mainly identified in blade leading, tip and trailing edge region. The green color indicates that the mean error of the remaining blade regions is much lower in the single digit micrometer range.

Within the second step of the model update procedure the finite element mesh has to be modified in order to match the experimentally identified surface mesh. The mesh modification shape is calculated evaluating the deviations for each mesh node of the reference mesh (Figure 1) along its element normal. The shape illustrated in Figure 7 is used to modify the initial finite element mesh (Figure 1) to ensure the best representation of the measured surface (Figure 5). Each node is shifted along its shape vector. A penetration of the measured surface mesh is not permitted in this context.

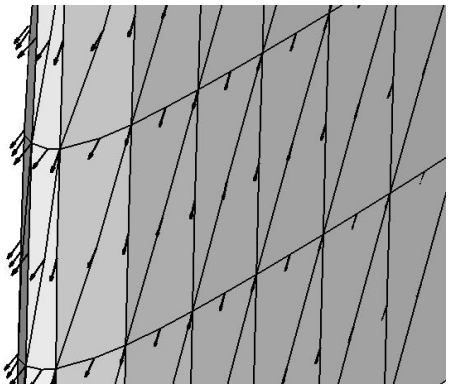


Figure 7: Mesh modification shape for exemplary leading edge

Figure 8: Deviation plot for a an exemplary blade ring section - updated finite element mesh versus measured part geometry

After finishing mesh morphing the updated model has to be compared with the measured part geometry. The corresponding deviation plot (Figure 8) shows nearly no visible differences between updated mesh and measured real part. On average the remaining deviations are in range of plus minus one digit micrometer. For this reason model modification is assessed to be sufficient.

4 MODEL VALIDATION METHOD

The mesh updating procedure presented in Chapter 3 affects both the calculated mass matrix of the investigated structure and the stiffness matrix. Compared to Chapter 2 the originally assumed cyclic symmetry of the blisk is disturbed due to the introduced geometric modifications. Commonly there is an influence on the resulting modal parameters evaluated by solving the undamped eigenvalue problem as described in Equation 3. These updated modal characteristics have to be checked for validity. Commonly modal parameters of the real part are identified by an experimental modal analysis and compared to their numerical counterparts. As shown in Figure 3 natural frequencies of a blade integrated disk are appearing very close to each other. Therefore the comprehensive natural frequency identification is dif-

ficult due to modal coupling effects. For this reason a vacuum test stand is used in order to minimize ambient air damping. In an experimental modal analysis this low damping condition helps to identify more natural frequencies of the investigated structure since modal coupling is decreasing.

4.1 Experimental Setup

The experimental determination of modal parameters within low pressure condition is carried out inside a vacuum test stand (Figure 9). Vibration excitation is realized with an impact mechanism. In addition laser scanning vibrometer measurements are applied to determine one frequency response function per blade in the frequency range of the blade modes 1 to 3 (Figure 3). A foam cuboid is used to support the blisk and ensure nearly free boundary conditions. Disturbing acoustical resonances are prevented by facing the inner surface of the chamber with nubby foam material.

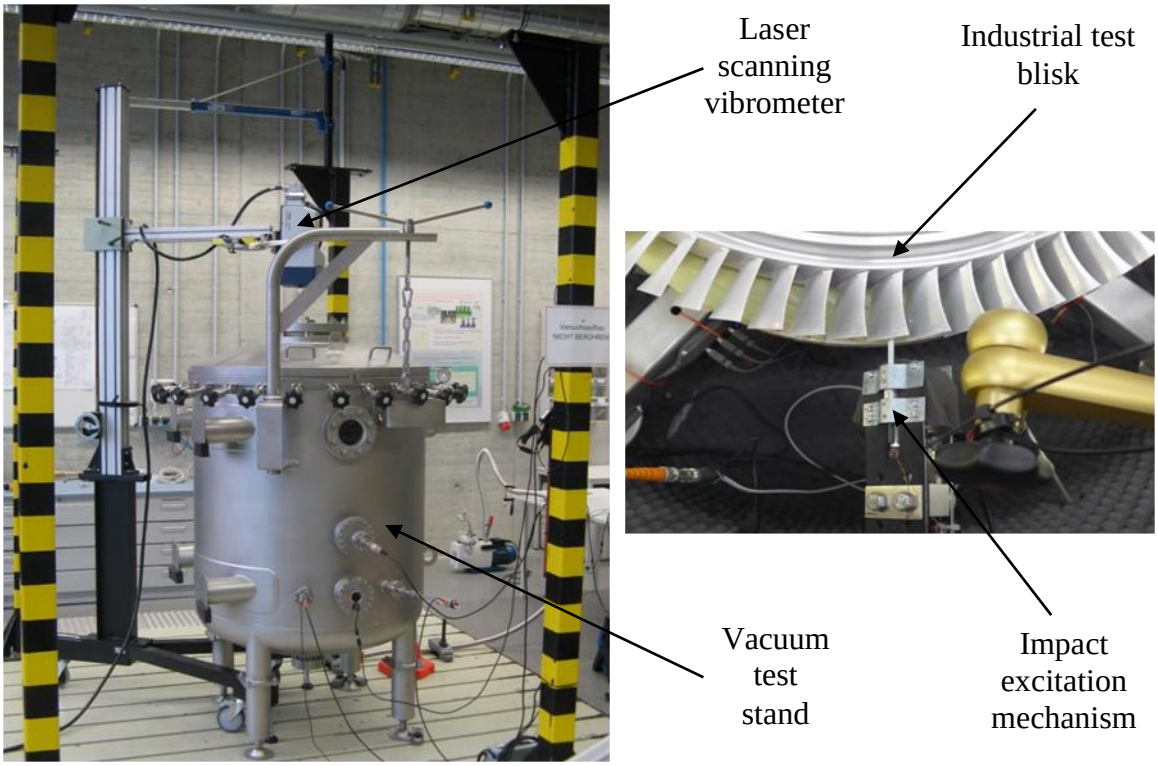


Figure 9: Vacuum test stand and experimental setup in detail

After initializing the test setup the ambient air pressure inside the vacuum champer is reduced to 100 Pa. Vibration measurements are carried out for all blades of the blisk successively at the leading edge tip (Figure 10). The location of impact does not change during the experiments.

To illustrate the influence of low pressure conditions on measured vibration response Figure 11 compares the normalized measured vibration velocity for an exemplary blade at 100 Pa and 100 kPa respectively. Apart from ambient pressure all experimental parameters remain unchanged. As already mentioned vacuum condition largly eliminates damping contribution of the surrounding fluid and results to a more gently decreasing amplitude response in time domain. Measurement data exemplary shown demonstrate that even if the damping contribution of the surrounding air is low it has major influence on system response. In order to emphasize the importance of this low damping condition for modal parameter identification the

time signal shown in Figure 11 is converted into frequency domain using a Fast Fourier Transformation within the next step.

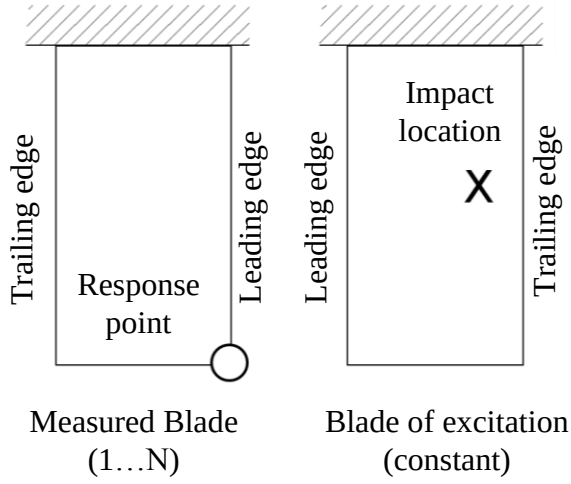


Figure 10: Experimental setup – location of response measurement (left) and impact excitation (right)

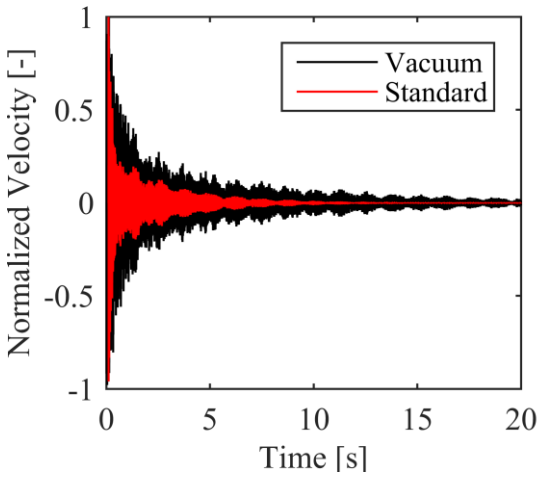


Figure 11: Comparison of normalized impact response at vacuum (100 Pa) and at ambient air pressure (100 kPa) for an exemplary blade

A detailed look at the results of this transformation (Figure 12) clarifies that decreasing ambient pressure affects a better separation of resonance peaks in frequency domain. Furthermore a small frequency shift appears due to modified damping values (cf. Equation 2). The Phase spectrum illustrated in Figure 13 confirms the impression of better identification opportunities at vacuum conditions.

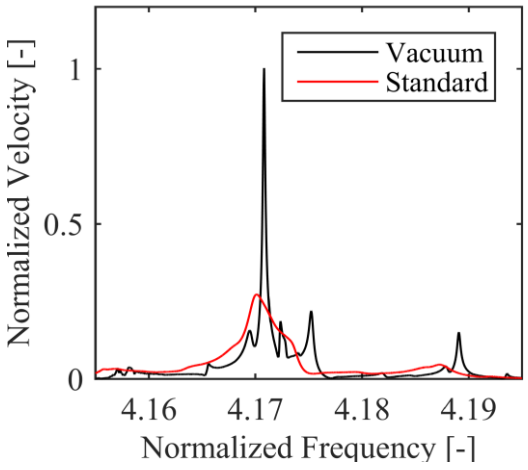


Figure 12: Comparison of normalized velocity spectrum at vacuum (100 Pa) and at ambient air pressure (100 kPa) for an exemplary blade

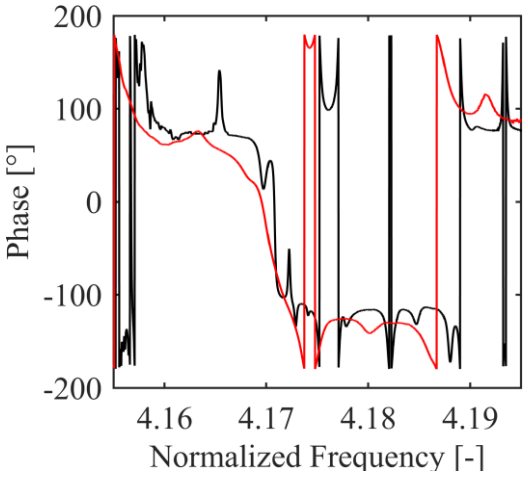


Figure 13: Comparison of phase spectrum at vacuum (100 Pa) and at ambient air pressure (100 kPa) for an exemplary blade

In context of a sufficient model validation it is advantageous to identify as much modal parameters as possible by an experimental modal analysis. The blade integrated disk considered here has extremely strong coupled resonance frequencies. Because of that the presented test facility helps to realize an effective simulation model validation.

4.2 Modal Parameter Identification

Within the next step model parameters have to be extracted from measurement data received by impact measurements described in Chapter 4.1. There are a number of publications dealing with modal parameter identification from vibration experiments. Due to strong coupled resonances as mentioned in the chapter before a multi degree of freedom approach (MDOF) is necessary for interpreting measurement results. The procedure applied here uses an optimization-aided nonlinear least square method in order to fit the frequency response function. A detailed description of this method is given in [20].

Figure 14 shows an exemplary frequency response function fit. Illustrated are the velocity and phase spectra resulting from the experiment (black) and fitted with the MDOF approach (red) in the frequency range of blade mode 1 (cf. Figure 3). Additionally the residuals are defined by

$$\mathrm{Error(f)} = \frac{\left|\mathrm{H}_1^{\mathrm{fit}} - \mathrm{H}_1^{\mathrm{exp}}\right|}{\left|\mathrm{H}_1^{\mathrm{exp}}\right|}. \tag{6}$$

Herein H_1 denotes the complex frequency response function. Amplitude and phase error are plotted separately on the top of Figure 14. Close to resonances the fitting error is in low one digit percentage range. Thus, a reliably natural frequency, damping value and eigenvector identification can be guaranteed.

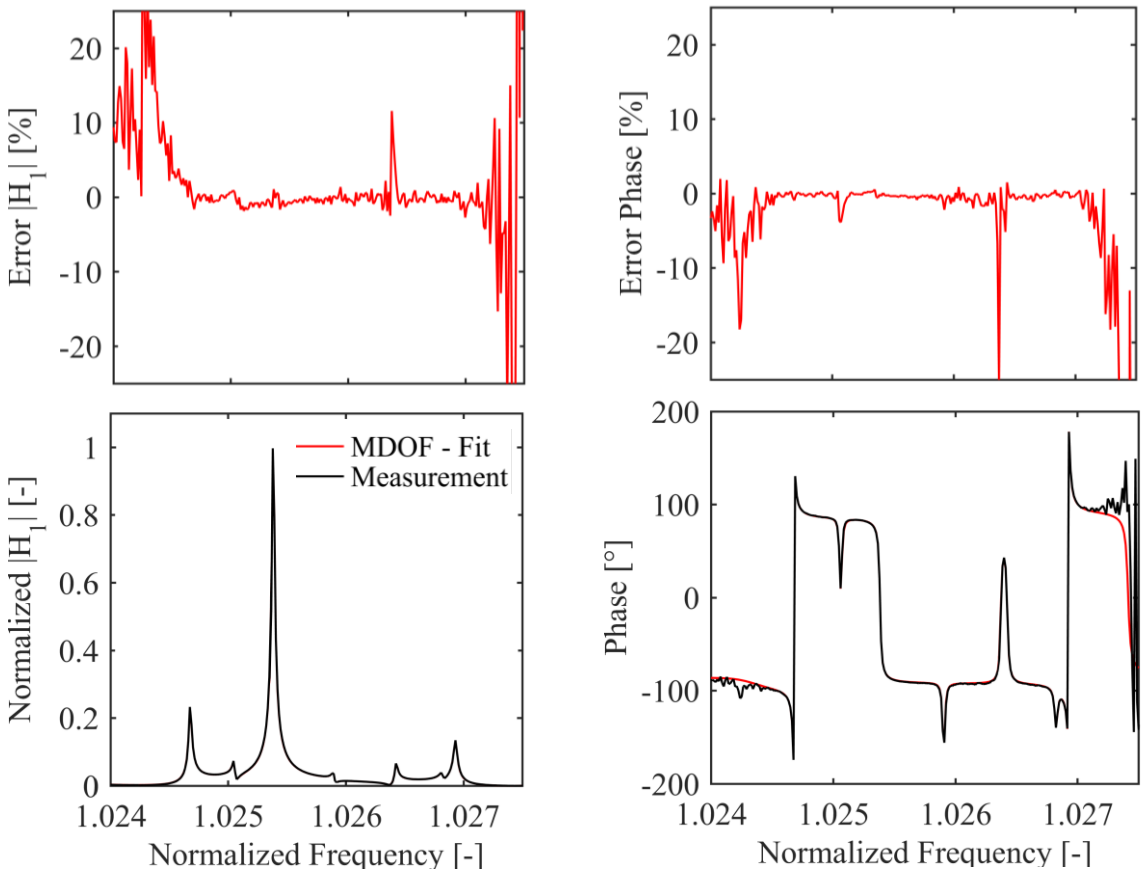


Figure 14: Comparison of measured and fitted forced response function for an exemplary blade in the frequency range of blade mode family 1

In case of the blade integrated disk rotor which is discussed here the introduced fitting algorithm allows to identify about 250 natural frequencies in the range of blade mode 1 to blade mode 3. Moreover 35 eigenvectors are evaluated within this frequency range. More specifically 21 circumferential system modes of blade mode 1, 6 system modes of blade mode 2 and 8 system modes of blade mode 3 are available for model validation. Each eigenvector considers one measurement degree of freedom per blade (cf. Figure 10). The comparably low number of identified eigenvectors can be explained due to the strong modal coupling of neighboring resonances. For these cases a comprehensive evaluation of eigenvectors is much more difficult than natural frequency identification.

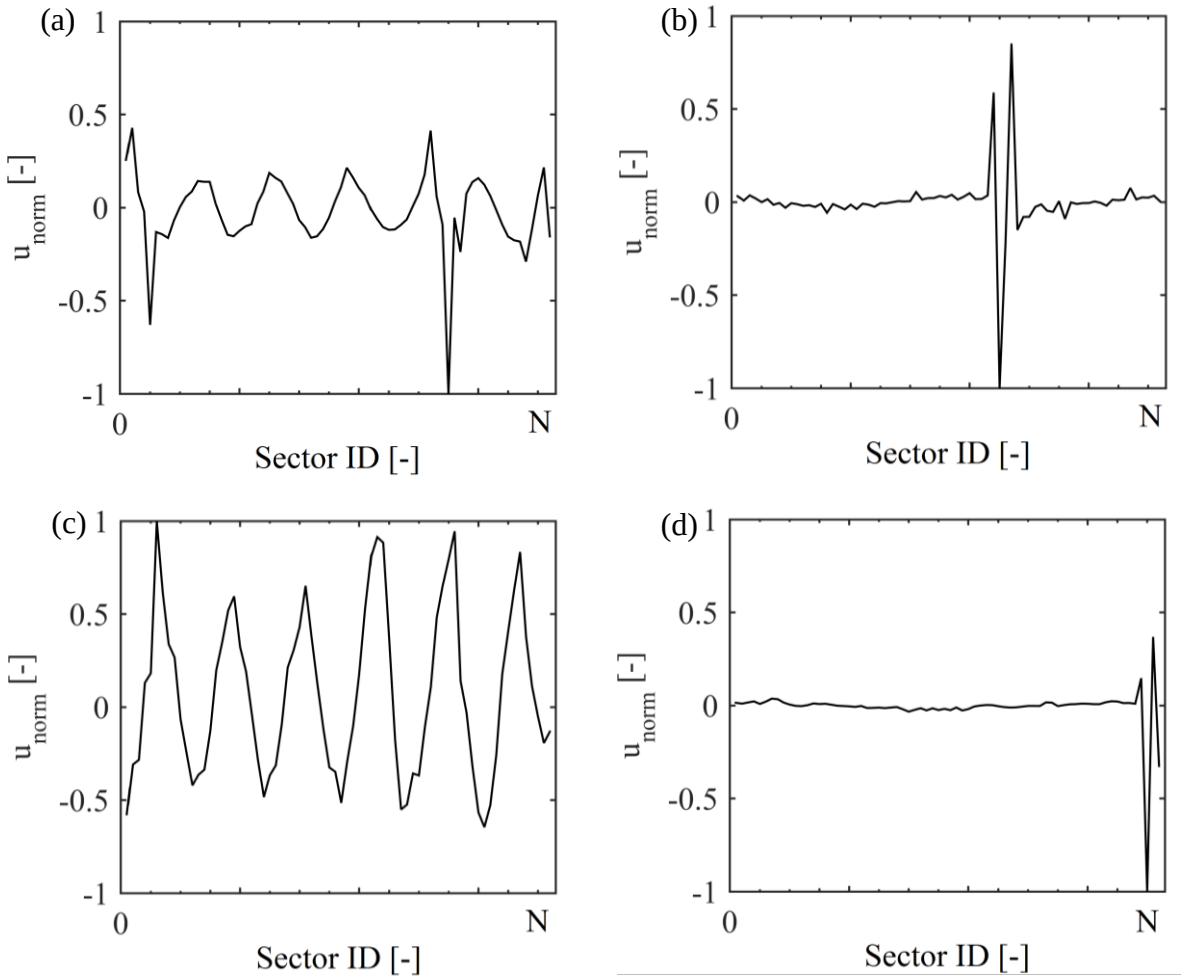


Figure 15: Experimental identified system modes reduced to one degree of freedom (4 modes out of 35)

Figure 15 (a-d) exemplary shows 4 circumferential system modes out of 35 which have been identified within the described measurement campaign. Even this first evaluation of experimental modal analysis impressively demonstrates mistuning effects. Especially modes plotted in Figure 15 (b) and (d) are characterized by a strong localization. That means only a few blades participate in this system mode. In general terms the presented natural frequencies and eigenvectors will be a purposeful data base validating the modified simulation model generated in Chapter 3.

4.3 Validation Results

Key topic of the present paper is to validate geometric modified finite element simulation model. This section uses the modal parameters extracted from vibration measurements as described in Chapter 4.2 and compares them against their numerical counterparts calculated with the finite element model updated in Chapter 3.

First of all the normalized frequencies of the idealized simulation model, the updated simulation model and the experimental identified frequency values are shown in Figure 16(a). The illustration focuses on the frequency range of the first blade mode. Corresponding blade mode shape as well as the circumferential system mode classification has been introduced in Chapter 2. For better understanding a detailed nodal diameter plot is added (Figure 16(b)).

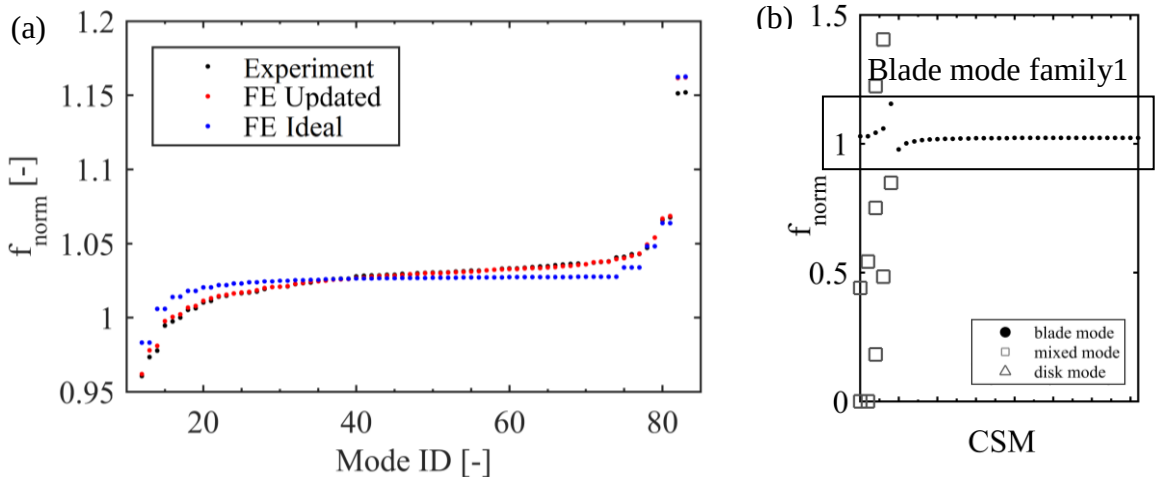


Figure 16: Normalized frequency values in the frequency range of blade mode 1 (a) and the corresponding detailed nodal diameter plot (b)

First of all Figure 16 (a) shows a good match of calculated and measured natural frequencies in case of using the geometry updated finite element model (cf. Chapter 3). Compared with that there are significant deviations in natural frequency prediction comparing the idealized finite element model (cf. Chapter 2) with experimental data. In order to characterize the differences between idealized and updated finite element model towards the experimental identified frequencies a mode individual frequency deviation is introduced

$$df_i = \frac{f_i^{exp/upd} - f_i^{ideal}}{f_i^{ideal}}. \tag{7}$$

Herein f_i^{exp} or f_i^{upd} respectively is the current frequency of interest. Both frequencies are related to the corresponding natural frequency of the idealized finite element model. The resulting deviation plot is presented in Figure 17. As can be seen the first impression of a sufficient model update is confirmed. Deviation characteristics of updated model and experimental data regarding the idealized reference model are nearly the same. In case of blade mode 1 frequency deviations up to 2.5 % are identified comparing the real rotor and the idealized finite element model. Same magnitude is predicted by the updated simulation model. Largest differences between measured and numerically predicted frequency characteristics occur in case of mode IDs corresponding to lower cyclic symmetry indices as for example $i = 82$ or $i = 83$. The comparably large differences are caused by simulation model update input. As described in Chapter 3 optical measurement results are only available for the blade ring section. That means geometric properties of the disk remain unchanged. This obviously influences structur-

al behavior in case of lower cyclic symmetry indices. Further it has to be mentioned that the resonances i = 70 and i = 79 could not be identified by vibration measurements. Because of that it is not possible to calculate the corresponding deviation values df_i.

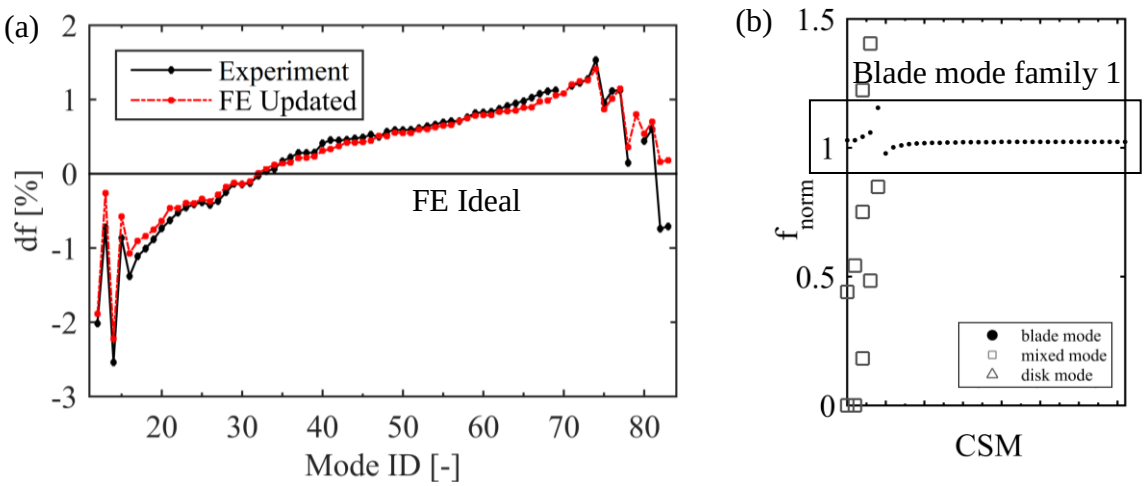


Figure 17: Deviations in natural frequencies between experiment, updated FE and idealized FE model in the frequency range of blade mode family 1 (a) and the corresponding detailed nodal diameter plot (b)

The model validation should be done for a comparable large frequency range. Thus, an identical frequency comparison has been carried out for blade mode 2 and blade mode 3 respectively. In case of blade mode 2 (Figure 18) the impressions of blade mode 1 evaluation are largely confirmed. The frequency deviation characteristics are predicted well by the updated simulation model. Only lower cyclic symmetry indices show comparably large differences.

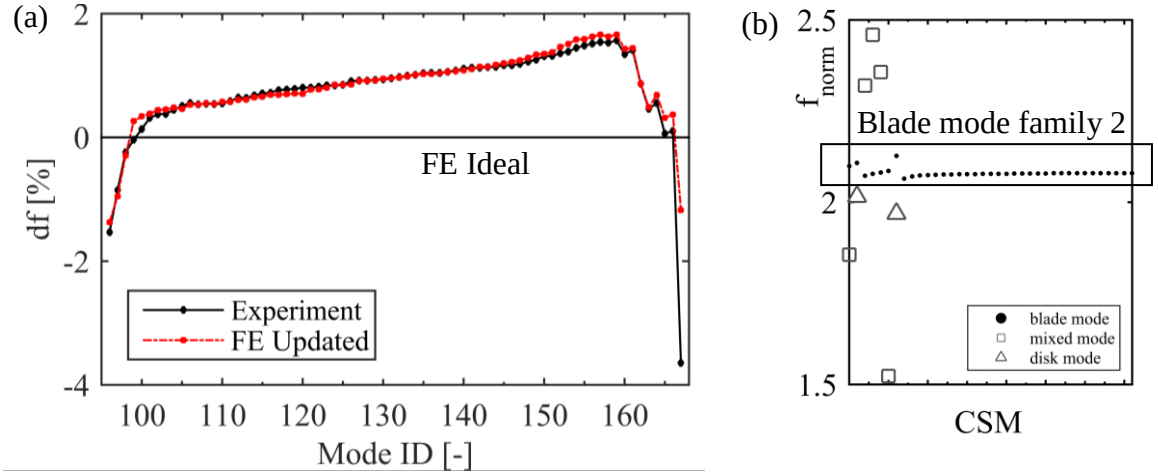


Figure 18: Deviations in natural frequencies between experiment, updated FE and idealized FE model in the frequency range of blade mode family 2 (a) and the corresponding detailed nodal diameter plot (b)

Finally a look at the evaluation results in the frequency range of blade mode 3 (Figure 19) generally confirms previous statements. In detail increasing differences between real blisk and updated simulation model can be observed for resonance IDs higher than i=260. In this context it has to be expected that the occurring differences between measurement results and up-

dated model are increasing by an increasing mode shape complexity. At this point a further discussion is necessary in order to quantify the influence of geometric errors caused by optical measurements. Further it has to be discussed if the finite element mesh is able to represent all relevant geometric features as accurate as necessary. A sufficient mesh study of the ideal model does not necessarily remain valid in case of local geometric deviations.

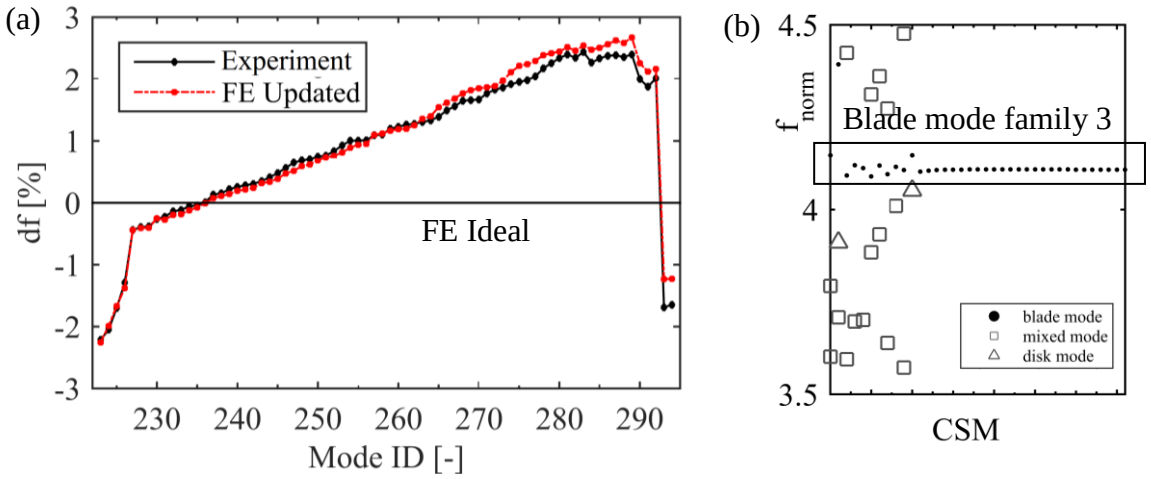


Figure 19: Deviations in natural frequencies between experiment, updated FE and idealized FE model in the frequency range of blade mode family 3 (a) and the corresponding detailed nodal diameter plot (b)

To complete model validation experimentally identified and numerically predicted system modes have to be compared using the Modal Assurance Criterion. As introduced in Chapter 4.2 a system mode is reduced to one degree of freedom per blade. In order to compare numerical and experimental results the numerical based evaluation of the deflection shape is reduced to one node per blade. It has been proven that the position tolerance between simulation model node and real part measurement point has a negligible influence to the resulting circumferential system mode. In the following the conventional MAC is described by

$$\mathrm{MAC}_{\mathrm{exp,fe}} = \frac{(\varphi_{\mathrm{exp}}^{\mathrm{T}}\varphi_{\mathrm{fe}})(\varphi_{\mathrm{fe}}^{\mathrm{T}}\varphi_{\mathrm{exp}})}{(\varphi_{\mathrm{exp}}^{\mathrm{T}}\varphi_{\mathrm{exp}})(\varphi_{\mathrm{fe}}^{\mathrm{T}}\varphi_{\mathrm{fe}})}. \tag{8}$$

In this connection φ is the current eigenvector of interest with one degree of freedom per blade. The main diagonal values of the resulting MAC matrix are shown in the bar chart of Figure 20 (a). It can be seen that the major number of modes is predicted well by the updated simulation model. Only 7 out of 35 modes show diagonal MAC values lower than 0.8. Correlation values for these suboptimal matching circumferential system modes ranging between 0.5 and 0.8. Figure 20 (d-e) exemplary shows three good matching system modes ID = 76, ID = 166 and ID = 224 as well as one suboptimal matching system mode ID = 16.

In order to rate this result the MAC is calculated a second time with the circumferential system modes of the idealized finite element model (cf. Chapter 2). Such a comparison shows nearly no correlation between the tuned finite element model and the measured modes of the real system. Only 5 out of 35 modes have diagonal MAC values greater than 0.5. Thus the efficiency of the model update procedure presented in Chapter 3 is emphasized.

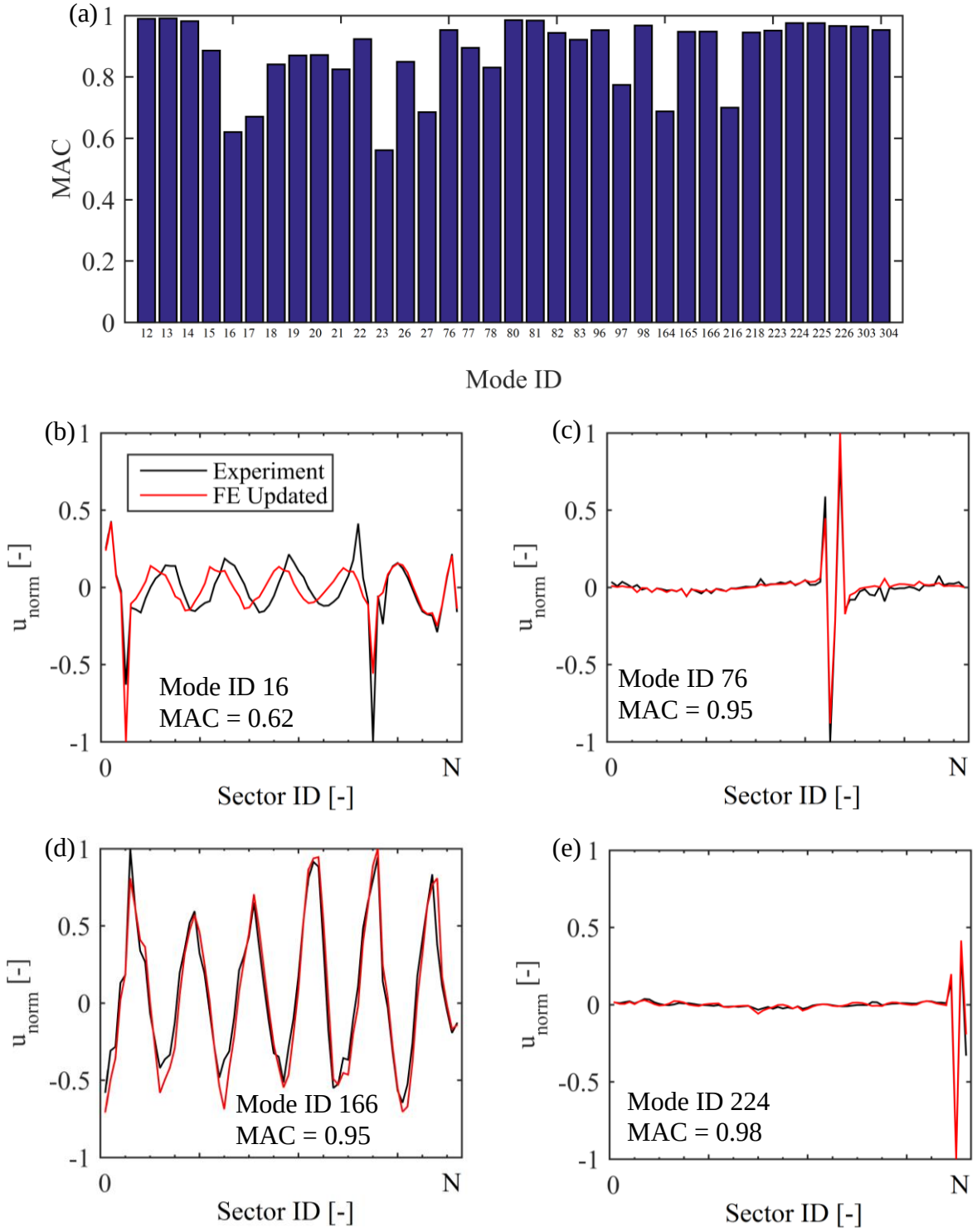


Figure 20: Modal Assurance Criterion for 35 circumferential system modes (a) and detailed comparison of 4 exemplary modes out of 35 (d-e)

Of course the introduced mode shape validation is not that comprehensive as the comparison of natural frequencies. The reasons for that are already discussed in Chapter 4.2. But nevertheless it becomes clear that geometric modifications of blisk simulation models are a sufficient way improving the prediction of structural dynamics behavior.

5 CONCLUSIONS

A finite element model of an idealized blade integrated disk has been introduced and its natural frequencies and mode shapes have been calculated. Within the next step the finite element model was updated using optical measurement data. In order to validate the updated model an experimental setup was presented which was used to identify modal parameters of the real blisk. Vacuum conditions have been proven advantageous in order to get validation data. Largely eliminated aerodynamic damping contribution of the surrounding fluid leads to a much better separation of resonance peaks and the number of evaluable natural frequencies increases dramatically. This finally results in a much more meaningful model validation.

Comparing the experimental identified model parameters with those of the geometric mistuned simulation model leads to the conclusion that geometric modifications of blisk simulation models are a sufficient way improving the prediction of structural dynamics behavior of these components. In this context the presented industrial test blisk has been successfully validated for fundamental blade mode families 1 to 3.

Nevertheless there are several topics which have to be addressed in future research work. From the numerical point of view it is not finally clear which criterion should be used to evaluate the suitability of a finite element mesh in order to consider all geometric features of interest. Furthermore it has to be discussed how to deal with unavoidable measurement errors in context of an optical surface capturing. In addition the vibration measurement setup has to be improved. Here a similarly comprehensive evaluation would be desirable for circumferential system modes as natural frequencies. Alternatively equivalent model validation procedures could be studied for reducing experimental efforts.

6 ACKNOWLEDGMENT

The work presented in this paper is part of the research project ’COOREFLEX-Turbo (Subproject 1.3.4: Robustness Analysis of Blade Vibration)’ which is funded by the German Federal Ministry of Economics and Technology (FKZ 03ET7021). The authors gratefully acknowledge this financial support. Furthermore the authors are grateful to Rolls-Royce Deutschland Ltd & Co KG for granting permission for its publication. Martin Burk, Joerg Weigand, Thomas Stopp and Martin Harding from Rolls-Royce Deutschland Ltd & Co KG in Oberursel are greatly acknowledged for supporting this work by providing the blisk hardware and optical measurement results.

REFERENCES

[1] D.S. Whitehead, Effect of Mistuning on the Vibration of Turbomachine Blades Induced by Wakes, *Journal Mechanical Engineering Science*, Vol. 8, pp. 15-21, 1966.

[2] S.-T. Wei, C. Pierre, Localization Phenomena in Mistuned Assemblies with Cyclic Symmetry Part I: Free Vibrations, *ASME Journal of Vibration, Acoustics, Stress, and Reliability in Design*, 110, pp. 429-438, 1988.

[3] S.-T. Wei, C. Pierre, Localization Phenomena in Mistuned Assemblies with Cyclic Symmetry Part II: Forced Vibrations, *ASME Journal of Vibration, Acoustics, Stress, and Reliability in Design*, 110, pp. 439-449, 1988.

[4] E. P. Petrov, D. J. Ewins, Analysis of the Worst Mistuning Patterns in Bladed Disk Assemblies, *Journal of Turbomachinery*, 125, pp. 623-631, 2003.

[5] J. Jugde, C. Pierre, O. Mehmed, Experimental Investigation of Mode Localization and Forced Response Amplitude Magnification for a Mistuned Bladed Disk, *Journal of Engineering for Gas Turbines and Power*, 123, pp. 940-950, 2001.

[6] M. P.Castanier, C Pierre, Modelling and Analysis of Mistuned Bladed Disk Vibration: Status and Emerging Directions, *Journal of Propulsion and Power*, 22, No. 2, pp. 384-396, 2006.

[7] D. Laxalde, F. Thouverez, J.-J Sinou, J.-P. Lombard, S. Baumhauer, Mistuning Identification and Model Updating of an Industrial Blisk (review Article), *International Journal of Rotating Machinery*, Volume 2007, Article ID 17289, 2007.

[8] D. M. Feiner, J. H. Griffin, Mistuning Identification of Bladed Disks Using a Fundamental Mistuning Model—Part I: Theory, *Journal of Turbomachinery*, 126, pp. 150-158, 2004.

[9] J. A. Judge, C. Pierre, S. L. Ceccio, Experimental mistuning identification in bladed disks using a component-mode-based reduced-order model. *AIAA journal*, *47*(5), 1277-1287, 2009.

[10] H. Schoenenborn, D. Grossmann, W. Satzger & H. Zisik, Determination of blade-alone frequencies of a blisk for mistuning analysis based on optical measurements. In *ASME Turbo Expo 2009: Power for Land, Sea, and Air* (pp. 221-229). American Society of Mechanical Engineers, 2009.

[11] A. A. Kaszynski, J. A. Beck, & J. M Brown. Automated Finite Element Model Mesh Updating Scheme Applicable to Mistuning Analysis. In *ASME Turbo Expo 2014: Turbine Technical Conference and Exposition* (pp. V07BT33A025-V07BT33A025). American Society of Mechanical Engineers, 2014.

[12] V. Ganine, M. Legrand, H. Michalska, & C. Pierre. A sparse preconditioned iterative method for vibration analysis of geometrically mistuned bladed disks. *Computers & Structures*, *87*(5), 342-354, 2009.

[13] D. L. Harris. Measurement of damping of composite materials for turbomachinery applications. *MSFC Center Director's Fund Final Report, Project*, (94-05) ,1998.

[14] J. J. Kielb, & R. S. Abhari. Experimental study of aerodynamic and structural damping in a full-scale rotating turbine. In *ASME Turbo Expo 2001: Power for Land, Sea, and Air* (pp. V004T03A028-V004T03A028). American Society of Mechanical Engineers, 2001, June.

[15] P. Hönisch, A. Kühhorn, & B. Beirow. Experimental and numerical analyses of radial turbine blisks with regard to mistuning. In *ASME 2011 Turbo Expo: Turbine Technical Conference and Exposition* (pp. 971-980). American Society of Mechanical Engineers, 2011, January.

[16] W. Heylen, & T. Janter. Extensions of the modal assurance criterion. *Journal of Vibration and Acoustics*, *112*(4), 468-472, 1990.

[17] K. Harding. Latest optical methods for industrial dimensional metrology. *In Optics East 2005* (pp. 600001-600001). International Society for Optics and Photonics, 2005.

[18] A. A., Kaszynski, J. A. Beck, J. M. Brown. Uncertainties of an automated optical 3d geometry measurement, modeling, and analysis process for mistuned integrally bladed

rotor reverse engineering. *Journal of Engineering for Gas Turbines and Power*, *135*(10), 102504, 2013.

[19] F. Remondino, & S. El-Hakim. Image-based 3D Modelling: A Review. *The Photogrammetric Record*, *21*(115), 269-291, 2006.

[20] B. Beirow, T. Maywald, F. Figaschewsky, C. R. Heinrich, A. Kühhorn, & T. Giersch. Simplified determination of aerodynamic damping for bladed rotors. Part 1: Experimental validation at rest. In *ASME 2016 Turbo Expo: Turbine Technical Conference and Exposition*. American Society of Mechanical Engineers, 2016, June.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

Optimum Design, Finite Element Model Updating and Dynamic Analysis of a Full Laminated Glass Panoramic Car Elevator

Dimitrios Giagopoulos[1], Iraklis Chatziparasidis[2]

[1] Dept. of Mechanical Engineering, University of Western Macedonia, Kozani, Greece

[2] Research & Development Department, Kleemann Hellas Complete Lifts SA, Kilkis, Greece

e-mail: dgiagopoulos@uowm.gr, ihatz@kleemann.gr

Keywords: Full Glass Car, Large scale structures, Model updating, Stress evaluation, Structural Dynamics.

Abstract. *A systematic optimum design procedure, including an accurate dynamic analysis of a full glass panoramic car elevator under real dynamic load conditions are presented in this work. The cabin is manufactured entirely of laminated glass (two glass layers and an interlayer of polivinyl butiral -PVB), except the roof and the platform. First, modal identification and structural model updating methods are applied, leading to develop high fidelity finite element model of the glass and its connection subsystems. The identification of modal characteristics of the glass is based on acceleration and stress time histories, which are obtained through an experimental investigation of its dynamic response, in two different states. First, in a support-free state by imposing impulsive loading and second in a fixed-free state by imposing random excitation with the use of an electrodynamic shaker. Single and multi-objective structural identification methods are used for estimating the parameters (material properties) of the FE model, based on minimizing the deviations between the experimental and analytical modal characteristics. Next, a "mixed computational-experimental" analysis method is applied, in order to simulate accurately the dynamic behavior of the complete elevator system, in emergency situations like safety gear engagement. A series of experimental tests were performed under real operating conditions, using an experimental device that was designed exactly for this purpose and aimed at recording the acceleration time histories at the connection points of the frame with the safety gears. These acceleration time histories are subsequently used as base excitation for the FE model of the complete elevator system and the stresses developed under these specific loading conditions are evaluated. On the basis of these numerical results, the critical points of the frame are selected, as corresponding to larger stresses and an optimum design procedure was applied. Finally, in order to test the reliability of the method applied, strain gauges are placed at the critical points of the optimum designed system and a series of measurements are carried out, in order to experimentally verify the developed stresses. Direct comparison of the numerical and experimental data verified the reliability and accuracy of the methodology applied.*

1 INTRODUCTION

Within an elevator system the car is one of the most significant components because it is used to carry goods and most importantly persons. Because of its use the car is subject of strict legislations and standards that differ to each Country or State. In the same time the car is, usually, the only functional component of the elevator that is visible by the final user. This fact alters the car to be also an architectural and decorative element of the building. The latest trends in Architecture is to use fully transparent glass building elements [17]. A frameless glass panoramic elevator could not only fit to this trend but also extend the aesthetic of the building. However, glass is a brittle material and this attribute made the verification of the materials load carrying capacity to be mostly based on experimental studies. Some research has been done on the field of modeling of the behavior of glass components, using finite element methodologies, but these are mainly studying static loads situations aimed at the building construction domain [16-18]. In this work we present a method for modeling glass components in dynamic situations that may occur during an elevator's function e.g. when the safety gear of the elevator is activated during an emergency stop. An efficient modelling method is required for the various elevator parts which are in contact with each other. To achieve these modelling issues, it is important to develop an accurate Finite Element Analysis procedure, in order to simulate the dynamic behaviour of these systems.

Another main issue addressed by the present work is the need for the development and application of new appropriate methodologies for investigating the dynamics of large scale mechanical models in a systematic and efficient way. The equations of motion of mechanical systems with complex geometry are first set up, applying classical finite element techniques. As the order of these models increases, the existing numerical and experimental methodologies for a systematic determination of their dynamic response become inefficient to apply. Moreover, in order to optimize the FE model of a structure, structural model updating methods [7], have been proposed in order to reconcile the numerical (FE) model, with experimental data. Structural model parameter estimation based on measured modal data [1-6] are often formulated as weighted least-squares estimation problems in which metrics, measuring the residuals between measured and model predicted modal characteristics, are build up into a single weighted residuals metric formed as a weighted average of the multiple individual metrics using weighting factors. Standard gradient-based optimization techniques are then used to find the optimal values of the structural parameters that minimize the single weighted residuals metric representing an overall measure of fit between measured and model predicted modal characteristics. Due to model error and measurement noise, the results of the optimization are affected by the values assumed for the weighting factors.

The work presented here is based on previous work where we demonstrated the advantages of applying appropriate numerical and experimental methodologies in order to accurately predict the dynamic response and the identification of the critical points in an elevator system [2]. Also examined the applicability and effectiveness of the updating methods, coupled with robust, accurate and efficient finite element analysis software.

2 FINITE ELEMENT MODEL UPDATING METHODS

Let $D=\{\hat{\omega}_r, \underline{\hat{\phi}}_r \in R^{N_o}, r=1,\cdots,m\}$ be the measured modal data from a structure, consisting of modal frequencies $\hat{\omega}_r$ and mode shape components $\underline{\hat{\phi}}_r$ at N_o measured DOFs, where m is the number of observed modes. Consider a parameterized class of linear structural models used to model the dynamic behavior of the structure and let $\underline{\theta} \in R^{N\theta}$ be the set of free structural model parameters to be identified using the measured modal data. The objective in a modal-based

structural identification methodology is to estimate the values of the parameter set $\underline{\theta}$ so that the modal data $\{\omega_r(\underline{\theta}),\ \underline{\phi}_r(\underline{\theta}) \in R^{N_0}, r=1,\ldots,m\}$ predicted by the linear class of models at the corresponding N_0 measured DOFs best matches the experimentally obtained modal data in D. For this, let

$$\varepsilon_{\omega_r}(\underline{\theta}) = \frac{\omega_r^2(\underline{\theta}) - \hat{\omega}_r^2}{\hat{\omega}_r^2} \quad \text{and} \quad \varepsilon_{\underline{\phi}}(\underline{\theta}) = \frac{\left\| \beta_r(\underline{\theta})\underline{\phi}_r(\underline{\theta}) - \underline{\hat{\phi}}_r \right\|}{\left\| \underline{\hat{\phi}}_r \right\|} \tag{1}$$

be the measures of fit or residuals between the measured modal data and the model predicted modal data for the r-th modal frequency and mode shape components, respectively, where $\|\underline{z}\|^2 = \underline{z}^{\mathrm{T}}\underline{z}$ is the usual Euclidean norm, and $\beta_r(\underline{\theta}) = \underline{\hat{\phi}}_r^T \underline{\phi}_r(\underline{\theta}) / \left\|\underline{\phi}_r(\underline{\theta})\right\|^2$ is a normalization constant that guaranties that the measured mode shape $\underline{\hat{\phi}}_r$ at the measured DOFs is closest to the model mode shape $\beta_r(\underline{\theta})\underline{\phi}_r(\underline{\theta})$ predicted by the particular value of $\underline{\theta}$.

To proceed with the model updating formulation, the measured modal properties are grouped into two groups. The first group contains the modal frequencies while the second group includes the mode shape components for all modes. For each group, a norm is introduced to measure the residuals of the difference between the measured values of the modal properties involved in the group and the corresponding modal values predicted from the model class for a particular value of the parameter set $\underline{\theta}$. For the first group, the measure of fit $J_1(\underline{\theta})$ is selected to represent the difference between the measured and the model predicted frequencies for all modes. For the second group, the measure of fit $J_2(\underline{\theta})$ is selected to represent the difference between the measured and the model predicted mode shape components for all modes. Specifically, the two measures of fit are given by

$$J_1(\underline{\theta}) = \sum_{r=1}^{m} \varepsilon_{\omega_r}^2(\underline{\theta}) \quad \text{and} \quad J_2(\underline{\theta}) = \sum_{r=1}^{m} \varepsilon_{\underline{\phi}}^2(\underline{\theta}) \tag{2}$$

The parameter estimation problem is traditionally solved by minimizing the single objective

$$J(\underline{\theta};\underline{w}) = w_1 J_1(\underline{\theta}) + w_2 J_2(\underline{\theta}) \tag{3}$$

formed by the two objectives $J_i(\underline{\theta})$, using the weighting factors $w_i \geq 0$, $i=1,2$, with $w_1 + w_2 = 1$. The objective function $J(\underline{\theta};\underline{w})$ represents an overall measure of fit between the measured and the model predicted characteristics. The relative importance of the residual errors in the selection of the optimal model is reflected in the choice of the weights. The results of the identification depend on the weight values used. The optimal solutions for the parameter set $\underline{\theta}$ for given $\underline{w}$ are denoted by $\underline{\hat{\theta}}(\underline{w})$ [8-10].

3 EXPERIMENTAL APPLICATION

In this section, the emphasis is placed on applying the methodology proposed to an elevator with a frameless glass car. The ultimate goal was to develop an accurately Finite Element Analysis procedure, in order to simulate the dynamic behaviour of this system. The FE model of the structure examined was created using mainly rectangular and triangular shell finite elements. Additionally, some other elements like solid (hexahedral) and rigid body elements were also used. After development of the overall FE model, the next step was to examine the dynamic response of system when safety gear is activated and the elevator stops. In order to accurately simulate this procedure, the first step is the identification of the braking forces that acting on the system. The next step, is the verification and validation of the

developed FE model. More specifically, the emphasis given in the developing of an accurate FE model of the glass panels and its supports.

3.1 Determination of Acceleration Levels Under Real Dynamic Load Conditions

In order to verify the accuracy of the FE model of the system and also identify the braking forces acting on the system, we select, to examine only the elevator chassis including the platform of the cabin with full load. On this system, triaxial accelerometers are placed at (6) selected positions. These positions along with the measurement directions are presented in Figure 1. The two connection points of the frame with the safety gear (A1, A2) are included along with four other locations (A3-A6) which are used as reference points. A series of experimental trials were performed under real operating conditions (free fall of the elevator frame), using an experimental device designed and constructed exactly for this purpose (by Kleemann Hellas S.A), aimed at recording the acceleration time histories at the selected six points. Also, in Figure 2 presented indicative photos of two measurement locations.

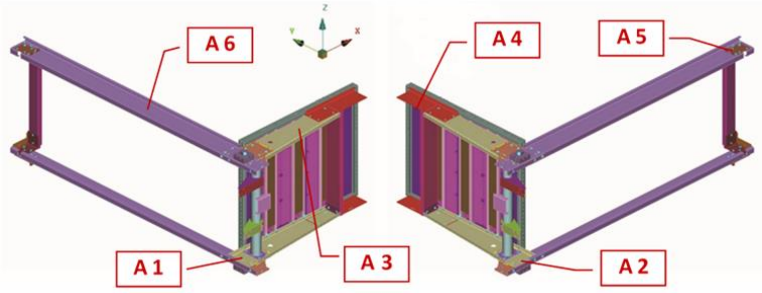


Figure 1: Measurement Locations of Acceleration Time Histories.

Figure 2: Acceleration Measurement Locations at a Connection Point (A1) and at a Reference Point (A4).

Next, the measured acceleration time histories are used in order to determine the braking force that acting on the system. After several experimental trials, the form of the corresponding time-varying braking forces, in each progressive safety gear, were calculated. Then, these force time histories were imported as base excitation in the finite element model of the system. Also, in order to solve this transient response analysis problem in a computationally effective way, a reduction in the dimensions of the original system is performed, so that the results are accurate in the frequency range 0-500 Hz. The total number of degrees of freedom in the reduced model was about 3,500, which is much smaller than the number of degrees of freedom in the original model (1,200,000). The reduced model was solved numerically in order to calculate the maximum stresses developed for the given loading. The identified critical points of the structure include mainly areas of the chassis arms on which the platform is based. Figure 3 shows selected results, in which some indicative

points of the superstructure with maximum stresses are presented.

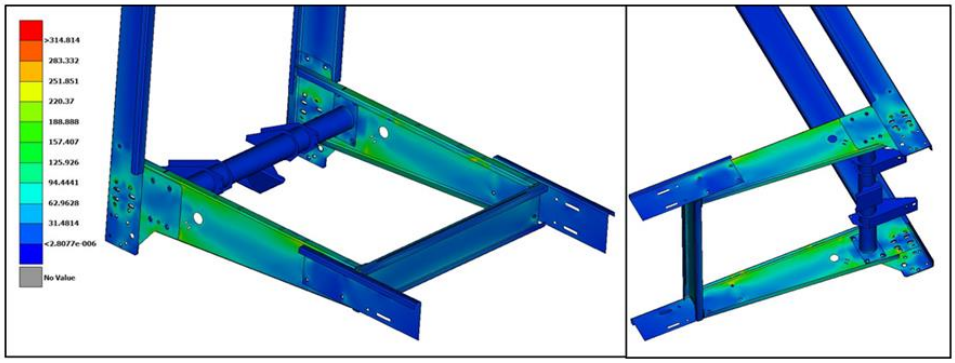


Figure 3: - Locations of the Chassis where Maximum Stresses Appear

3.2 Validation of the Applied Methodology

In order to test the reliability of the method applied, strain gauges were placed at (5) selected critical points of the chassis and a new set of measurements was carried out under similar dynamic loading conditions, to experimentally verify the stress levels developed. These positions are presented in Figure 4 and include locations on the right side of the chassis arms (SG1, SG2 and SG3) and on the left side (SG4 and SG5).

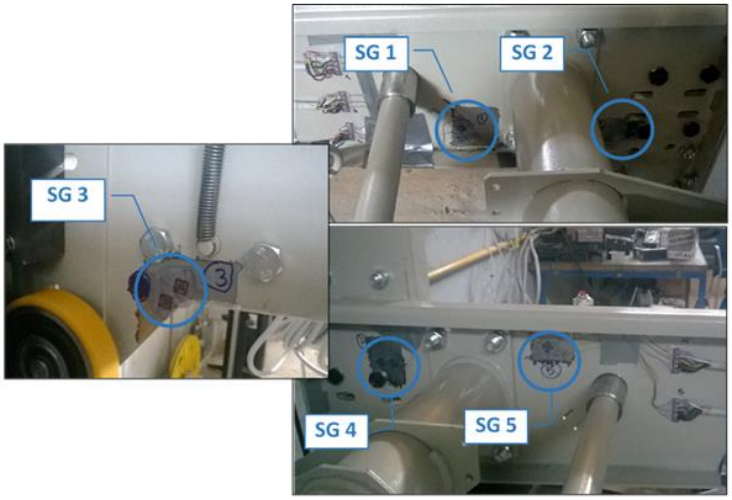


Figure 4: Measurement Locations where Maximum Stress Appears

For a complete monitoring of the stress state, three bridges with a 120o angle rosette were placed at each of the selected points. Some of the experimental and numerical results are summarized in Table 1. This table presents the maximum values of the von Mises stress obtained in four tests (indicated by PG1-PG4) and for all the points where measurements were taken (denoted by SG1-SG5). More specifically, the third from the end column presents the maximum values for all tests, for each measurement location, the penultimate column presents the corresponding maximum values obtained by the finite element model and the last column presents their percentage difference. A comparison of the numerical and experimental data presented in Table 1 verifying that the proposed method is reliable.

	Maximum Value (peak) of Equivalent Stress Von Mises [MPa]				Maximum Value of all Structural Tests	Maximum Value of FEM Solution	Error (%)
Structural Tests	PG1	PG2	PG3	PG4			
Strain Gage Positions							
SG1	90.47	85.48	97.22	92.56	**97.22**	105.37	**7.73**
SG2	198.64	193.72	201.14	196.35	**201.14**	209.82	**4.14**
SG3	168.61	165.40	170.85	173.21	**173.21**	170.53	**1.57**
SG4	160.49	148.25	154.71	152.16	**160.49**	155.61	**3.14**
SG5	253.75	246.53	263.22	255.91	**263.22**	278.24	**5.40**

Table 1: Maximum Value of Equivalent Von Mises Stress [MPa]

3.3 FE Model Verification of a Glass Panel with the Support System

Next the emphasis given to develop a high fidelity FE model of the glass panels. To achieve this, it is necessary to optimize the numerical FE model of a glass panel including the bottled support. Basic structural model updating methods have been proposed [7], in order to reconcile the numerical (FE) model, with experimental data. Structural model parameter estimation based on measured modal data [1-6] are often formulated as weighted least-squares estimation problems in which metrics, measuring the residuals between measured and model predicted modal characteristics.

3.3.1. FE Model

First, the geometry of the glass panel with the test support device is discretized mainly by solid tetrahedral elements. For the development and solution of the finite element model some appropriate software was used [20, 21]. The detailed FE Model of the experimental device presented in Figure 6. Two typical eignmodes predicted by the nominal finite element model, presented in Figure 5.

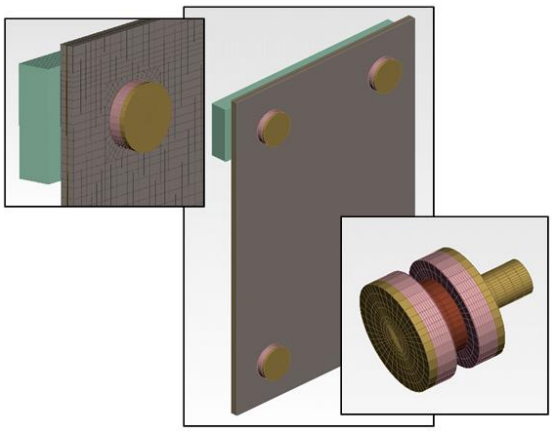


Figure 5: Finite Element Model of the Glass Panel with Support

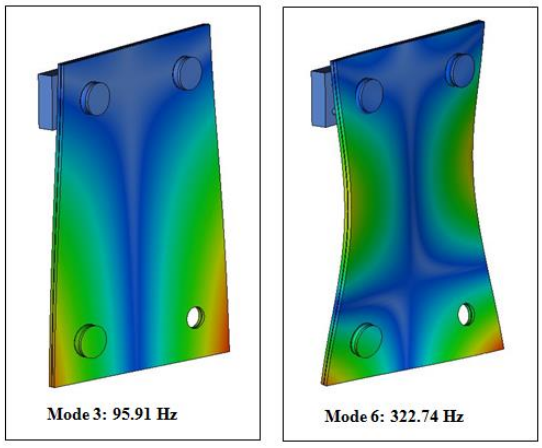


Figure 6: Typical Eigenmodes Predicted by the Nominal Finite Element Model

The FE models of the main frame and the fork parts are updated using the identified modal frequencies and mode shapes shown in Tables 1 and 2. The identified mode shapes include components at all sensor locations. Additionally, we define as design response the total weight of the model, in order to be taken into consideration during the optimization process.

3.3.2. Experimental Modal Analysis

After development of the nominal FE model, an experimental modal analysis of the devise was performed to quantify its dynamic characteristics. The system was tested in fixed-free boundary condition. First, all the necessary elements of the FRF matrix required for determining the response of the frame substructure were determined by imposing impulsive loading [1, 2, 8, 9]. The measured frequency range was 0-2048 Hz, which includes the analytical frequency range of interest, 0-400 Hz. An initial investigation indicated that the frame has six natural frequencies in this frequency range. A schematic illustration of the measurement geometry of the experimental device is presented in Figure 7. In this figure, presented the locations of the tri-axial accelerometers, strain gauges and of the electrodynamic shaker.

For instance, Figure 8 shows the magnitude of two typical elements of the FRF matrix before (continuous line) and after (dashed line) application of the Welsh's smoothing method.

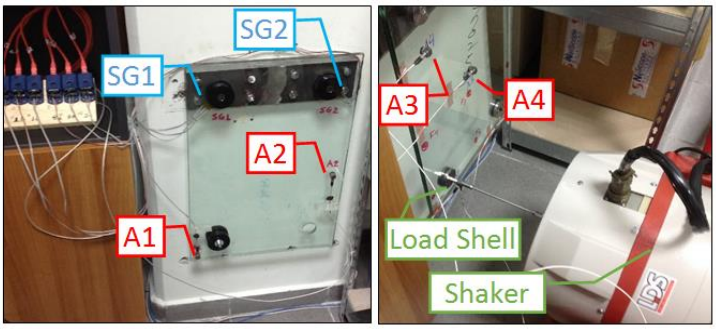


Figure 7: Schematic Illustration of the Experimental Device, Fixed-Free Arrangement with Electrodynamic Shaker, Accelerometers and Strain Gauges Locations

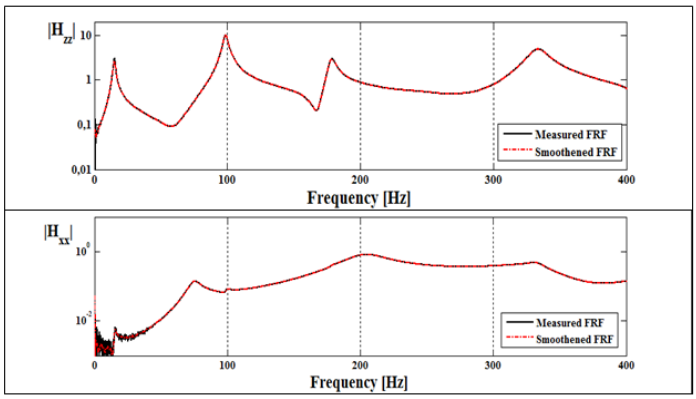


Figure 8: Typical Elements of the FRF Matrix

Based on the measured FR functions, the natural frequencies and the damping ratios of the frame substructure were estimated. As an outcome of the above procedure, the first column of Table 2 presents the values of the lowest 6 natural frequencies (ω_{rE}) of the system examined, while the corresponding damping ratios are included in the fourth column. In the same table,

the second column presents the values of the natural frequencies obtained from the analysis of the nominal finite element model ($\omega_{rN_{FE}}$) and the third column compares these frequencies with the corresponding frequencies obtained by the experimental data. The errors determined between the nominal FE model and the experimental measurements are not insignificant, indicating that the FE model updating process is necessary.

Mode	Identified Modal Frequency	Nominal FE Predicted Modal Frequency	Difference between Identified and FE Predicted Modal Frequencies	Identified Modal Damping Ratio
	$\omega_{rE}[Hz]$	$\omega_{rN_{FE}}[Hz]$	$\frac{\omega_{rN_{FE}}-\omega_{rE}}{\omega_{rN_{FE}}}100\%$	$\zeta_{rE}(\%)$
1	15.15	17.06	11.20	4.78
2	73.69	87.23	15.52	7.14
3	98.70	95.91	2.91	1.74
4	177.80	170.94	4.01	1.56
5	203.01	227.32	10.69	7.90
6	333.10	322.74	3.21	2.29

Table 1: Modal Frequencies and Modal Damping Ratios

3.3.3. FE model parameterization and Updating Results

The parameterization of the FE model of the experimental device are introduced in order to demonstrate the applicability of the proposed FE model updating method. The parameterized model consisting of six parts which are shown in Figure 9. At each of these parts are used as design variables the Young's modulus and the density. Thus, the final number of the design parameters are twelve (12) variables. In Table 2 presented the initial values that have been set in each parameter, which are identical to the nominal FE model, with the upper and lower limits, which were selected to be used for the optimization process. The last column of the table shows the step of design, which is set at 1% of the respective previous value for all cases. The finite element model is updated using the lowest six identified modal frequencies and mode shapes shown in Table 1. The identified mode shapes include components at all 4 sensor locations.

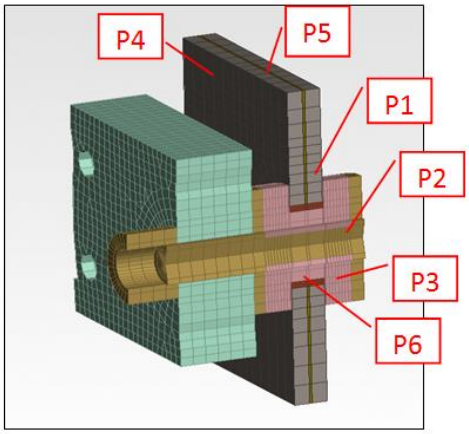


Figure 9: Parts of the Parameterized FE Model

Part	Initial Density [kg/m^3]	Initial Young's Modulus [Gpa]	Move Limit
	LB - UB	LB - UB	
P1	1840 **1000 – 2500**	0.6 **0.3 – 0.9**	1%
P2	7850 **7200 – 8500**	210.0 **190 – 220**	1%
P3	1100 **800 – 2000**	0.018 **0.006 – 0.05**	1%
P4	2500 **2000 – 3000**	63.0 **60 – 70**	1%
P5	1840 **1000 – 2500**	0.6 **0.3 – 0.9**	1%
P6	2500 **1500 – 3000**	0.6 **0.3 – 0.9**	1%

Table 2: Design Variables and Optimization Design Limits

The results from the FE model updating method are shown in Table 3. In this table presented a comparison between identified (ω_{rE}) and $(\omega_{rO_{FE}})$ optimal FE predicted modal frequencies.

Mode	Identified Modal Frequency	Optimal FE Predicted Modal Frequency	Difference between Identified and FE Predicted Modal Frequencies
	$\omega_{rE}[Hz]$	$\omega_{rO_{FE}}[Hz]$	$\frac{\omega_{rO_{FE}}-\omega_{rE}}{\omega_{rO_{FE}}}100\%$
1	15.15	15.29	0.92
2	73.69	76.25	3.36
3	98.70	99.30	0.60
4	177.80	177.00	0.45
5	203.01	202.50	0.25
6	333.10	319.10	4.39

Table 3: Comparison Between Identified and Optimal FE Predicted Modal Frequencies

The acceleration time history and the FRF predicted by the optimal FE model (red dashed dot line) of the glass panel are compared in Figure 10 with the acceleration time history and the FRF computed directly from the measured data (blue continuous line) at one indicative measurement locations of the glass panel (A1) in the frequency range [0Hz, 400Hz]. The acceleration time history and the FRF of the initial nominal model (black dashed line) is also shown in these figures to be inadequate to represent the measured acceleration time history and the FRF. Compared to the FRF of the initial nominal model, it is observed that the updated optimal model tend to considerably improve the fit between the model predicted and the experimentally obtained FRF close to the resonance peaks.

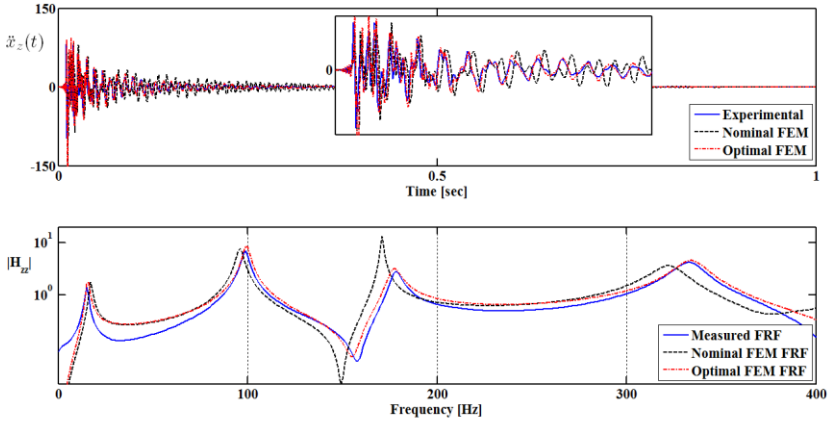


Figure 10: Comparison between measured, nominal and optimal acceleration time histories and FRF in the measured location A1

3.4 Analysis of the FE Model of Full Elevator System - Frameless Full Glass Cabin

Finally, a detailed finite element model of a full elevator system using a frameless full glass car was build. The model was solved numerically in transient response analysis, using the method which was presented in previous sections 3.1-3.2, in order to calculate the maximum stresses developed [19]. Based on the results of this analysis the elevator chassis

and the bottled glass supports was redesigned and optimized, in order to achieving the minimum design stresses at the glazing components during dynamic load conditions caused by e.g. emergency safety gear engagement. The final FE model with indicative numerical results presented in Figure 11.

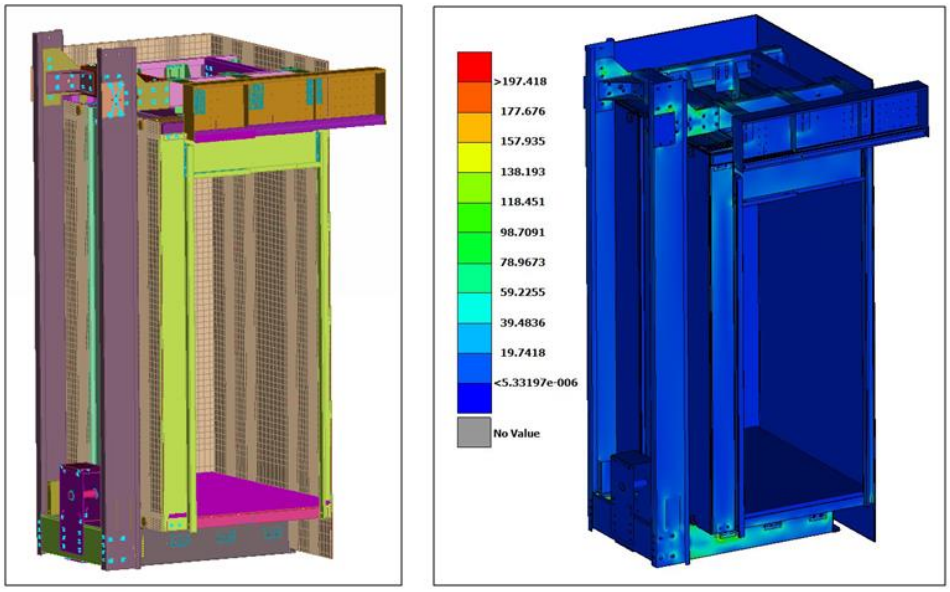


Figure 11: Frameless Full Glass Car FE model and Transient Analysis Results

Based on the results of all the above procedure, a full glass frameless panoramic elevator (Figure 12) was designed and developed. In order to test the reliability of the method applied, strain gauges were placed at (4) selected critical points of the glass panels and a new set of measurements was carried out under similar dynamic loading conditions, to experimentally verify the stress levels developed. A comparison of the numerical and experimental data verifying that the proposed method is quite reliable.

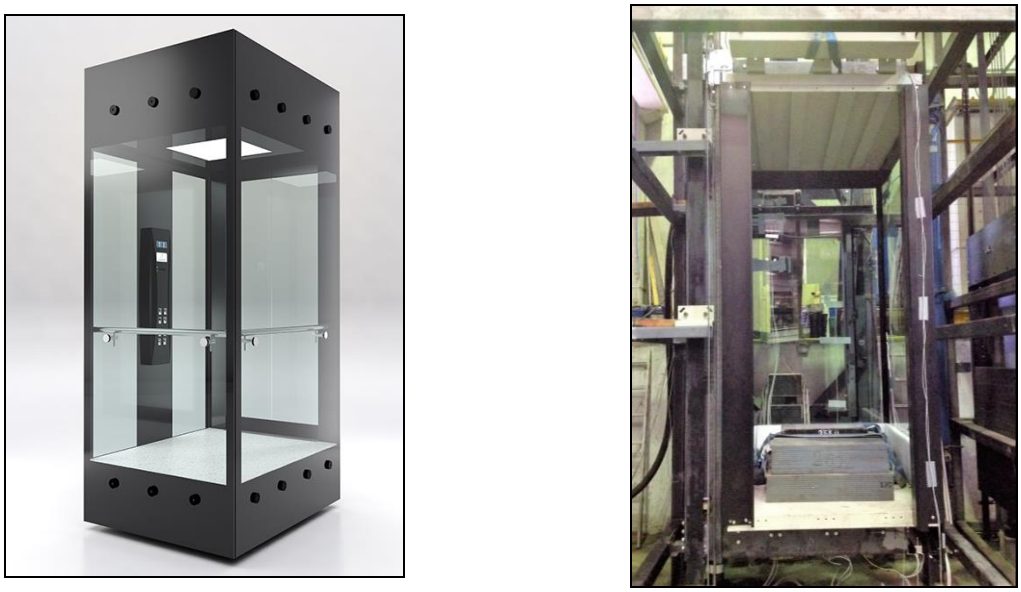


Figure 12: Full Glass Frameless Car with Experimental Verification Test (Patent Pending)

4 SUMMARY

A systematic method was presented for determining the dynamic response and identifying the critical points of a full glass panoramic car elevator system when subjected to dynamic load conditions. The special feature in this work is that the cabin is manufactured entirely of laminated glass, except the roof and the platform. First, modal identification and structural model updating methods are applied, leading to develop high fidelity finite element model of the glass and its connection subsystems. Next, a "mixed computational-experimental" analysis method is applied, in order to simulate accurately the dynamic behavior of the complete elevator system, in emergency situations like safety gear engagement. Based on the results of the method applied the elevator chassis and the bottled glass supports was redesigned, optimized. Finally, a full glass frameless panoramic elevator was designed and developed. Comparison of numerical and experimental results indicated that the methodology applied gives accurate results and provides a useful tool in predicting the critical stress levels developed in the elevator.

REFERENCES

[1] Giagopoulos, D. and Natsiavas, S., Hybrid (numerical-experimental) modeling of complex structures with linear and nonlinear components, Nonlinear Dynamics, Vol. 47, pp. 193-217, 2007.

[2] Giagopoulos, D. and Natsiavas, S., Dynamic Response and Identification of Critical Points in the Superstructure of a Vehicle using a Combination of Numerical and Experimental Methods, Experimental Mechanics, DOI 10.1007/s11340-014-9966-z, 2014.

[3] Ewins, D.J., 1984, Modal Testing: Theory and Practice, Research Studies Press, Somerset, England.

[4] Mohanty, P. and Rixen, D.J., Identifying mode shapes and modal frequencies by operational modal analysis in the presence of harmonic excitation, Experimental Mechanics, Vol. 45, pp. 213-220, 2005.

[5] Spottswood, S.M. and Allemang, R.J., On the investigation of some parameter identification and experimental modal filtering issues for nonlinear reduced order models, Experimental Mechanics, Vol. 47, pp. 511-521, 2007.

[6] Richardson, M.H. and Formenti, D.L., Global curve fitting of frequency response measurements using the rational fraction polynomial method, Third IMAC Conference, Orlando, Florida, 1985.

[7] Mottershead, J.E. and Friswell, M.I., Model updating in structural dynamics: A survey, Journal of Sound and Vibration, Vol. 167, pp. 347-375, 1997.

[8] Papadimitriou, C., Ntotsios, E., Giagopoulos, D. and Natsiavas, S., Variability of updated finite element models and their predictions consistent with vibration measurements, Structural Control and Health Monitoring, Vol. 19, pp. 630-654, 2012.

[9] Giagopoulos, D., Papadioti, D.-Ch., Papadimitriou, C. and Natsiavas, S., Bayesian uncertainty quantification and propagation in nonlinear structural dynamics, Proceedings of the IMAC-XXXI 2013, Garden Grove, California, USA, 2013.

[10] Christodoulou K, Ntotsios E, Papadimitriou C, Panetsos P. Structural model updating and prediction variability using Pareto optimal models. Computer Methods in Applied Mechanics and Engineering 2008; 198 (1):138-149.

[11] Mottershead J, Link M, Friswell M. The sensitivity method in finite element model updating: A tutorial. Mechanical Systems and Signal Processing. DOI 10.1016/j.ymssp.2010.10.012, 2010

[12] Grafe H Model Updating of Large Structural Dynamics Models Usong Measured Response Functions. Imperial College of Science, Department of Mechanical Engineering, Dynamics Section, Phd Thesis, 1998

[13] Fritzen CP, Jennewein D, Kiefer T. Damage detection based on model updating methods. Mechanical Systems and Signal Processing 1998; 12(1):163–186.

[14] Allemang, R. and D. Brown (1982). A correlation coefficient for modal vector analysis. In 1st International Modal Analysis Conference, Orlando, USA.

[15] Grafe, H. (1995, May). Review of frequency response function updating methods. Technical report no. 1.01, BRITE-URANUS BRE2-CT94-0946.

[16] Maniatis, I., 2006. Numerical and experimental investigations on the stress distribution of bolted glass connections under in-plane loads. Technisce Universität München.

[17] Mocibob, D., 2008. Glass panel under shear loading-Use of glass envelopes in building stabilization. École polytechnique fédérale de Lausanne.

[18] Mocibob, D. & Crisinel, M., 2007. Glass panel under in-plane shear loading: Experimental investigation on structural glass panel point support.

[19] Giagopoulos, D., Chatziparasidis, I. & Sapidis, N.S., 2015. Structural integrity analysis and optimization of an elevator frame, through fe modeling and experimental tests. In COMPDYN 2015 - 5th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering, pp. 3194–3204.

[20] DYNAMIS 3.1.1, Solver Reference Guide, DTECH, Thessaloniki, Greece, 2013.

[21] ANSA and META-Post, BETA CAE Systems S.A., Thessaloniki, Greece.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

INVESTIGATION IN THE MECHANICAL BEHAVIOR OF LOW-VELOCITY IMPACTED CFRP PLATES UNDER DIFFERENT BOUNDARY CONDITIONS

S. Arslan[1], **M.Z. Sadeghi**[1], **A. Dafnis**[1] **and K.-U. Schroeder**[1]

[1] Institute of Structural Mechanics and Lightweight Design (SLA) RWTH Aachen University
Wuellnerstr. 7, 52062 Aachen, Germany
sueleyman.arslan@rwth-aachen.de, zamaan.sadeghi@sla.rwth-aachen.de, dafnis@sla.rwth-aachen.de, kai-uwe.schroeder@sla.rwth-aachen.de

Keywords: Low Velocity Impact, Carbon Fiber Reinforced Polymer (CFRP), Delamination, Finite Element Method, Boundary Condition (BC), NDT techniques

Abstract. *The mechanical behavior of Carbon Fiber Reinforced Polymer (CFRP) plates with different boundary conditions (BC) under low velocity impact is investigated experimentally using a drop tower test. To achieve different strain rates, the drop tower's impactor is released from different heights. Two NDT techniques namely the Infrared Lock-In Thermography (IR) and the Ultrasonic Test Method (UT) are applied to detect the damage, i.e. invisible delamination patterns in the plates, which are induced by a metallic hemispherical impactor.*

A numerical model is developed to simulate the impact tests using the commercial software LS-Dyna. To validate the developed numerical model experimentally, two BCs namely fixed and quasi-freely supported are designed. However, numerical modeling of realistic fixed BCs (by using steel frames with screws) is difficult to be realized (due to undesirable local clumping effects, which induce initial inhomogeneous stress distributions, etc.) and also designing such a test-rig is a time consuming and expensive process. Therefore, another experimental test-rig is designed as an alternative to overcome such problems in the validation process of the developed numerical model. In the proposed test-rig, plates are directly supported on a specific elastomeric pad to represent a quasi-free BC.

A good agreement between the experimentally detected damage and numerical results is observed. The proposed quasi-free BC shows a high potential to be an alternative for the fixed BCs in the validation process especially for parametric studies. Meanwhile, comparison between two different BCs designed in this work leads to the following results: Using quasi-free BC results to the reduction in the impact load and consequently smaller shear stress distribution and bending moment comparing to plates with fixed supports. Such an effect leads to reduction in the damage size and slight change in the damage pattern in freely supported plates.

1 INTRODUCTION

Composite panels are attracting increasing attention in a great variety of transportation industries (such as aerospace, marine, automotive, etc.) due to their high strength/weight ratio. Such structures might experience the impact event (low or high velocity) by external objects while they are in service. The impact event induces different damages in the composite panels such as matrix cracks, fiber breakage and more importantly delamination (inter-laminar cracking), which are called Barely Visible Impact Damage (BVID). Such damages can drastically reduce the strength of the laminate. The mechanical behavior of composite structures under impact has been well studied over last decades [1-3].

Some parts of the studies on this research domain are focused on the experimental observations combined with measurement techniques (i.e. Ultrasonic C-scan, X-ray Computed Tomography (CT), etc.) in order to provide a better understanding of the damage mechanism happening during the impact event [4-10]

Alternatively, since testing composites experimentally under impact loadings is an expensive and troublesome task, there have been many efforts within last few years in investigating the damage behavior of such structures numerically and analytically. These works mainly provide numerical frameworks for inter-laminar and intra-laminar caused due to the impact event using stress/strain based criteria (for the onset of the damage) and fracture mechanics (for the propagation of the damage) [11-13] or modeling discrete damage by using interface cohesive zone models (CZM) [14-16].

In spite of the efforts carried out to model the damage mechanism numerically, however, experimental testing should be considered to validate developed numerical frameworks. In this contribution, as a part of validation process of a developed numerical model, the effect of BCs on the damage response of CFRP plates under low velocity impact has been investigated. Designing realistic fixed BCs (using steel frames with screws) is an expensive and time consuming process especially for parametric studies. Meanwhile, considering different effects existing in the fixed BC, such as undesirable local effects which leads to initial inhomogeneous stress distributions in the edges of the laminate, in the numerical model is very difficult. Therefore, another alternative test-rig (quasi free BC using elastomeric pads) is proposed for the validation of the developed numerical model in order to overcome the abovementioned problems for the fixed BCs.

A good agreement between the experimental and numerical results is observed. Considering the effort needed for the designing the quasi-free BC – which is much less than the fixed BC, quasi-free BCs have a good alternative for validation process. Meanwhile, the effect of the different BCs under mechanical behavior under low velocity impact ca be summarized as follow: The quasi-freely supported plates can dramatically decrease the impact load leading to smaller bending moment as well as shear stress distribution and consequently reduction in the damage size delamination comparting to the fixed BCs. It is also observed that BC can affect the damage pattern slightly especially in the lower impact velocities.

2 EXPERIMENTAL SETUP

The low velocity impact event is created by using a drop tower test. The drop tower test setup was designed and calibrated already in SLA, RWTH Aachen University and was validated in previous works [17]. The setup has the capability to release the impactor (with the weight range of 0.7 to 2.4 kg) up to 5 meters height on the specimens with maximum dimension of 1400x600x70 mm³ (Figure 1).

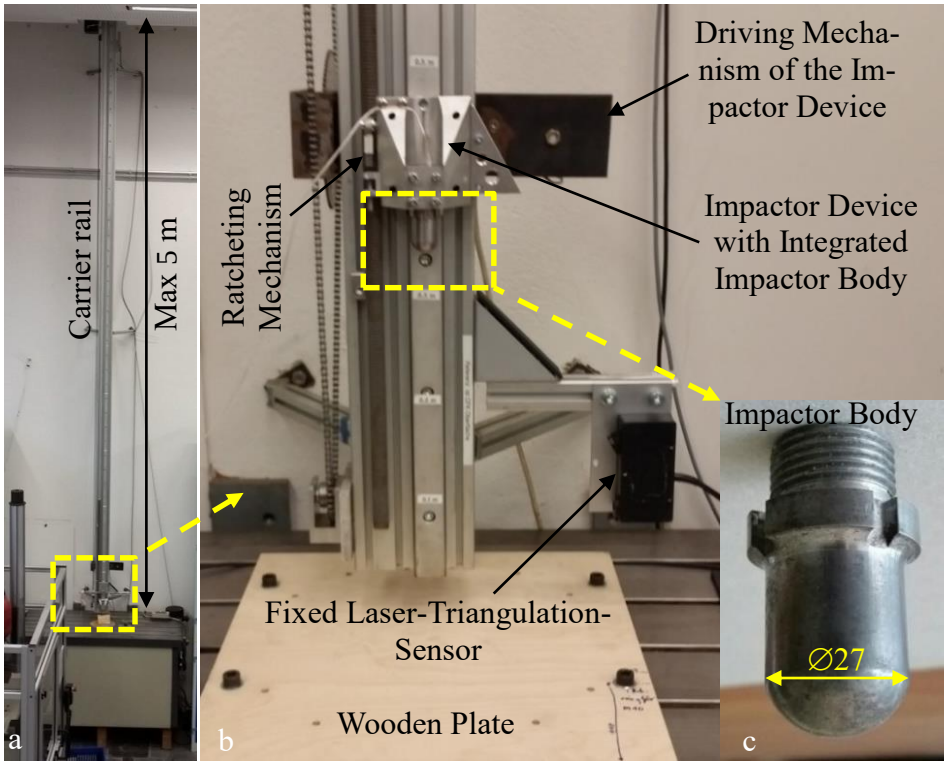


Figure 1: Experimental setup: (a) drop tower, (b) impactor support setup, (c) impactor body

In this work, a cylindrical steel impactor body (Figure 1 (c)) is integrated in an aluminum driving mechanism being able to be released from different heights on the composite panel to achieve strain rates from 480 to 1400 1/s, with respect to the thickness of the plate. The impactor body with the weight of 154 g and diameter of 27 mm has a half-hemispherical head. During the impact event, acceleration and vertical distance time histories are recorded using an accelerometer integrated in the impactor body and a laser-triangulation displacement sensor fixed to the carrier rail. The impactor driving device can be moved electrically to the desired height by using a cable. The corresponding heights are controlled by a distance sensor. To avoid a repeated impact on the plate due to the elastic rebound of the impactor device, a ratcheting mechanism is used (Table 1).

Name of the Component	Commercial Name	Capability
Piezoelectric Accelerometer	ICP Type M350A04	5,000g
Laser-Triangulation Displacement Sensor	Micro-Epsilon Type optoNCDT LD 1605-50	Max. range = ±25 mm Voltage range = ±10 V
Cable Pull with Trip Wire Distance Sensor	Micro-Emsilon Type WDS-5000-P501-S-U	Max. 5 m

Table 1: Measurement system used in the drop tower setup

The properties of the plate are shown in Table 2.

Material	Dimension [mm³]	Number of layers	Stacking sequence
HTA/6376	250x250x6.24	48	$[(90/45_2/-45_2/90/0_2)s(0/45/-45/0)_2]s$

Table 2: Properties of the plate

Two different BCs, fixed and quasi-free, are designed to support the impacted plate in the experimental setup, 5 plates for quasi-free and 5 fixed BC.

2.1 Fixed BC

In order to create a fixed BC for the plates, a specific setup is designed. The plate is clamped between two rectangular steel frames (steel-S235JR) having the same size of the composite panels. In order to ensure constant pressure distribution over the pre-stressed panels' area, 32xM8 screws with washers are well distributed around the steel frames (Figure 2).

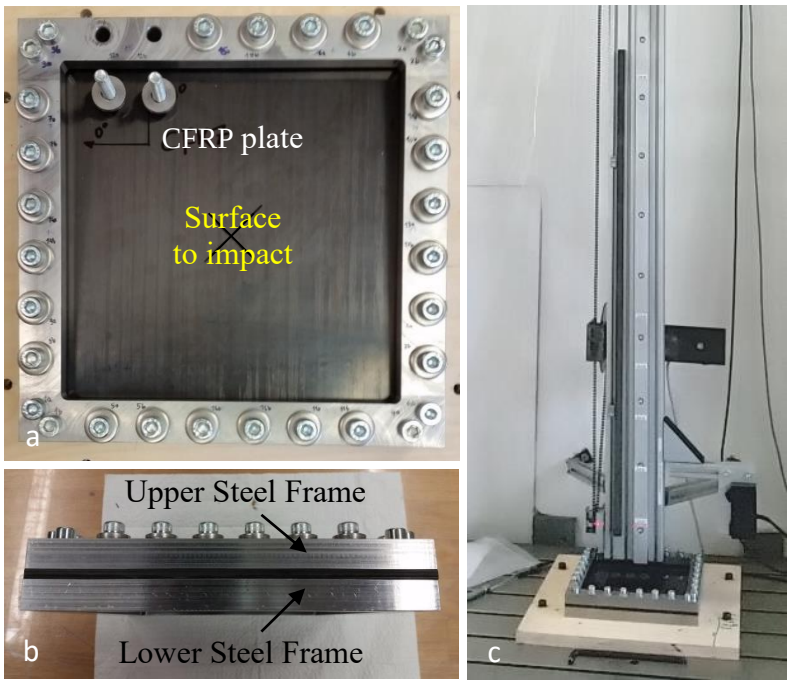


Figure 2: Preparation of the fixed BC: (a) clamping the laminate between two steel frames, (b) side view, (c) attaching the clamped laminate to the wooden part of the drop tower

This arrangement allows acceptable comparison with the numerical model developed for fixed BC. To achieve the constant pressure distribution over the pre-stressed area of the laminate, 70% of the allowable torsional moment for the M8 screws is applied for any screws in a gradual increasing manner. Once the laminate is clamped between the steel frames, the whole frame is bonded to a wooden plate by using 8xM6 screws. The wooden plate attached to the drop tower is used as an interface in order to damp wave transmissions initiated during the impact event.

2.2 Quasi-free BC

Quasi-free BC on the plates is achieved by the use of the silicon elastomer named Protectelast®. The behavior of Protectelast under different strain rates was experimentally investigated in previous study [18]. Having had a high capability of elasticity range, 80 % in compression and 640 % in tension, such elastomers can be used as alternative supporting material for representing a quasi-free condition. Using such a material in the BC designed for the plates leads to less constraints in the plate edges and consequently reduction in the stress and strain fields in the laminate edges in comparison to fixed BC. To this end, two rectangular

aluminum frames with the dimension 305x305x5 mm³ are manufactured to carry the elastomer pads with thickness of 20 mm (Figure 3).

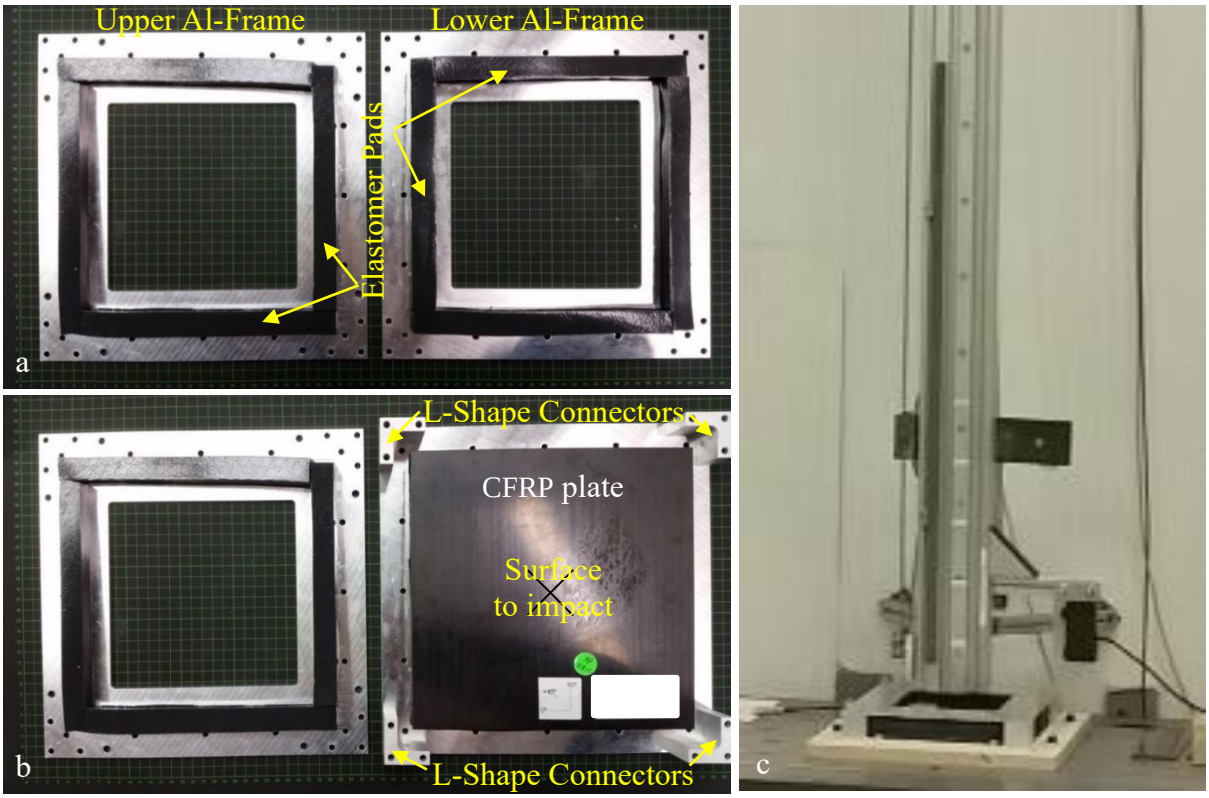


Figure 3: Preparation of the quasi-free BC: (a) inserting elastomers in the aluminum frames, (b) putting the plate on the lower elastomer pads, (c) attaching the clamped laminate to the wooden part of the drop tower

The surface of the aluminum frames are nicely milled to make sure that elastomer pads can be fit inside of the frames. After putting the laminate on the elastomer pad of lower aluminum frame, the upper aluminum frame is assembled on the lower one by using four aluminum L-shape connectors to keep a constant distance between two frames. This distance allows the laminate moves freely between the elastomer pads during the impact event. The final assembly is attached to the wooden plate by using 8xM6 screws.

3 FINITE ELEMENT MODEL

The numerical model in this work is developed by using LS-Dyna software. The FE model includes the modeling part for the impactor body, the plate and two BCs applied in the experiments.

3.1 Impactor

The steel impactor body (0.154 kg) and its aluminum impactor device (0.537 kg) are modeled as half-hemispherical solid body with steel material property. However, the weight of the aluminum is considered in the density of the impactor body model. The mesh size at the front side of the impactor, which is in contact with the plate, is about 1.6 mm. The desired velocity is applied to whole impactor body. For the impact event, a node-to-surface contact is used between the impactor and first sub-laminate (explained later). Due to the hemispheric shape of

the impactor, the impactor is selected as the slave and the sub-laminates of the plate as the master.

3.2 CFRP plate

The plate is modeled as shell elements using material-enhanced-composite-damage (MAT_54) as material card in LS-Dyna. This material card gives the possibility to consider the reduction of compressive and tensile strength of the fiber and matrix after intra-laminar failure. Meanwhile, this material card provides the softening reduction factor for material strength in crash front elements. MAT_54 uses Chang-Chang failure criteria [19]. Material properties used for the plate are listed in Table 3.

Description	Values	Unit
Density	1620	kg/m³
Longitudinal Young's Modulus	145	GPa
Transverse Young's Modulus	10.3	GPa
Through-Thickness Young's Modulus	12.1	GPa
Major Poisson's Ratio	0.301	--
Major Transverse Poisson's Ratio	0.5	--
In-Plane Shear Modulus	5.3	GPa
Transverse Shear Modulus	5.275	GPa
Through-Thickness Shear Modulus.	3.95	GPa
Longitudinal Compressive Strength	1600	MPa
Longitudinal Tensile Strength	2250	MPa
Transverse Compressive Strength	290	MPa
Transverse Tensile Strength	64	MPa
In-Plane Shear Strength	98	MPa

Table 3: Material properties for HTA/6376 [22], [23]

Modeling of the Inter-laminar Behavior

As for the modeling of the delamination pattern induced by the impact event, interface layer based on CZM is used. Interface CZM layers are widely being used by researchers due to many advantages they offer in comparison with other damage modeling techniques [20]. For instance, they are able to predict both the onset and propagation of the damage and they are easy to be implemented in most of simulation software. Since modeling the entire thickness of the CFRP plate with interface layers is quite time consuming (in this case 47 interface layers should be implemented), a special approach is used to model the delamination behavior of this plate. It was previously observed that under impact event, likelihood of delamination is higher between two plies with different fiber orientation [17]. Based on this approach and in order to save the analysis time, we only consider the likelihood of the delamination happening during the impact event only in the layers with the highest differences in their fiber orientations (in this case 90 degree). In this way, for the plate discussed in this work including 48 plies, 14 interface CZM layers have been proposed (Figure 4).

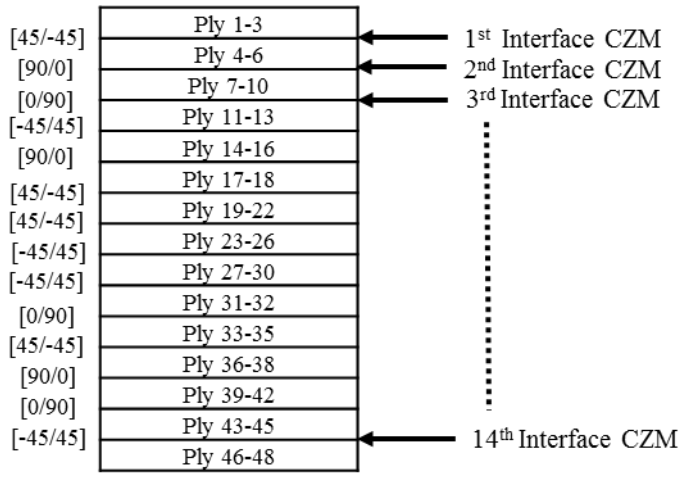


Figure 4: Arrangement of interface CZM layers for the plate

Two numerical methods are available for modeling the interface CZM: the cohesive-zone-element- and the tiebreak-contact-method. In this work tiebreak-contact-method is implemented for modeling the inter-laminar damage, since the cohesive-zone-element-method is too time consuming and therefore unsuitable. Tiebreak option 11 is chosen since it allows using coarser mesh. This contact uses B-K criteria for damage propagation (equation 1). For damage initiation, quadratic nominal stress criterion is applied (equation 2).

$$G_{IC} + (G_{IIC} - G_{IC}) \left(\frac{G_{II}}{G_I + G_{II}}\right)^{\mu} = G_C \tag{1}$$

$$\left(\frac{\langle \sigma_n \rangle}{N_{max}}\right)^2 + \left(\frac{\langle \sigma_s \rangle}{S_{max}}\right)^2 + \left(\frac{\langle \sigma_t \rangle}{T_{max}}\right)^2 = 1 \tag{2}$$

The mechanical properties for the onset and propagation of the damage for the interface CZM layers are shown in Table 4.

Description	Symbol	Values	Unit
Strain Energy Release Rate for Mode I	G_{IC}	260	J/m²
Strain Energy Release Rate for Mode II	G_{IIC}	1002	J/m²
Normal Failure Stress/Peak for Mode I	T	30	MPa
Shear Failure Stress/Peak for Mode II	S	60	MPa
B.K parameter	μ	1.2	--

Table 4: Required parameters for the interface cohesive zone behavior of HTA/6376 [22], [23]

Two different quadratic mesh sizes are considered for the plate. A circular hotspot area with a finer mesh (1.25 mm) is designed with 1.5 time of the biggest delamination area diameter induced by the highest velocity. Coarser mesh is considered outside of this area (2.5 mm). The plate consists of 13,792 elements per sub-laminate (total of 206,880 elements for the whole plate). Since the minimum number of elements needed for CZM should be three elements to capture the process zone happening in the damage propagation [21], a special scaling factor is used to calculate the peak tractions (mode I and II) for coarser mesh sizes.

More explanation about the element sizes and the corresponding scaling factors was provided in a previous work [18].

3.3 Boundary Condition (BC)

The numerical model for both BCs is developed with the same dimensions of the original specimens explained in the section 2.1 and 2.2. For more details, a section cut of the model is shown in Figure 5. In fixed BC, the upper and lower steel frames are modeled with solid elements (mesh size of 2.5 mm). The contact between the upper steel frame and the first sub-laminate as well as the contact between the lower steel frame and the last sub-laminate are both considered tied contact. In this way, the constant pressure caused by screws in the specimens is guaranteed. For quasi-free BC, the aluminum frames are modeled by using shell elements whereas the elastomers are modeled as solid elements (mesh size for both elastomers and aluminum frames is 2.5 mm). The mechanical behavior of the elastomer pads under different strain rates were already investigated in a previous study [18].

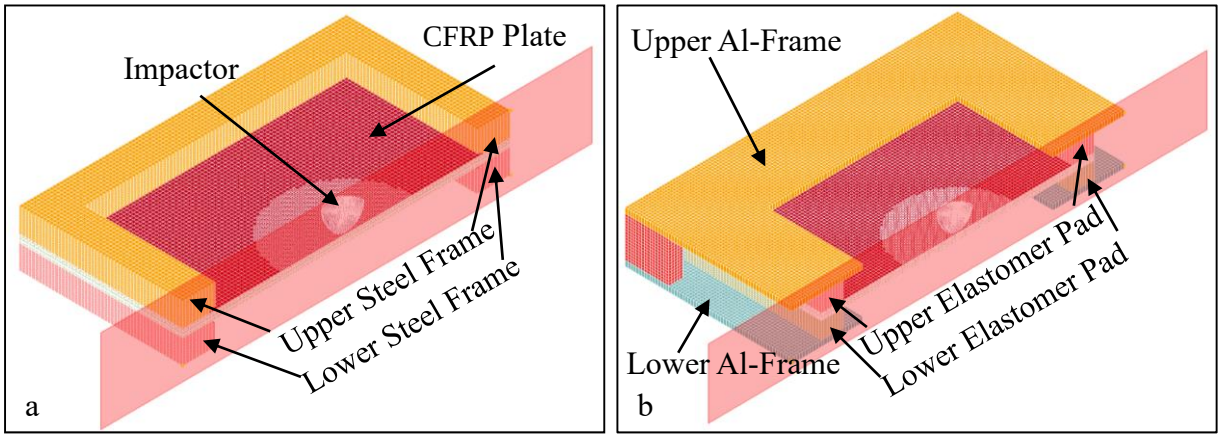


Figure 5: FE Model developed for (a) fixed and (b) quasi-free BC

4 RESULTS AND DISCUSSION

As previously mentioned in section 2, 10 sets of experiments in total are carried out in order to investigate the effect of fixed and quasi-free supported plates under different velocities ranging from 3 to 9 m/s. In this section the effect of these two BCs on the impact response of the plates under different impact velocities have been discussed numerically and experimentally.

Impact velocity and rebound behavior

Some experimental results of the impact test are listed in Table 5 and a comparison of velocities is made in Figure 6. The velocities V_I and V_R are estimated by measuring the slope of the displacement-time curve of the vertical movement of the impactor (measured by the laser-triangulation sensor). The residence time t_R is the period between the time at which the impactor hits the plate (V_I) and the time at which the impactor separates from the plate (V_R).

BC–CFRPx	Drop height h [m]	Impact velocity V_I [m/s]	Residence time of the impact event t_R [ms]	Resultant Deflection d_{max} [mm]	Rebound velocity V_R [m/s]
Quasi-free–CFRP1	1	4.33	4.865	-3.54	1.86
Quasi-free–CFRP2	2	6.22	4.870	-4.92	2.79
Quasi-free–CFRP3	3	7.60	4.745	-6.03	3.12
Quasi-free–CFRP4	4	8.54	4.660	-6.80	3.37
Quasi-free–CFRP5	4.61	8.79	4.645	-6.82	3.44
Fixed–CFRP6	0.5	3.05	<1.755*	<-2.57*	2.50
Fixed–CFRP7	1	4.37	1.755	-2.57	3.71
Fixed–CFRP8	2	6.17	1.795	-3.48	5.18
Fixed–CFRP9	3	7.52	1.900	-4.37	5.90
Fixed–CFRP10	4	8.74	1.890	-5.81	6.61

Table 5: Results for the impact tests

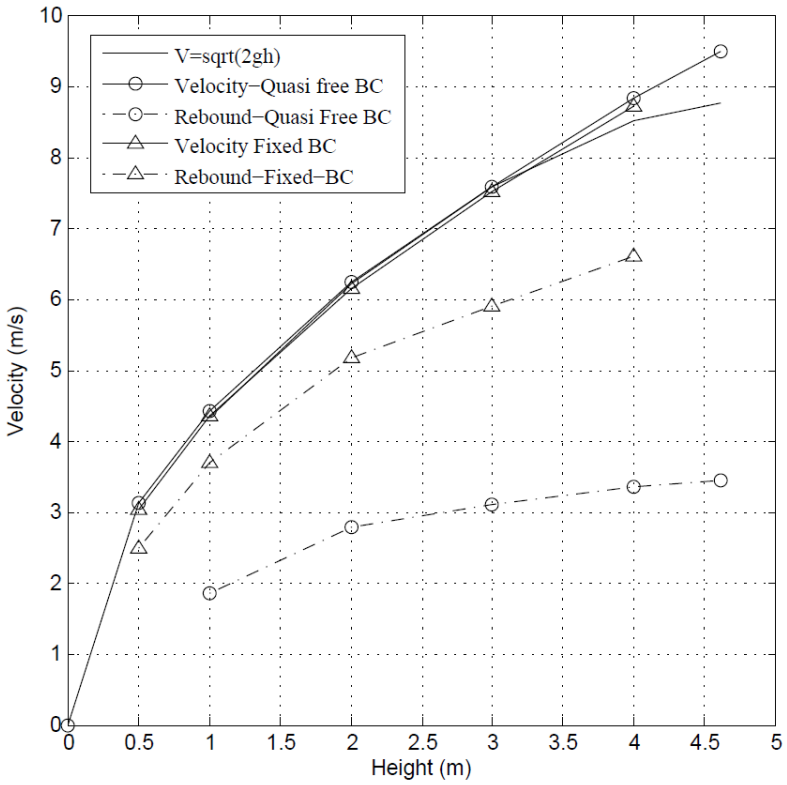


Figure 6: Comparison of the impact and rebound velocities for different BCs

As expected, almost all 10 impact velocities are close to the physical law $V = \sqrt{2gh}$, which states that friction effects and other disturbance have a very small influence on the velocity. However, as it is shown in Figure 6 the rebound velocity is highly influenced by the type of the BCs. Existence of elastomer in quasi-free BC leads to reduction in the rebound velocity comparing to corresponding plate with the fix support. This reduction mainly induced by the compression of the elastomer within the impact event.

The residence time t_R for the fixed BCs is shorter than that of the quasi-free. Meanwhile, the resultant deflection d_{max} is for the fixed BC smaller than for the quasi-free.

---

*Could not be measured exactly

In Figure 7, the resultant deflection d_{max} and residence time t_R for CFRP9 and CFRP3 is shown. Such a response is also observed in the other tests. The fixed BC shows a simple impact-rebound-behavior whereas the quasi-free shows a delayed response with a greater resultant deflection, because of the existing of the elastomer. After reaching the maximum value of -6.03 mm the quasi-free shows two rebounding behaviors.

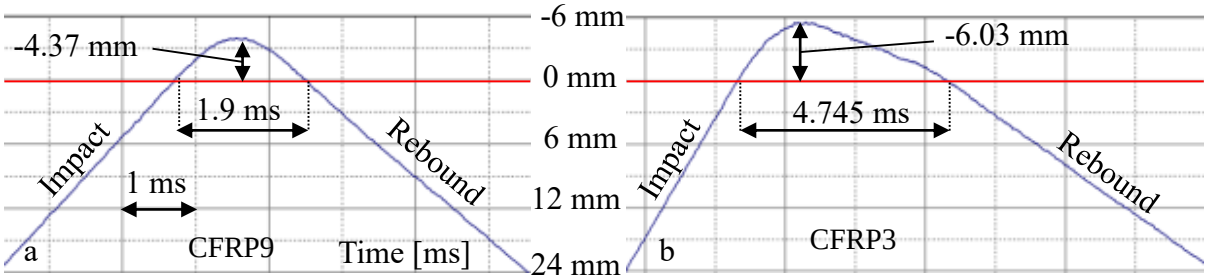


Figure 7: Comparison of the resultant deflection d_{max} and residence time of the impact event t_R for CFRP9 (a) (quasi-free) and CFRP9 (b) (fixed) (red line: CFRP plate surface on the impact side)

Impact Force

Comparison of the impact force – FE result – between the impactor and the plate for both quasi-free and fixed BC (CFRP5 and 10) is shown in Figure 8.

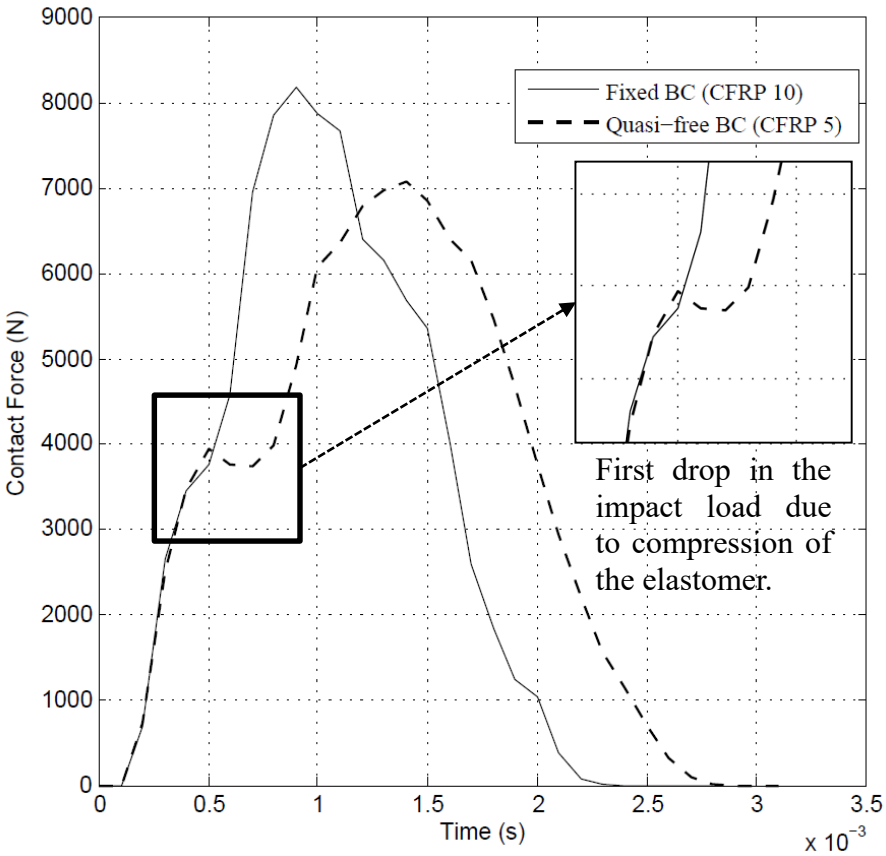


Figure 8: The impact force for fixed and quasi-free BC ($V_I \cong 9$ m/s)

In quasi-freely supported plates, elastomer undergoes sever compression during the impact event causing a drop in the contact force comparing to the plates supported with fixed BC (the frame is rigid). At the same time at which elastomer causes a drop in the contact force for quasi-free BC, contact load for the plate with fixed support shows a disturbance in the load response, which clearly indicate the effect of fixed BC and introducing more bending moment to the system, which dramatically increases the contact force. At the peak point of the both curves, the maximum delamination area is observed in which the damage area remains constant over time.

Damage size

For experimental observations two measurement techniques namely Ultrasonic and active Lock-in Thermography technique are used in order to determine the damage induced by the impact event in plates. Ultrasonic method is applied on top and through the thickness of the impacted plates to spot the damage area. This technique provides us qualitative and distinctive images of the damaged area but not quantitatively. This technique only provides the concentration of the damage on the first damaged area, which is detected by the device and the damage on the other layers are not shown in the results. Thermography technique provides us top view of the damage accumulation of the impacted plates. This technique is only able to detect damages happening up to around 4 mm from the surface of the plate.

In Figure 9, a comparison for the induced damage between two panels (CFRP5 and 10), which are impacted by the same velocity ($\cong 9\, m/s$) but different boundary conditions is shown, which includes Ultrasonic, Thermography (phase image) and FE observations.

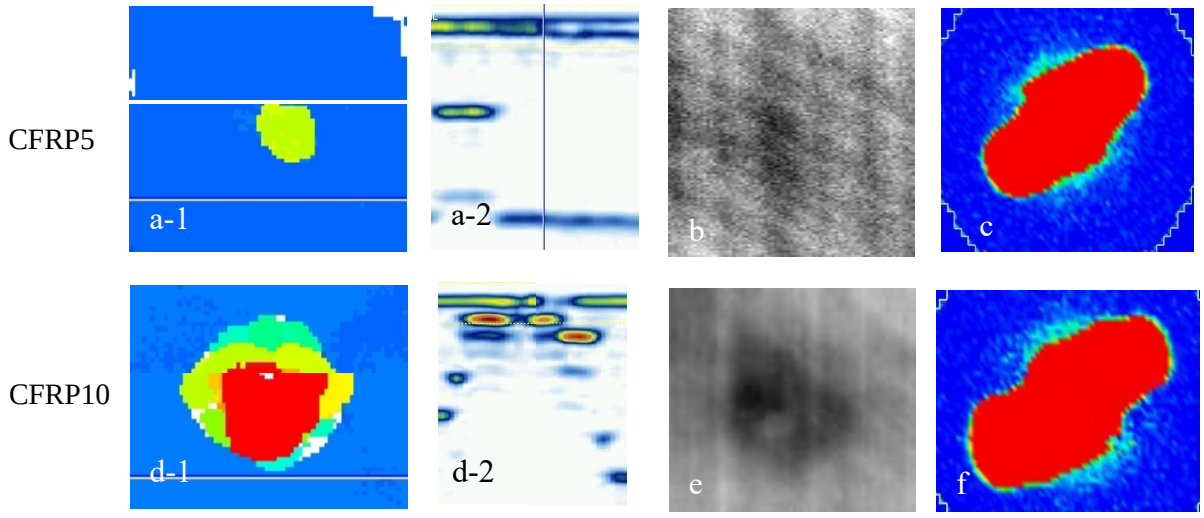


Figure 9: Comparison of the delamination: (a) and (d) ultrasonic, (b) and (e) thermography, (c) and (f) numerical

In Ultrasonic observations shown in Figure 9 the color red (the images on the top view) represents the damages exists on the first interface layer close to the surface of the plates, yellow represents the damages exist in the middle, green represents the damages in the sub laminates close to the bottom of the plates, and blue represent the undamaged area (bottom of the plate).

At the first glance, the comparison between the plates shows that the size of the induced damage by the impactor in fixed BC is bigger than the same panel with quasi-free BC. The same result is observed for other panels too. It can be attributed to the bigger bending moment

(and also shear stress distribution) existing in the panels supported with fixed BC comparing to quasi-freely supported panels leading to bigger damage in the panels.

For the CFRP5 (quasi-freely supported) the biggest induced damage is in the middle, the same result in the FE model in which interface layer number 7 shows the biggest damage area comparing to other interface layers. The damage induced in the middle of the CFRP5 is shown in the through the thickness Ultrasonic image (Figure 9 (a-2)). In the CFRP10, the top view of the Ultrasonic image shows that damage in the middle of the plate is bigger (yellow color) showing a good agreement with the FE results (7th interface layer). The through thickness Ultrasonic image (Figure 9 (d-2)) shows that the damage is only on the surface since the system is only able to spot the first damaged region (region close to the surface of the plate (red color in (d-1) and the rest of the damaged area – in this view – cannot be spotted. Thermography observations (Figure 9 (b) and (e)) show a good agreement with the other results. Comparison between Figure 9, (b) and (e) clearly shows that the damage due to the impactor under fixed BC is bigger that quasi-free BC.

Distribution of the shear stress for both BCs is shown in Figure 10 showing greater stress contour in the fixed BC especially at the edges.

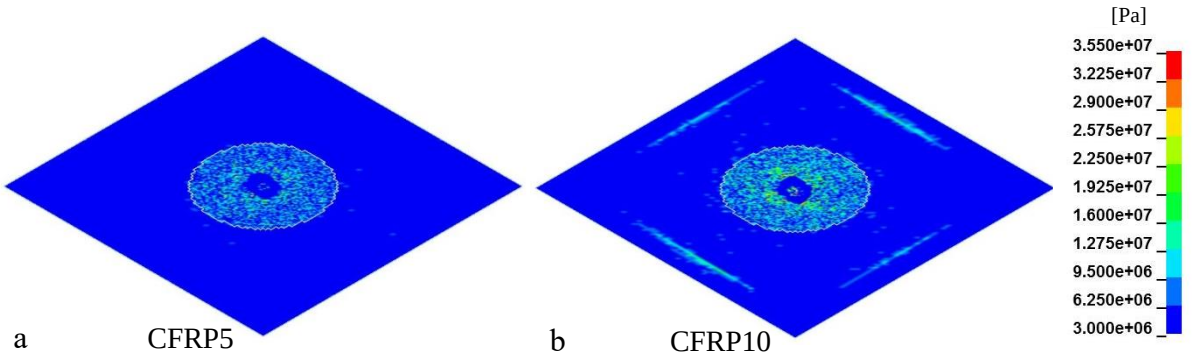


Figure 10: Shear distribution of the first interface layer: (a) quasi-free, (b) fixed BC

The evolution of the damage area over time for the composite panel for different BCs under different impact velocity is shown in Figure 11.

It can be clearly seen that for a certain velocity, the damage area is always bigger for the fixed supported panels. In both boundary conditions, a frustum-conical damage pattern is observed. For the plates impacted under fixed boundary condition this pattern is larger (in both upper and lower area of the frustum). However, when it comes to the quasi-free BC, the differences between the upper base and lower base of the frustum-conical damage pattern is becoming smaller (especially under lower velocities (CFRP1)) and damage pattern is more circular like. As it was previously discussed, Lock-in Thermography camera works based on the change of the frequency. In this method, the values of the frequency should be changed till a visible damage accumulation is achieved. However, this method is only able to detect the damage exists up to around 4 mm thickness (from the surface of the plate). Since in this work, all the plates have the same thickness of 6.24 mm, detecting the damage accumulation existing in the lower layers (bigger than 4 mm depth) is very difficult.

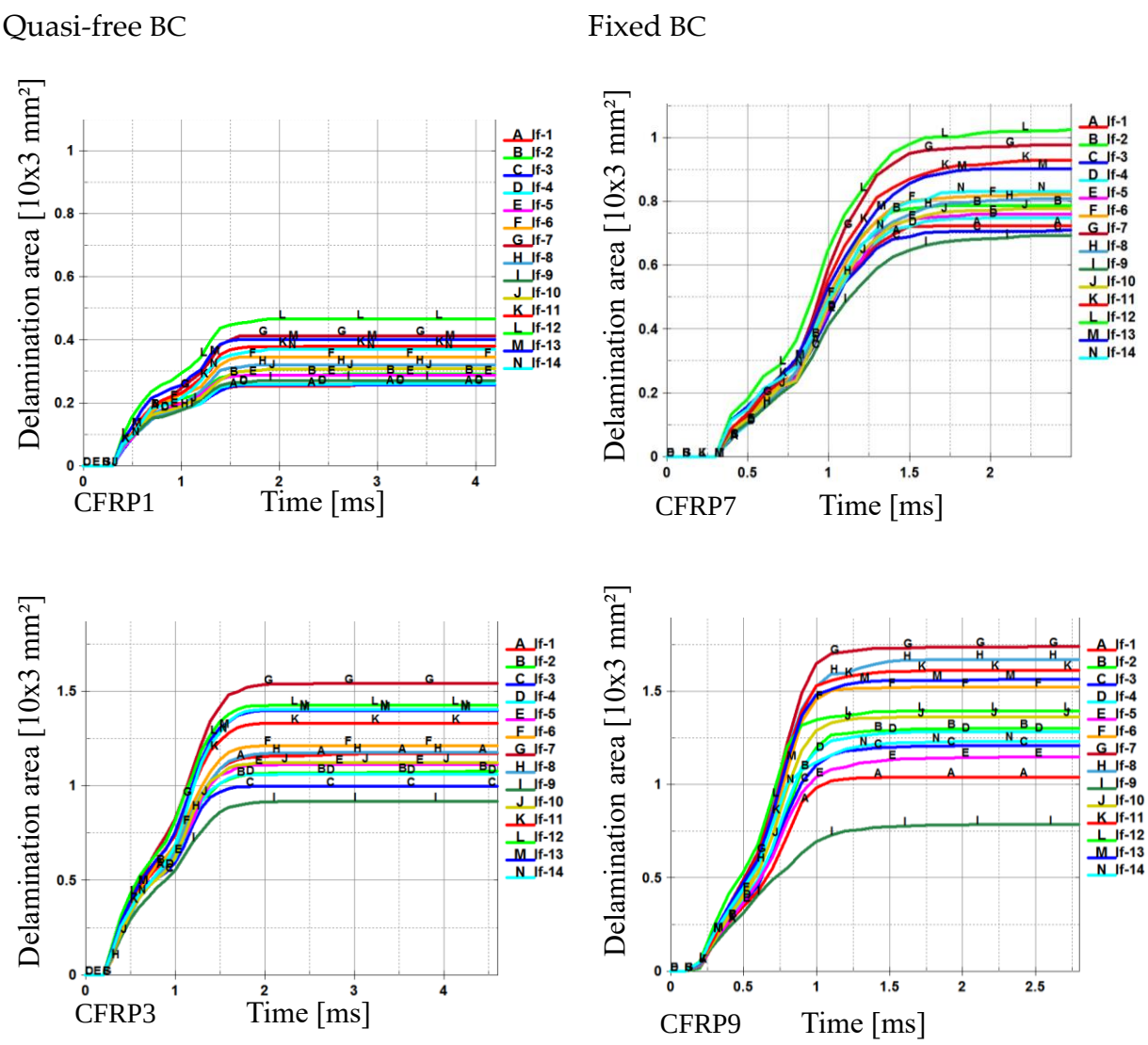


Figure 11: Delamination area from numerical results (A lf-1=1st and N lf-14=14th interface CZM layer)

A good example is the comparison between CFRP9 and CFRP3, same panel impacted under velocity of about 7.6 m/s but with different BCs. Due to the constrained applied in CFRP9, the induced impact damage is bigger comparing to the other plate. Figure 12 shows that the suitable image of the damage accumulation can be simply detected in such a panel by changing the lock-in frequency.

In contrast, CFRP3 – which is impacted under quasi-free BC – experiences smaller damages in the layers. In Figure 13, the growth of the area of the delamination versus elapsed time is shown. The induced damage in most of the interface layers is very small and there is only in the interface layer number 12 in which the damage area is slightly bigger. However, it is difficult in Lock-in Thermography method to detect such a damage, which is located in a depth of about 5 mm from the surface of the plate. Hence the reason, Lock-in Thermography is not able to spot a decent damage accumulation in the plate. Figure 14 visibly shows that even changing the frequencies does not detect a good damage accumulation.

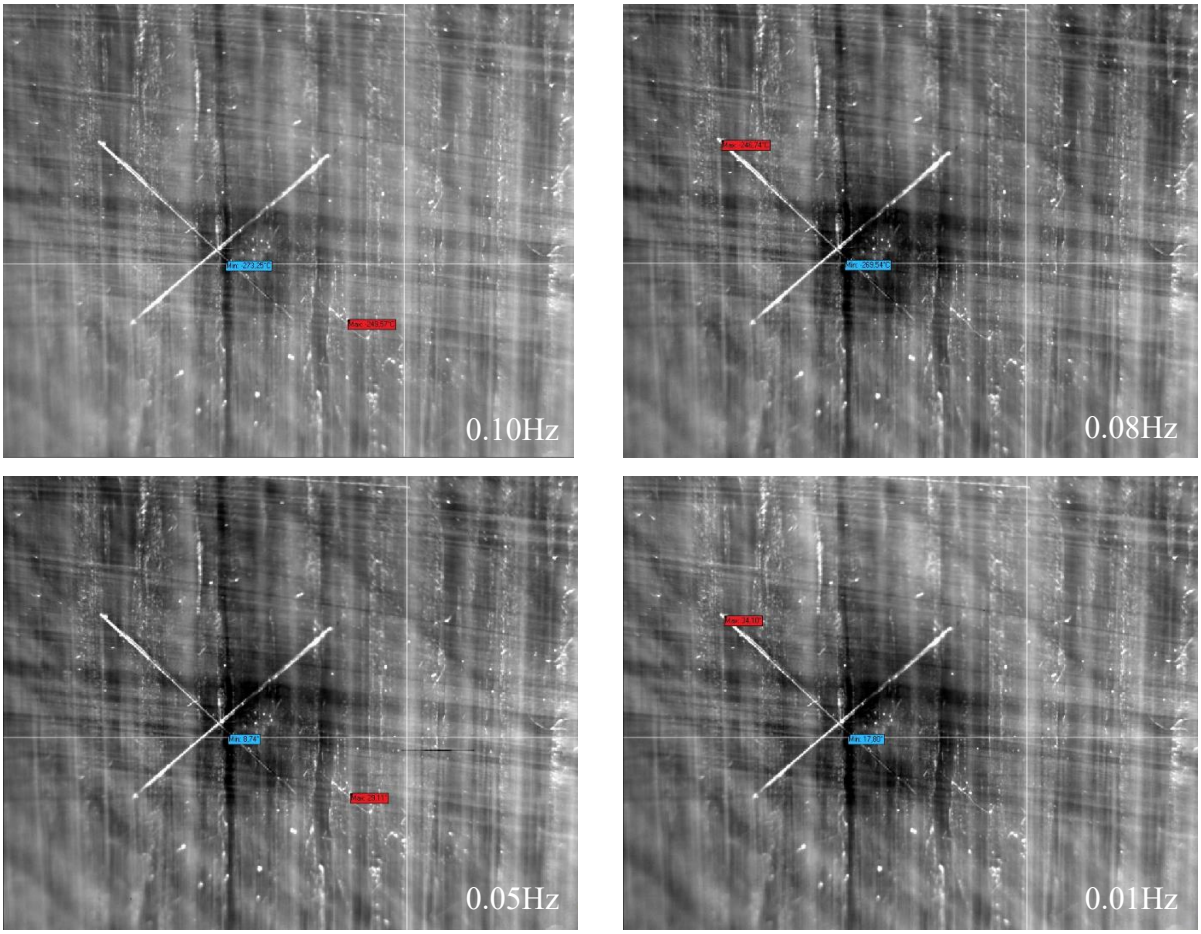


Figure 12: Damage measurement using Lock-in Thermography camera with different lock-in frequencies-CFRP9

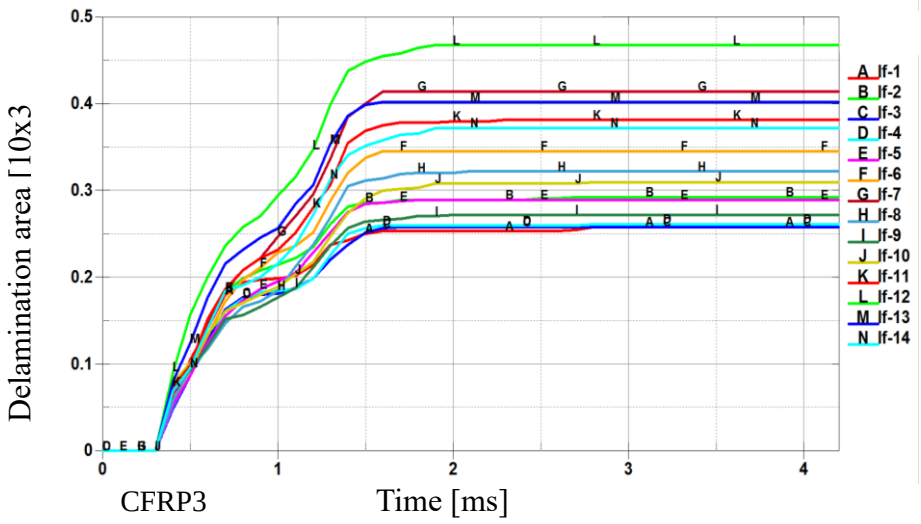


Figure 13: Delamination area of numerical results (A lf-1=1st and N lf-14=14th interface CZM layer)

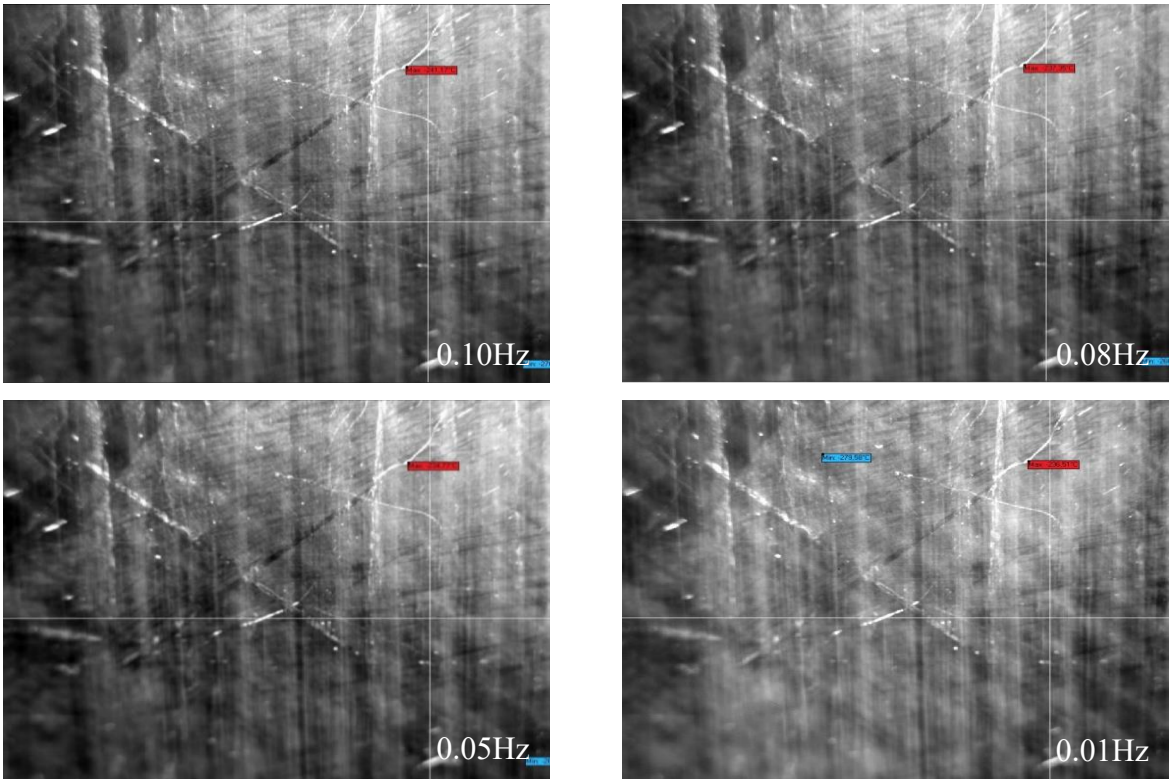


Figure 14: Damage measurement using Lock-in Thermography camera with different lock-in frequencies-CFRP3

5 CONCLUSIONS

The mechanical behavior of Carbon Fiber Reinforced Polymer (CFRP) plates (made of HTA/6376 with the dimension of 250x250x6.24 mm³) and with different boundary conditions (BC) under low velocity impact is investigated experimentally using a drop tower test. Two NDT techniques namely the Infrared Lock-In Thermography (IR) and the Ultra-sonic Test Method (UT) are applied to detect the damage induced by a metallic hemispherical impactor.

A numerical model is developed to simulate the impact tests using the commercial software LS-Dyna. In order to validate the developed numerical model experimentally, two boundary conditions namely fixed and quasi-freely supported are designed. A good agreement between the experimental observations using NDT methods and numerical results is achieved. The proposed quasi-free BC shows a high potential to be an alternative for the fixed BCs in the validation process especially for parametric studies. Meanwhile, comparison between two different boundary conditions designed in this work leads to the following results: Using quasi-free BC reduces the impact load and consequently smaller shear stress distribution and bending moment comparing to plates with fixed supports. This leads to reduction in the damage size and slight change in the damage pattern in freely supported plates.

ACKNOWLEDGEMENT

The authors wish to thank the European Commission Research Directorates financial support in the framework of Clean-Sky project JTI-CS-2011-03-SFWA-02-019. We also would like to give a special thanks to Werkzeugmaschinenlabor (WZL) RWTH Aachen who provided us with the Ultrasonic images.

REFERENCES

[1] M.O.W. Richardson, M.J. Wisheart, Review of low-velocity impact properties of composite materials, Composites Part A: Applied Science and Manufacturing, Volume 27, Issue 12, 1996, Pages 1123-1131.

[2] W.J. Cantwell, J. Morton, The impact resistance of composite materials — a review, Composites, Volume 22, Issue 5, 1991, Pages 347-362.

[3] Agrawal, Sandeep, Kalyan Kumar Singh, and P. K. Sarkar. "Impact damage on fibre-reinforced polymer matrix composite–a review." Journal of Composite Materials (2013): 0021998312472217.

[4] M.V. Hosur, C.R.L. Murthy, T.S. Ramamurthy, Anita Shet, Estimation of impact-induced damage in CFRR laminates through ultrasonic imaging, NDT & E International, Volume 31, Issue 5, October 1998, Pages 359-374.

[5] Catherine Potel, Thierry Chotard, Jean-François de Belleval, Malk Benzeggagh, Characterization of composite materials by ultrasonic methods: modelization and application to impact damage, Composites Part B: Engineering, Volume 29, Issue 2, 1998, Pages 159-169

[6] F Aymerich, S Meili, Ultrasonic evaluation of matrix damage in impacted composite laminates, Composites Part B: Engineering, Volume 31, Issue 1, January 2000, Pages 1-6.

[7] B.S. Ben, B.A. Ben, K.A. Vikram, S.H. Yang, Damage identification in composite materials using ultrasonic based Lamb wave method, Measurement, Volume 46, Issue 2, February 2013, Pages 904-912, ISSN 0263-2241.

[8] M. Tehrani, A.Y. Boroujeni, T.B. Hartman, T.P. Haugh, S.W. Case, M.S. Al-Haik, Mechanical characterization and impact damage assessment of a woven carbon fiber reinforced carbon nanotube–epoxy composite, Composites Science and Technology, Volume 75, 11 February 2013, Pages 42-48.

[9] S.A. Grammatikos, E.Z. Kordatos, T.E. Matikas, C. David, A.S. Paipetis, Current injection phase thermography for low-velocity impact damage identification in composite laminates, Materials & Design, Volume 55, March 2014, Pages 429-441.

[10] E. Abisset, F. Daghia, X.C. Sun, M.R. Wisnom, S.R. Hallett, Interaction of inter- and intralaminar damage in scaled quasi-static indentation tests: Part 1 – Experiments, Composite Structures, Volume 136, February 2016, Pages 712-726.

[11] P. Maimí, P.P. Camanho, J.A. Mayugo, C.G. Dávila, A continuum damage model for composite laminates: Part II – Computational implementation and validation, Mechanics of Materials, Volume 39, Issue 10, October 2007, Pages 909-919

[12] M.V. Donadon, L. Iannucci, B.G. Falzon, J.M. Hodgkinson, S.F.M. de Almeida, A progressive failure model for composite laminates subjected to low velocity impact damage, Computers & Structures, Volume 86, Issues 11–12, June 2008, Pages 1232-1252

[13] Y. Shi, T. Swait, C. Soutis, Modelling damage evolution in composite laminates subjected to low velocity impact, Composite Structures, Volume 94, Issue 9, September 2012, Pages 2902-2913.

[14] F. Aymerich, F. Dore, P. Priolo, Prediction of impact-induced delamination in cross-ply composite laminates using cohesive interface elements, Composites Science and Technology, Volume 68, Issue 12, September 2008, Pages 2383-2390.

[15] M.F.S.F de Moura, J.P.M Gonçalves, Modelling the interaction between matrix cracking and delamination in carbon–epoxy laminates under low velocity impact, Composites Science and Technology, Volume 64, Issues 7–8, June 2004, Pages 1021-1027.

[16] Stephen R. Hallett, Wen-Guang Jiang, Bijoysri Khan, Michael R. Wisnom, Modelling the interaction between matrix cracks and delamination damage in scaled quasi-isotropic specimens, Composites Science and Technology, Volume 68, Issue 1, January 2008, Pages 80-89

[17] S. Woeste, Impact investigation of dog-bone CFRP specimens, PhD thesis, Institute for lightweight design, RWTH Aachen, 2005.

[18] S. Arslan, A. Dafnis, K.-U. Schroeder, Simulation of Bird Strikes on CFRP Plates as a Part of a Validation Process, ICEAF IV, Skiathos, Greece, 2015.

[19] Chang, Fu-Kuo; and Chang, Kuo-Yen: A Progressive Damage Model for *Laminated* Composites Containing Stress Concentrations. *J. Compos. Mater.*, vol. 21, Sept. 1987, pp. 834-855.

[20] M.R. Wisnom, Modelling discrete failures in composites with interface elements, Composites Part A: Applied Science and Manufacturing, Volume 41, Issue 7, July 2010, Pages 795-805.

[21] Turon, A., Davila, C.G., Camanho, P.P.; "An Engineering Solution for Using Coarse Meshes in the Simulation of Delamination with Cohesive Zone Modeling", NASA Technical Report NASA/TM-2005-213547. Langley Research Center, Hampton, Virginia, March 2005.

[22] M. Ericsson, Simulating Bird Strike on Aircraft Composite Wing Leading Edge, Master Thesis, KTH Engineering Sciences, Stockholm, Sweden, 2012

[23] K.I. Tserpes, A three-dimensional progressive damage model for bolted joints in composite laminates subjected to tensile loading, Blackwell Science Ltd. Fatigue Fract Engng Mater Struct 24, 663–675, 2001

[24] LS-Dyna Aerospace Working Group, Modeling Guidelines Document, 2015

[25] F. Dolce F, Blast Impact Simulation on Composite Military Armours, Department of Mechanical Engineering, University of BATH, 2009

[26] LS-Dyna User's Manual I and II, 2015

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A NEW SHELL FINITE ELEMENT WITH DRILLING DEGREES OF FREEDOM AND ITS RELATION TO EXISTING FORMULATIONS

Robert Winkler[1], **Dimitrios Plakomytis**[1,2]

[1]Institute of Mechatronics, University of Innsbruck
Technikerstrasse 13, 6020 Innsbruck
e-mail: Robert.Winkler@uibk.ac.at

[2] INTALES GmbH
Innsbrucker Strasse 1, 6161 Natters
e-mail: Dimitrios.Plakomytis@uibk.ac.at

Keywords: Shell theory, drill rotation, rotation constraint, Allman shape functions

Abstract. *This contribution deals with shell element formulations involving three rotational degrees of freedom. Its purpose is twofold. Firstly, a review of relevant non-linear shell theories as well as finite element implementations is given and corresponding defects are addressed. Secondly, a geometrically linear shell element formulation is presented which is free of any of these defects. Conventional shell theories involve two rotational degrees of freedom, the drilling rotation being excluded. Three standard methods of incorporating the drilling degree of freedom are considered: (a) The application of a proper rotation constraint condition, (b) shell theories intrinsically involving three rotational degrees of freedom (here called 'micropolar theories') and (c) introducing the drilling degree of freedom at the level of the finite element discretization (Allman-type shape functions). Here, reference implementations relying on approaches (a) to (c) as well as their reasonable combinations are considered. It is demonstrated that each of these implementations reveals at least one questionable feature: Formulations based on a micropolar approach involve ad hoc assumptions related to the constitutive law. Formulations involving a rotation constraint imply the application of a problem-dependent penalty or regularization parameter. Finally, Allman-type shell elements suffer from a lower convergence rate for bending dominated problems compared to isoparametric ones. The proposed shell element relies on a conventional (nonpolar) shell theory and applies a drill rotation constraint via a penalty formulation. A crucial point is the application of properly designed enhanced strain fields to avoid in-plane locking phenomena. It is demonstrated numerically that the resulting implementation overperforms existing shell elements. Most importantly, the way of incorporating the rotation constraint proves to constitute a proper penalty formulation in the sense that for sufficiently large values of the penalty parameter the results are independent of the latter. Moreover, this threshold is independent of the problem. The relation to existing implementations of the drill rotation constraint, all of them requiring the application of problem dependent penalty parameters, is discussed. Concluding, it is explicated how the present formulation derives from a micropolar approach.*

1 Introduction

From a practical point of view, it is generally agreed that shell finite elements with six degrees of freedom (dofs) at each node exhibit substantial advantages compared to those with just five dofs or even those with a varying number (5/6 dof formulations). From a theoretical or numerical point of view, some aspects of 6 dof formulations still deserve additional clarification. The implementation of the additional dof, i.e. the so called 'drilling' rotation, is achieved either by constraining the drilling dof to the in-plane rigid-body rotation (a), applying shell theories intrinsically involving three rotational degrees of freedom (b), or on the level of the finite element discretization applying so-called Allman-type shape functions (c).

State-of-the-art shell finite element formulations are based on mixed variational principles, the Hu-Washizu principle, e.g. In contrast to formulations relying on the total potential energy, these mixed formulations involve incompatible stress and/or strain fields, leading to hybrid, enhanced strain, or mixed formulations which are largely free of locking effects. The comparison of the different approaches if performed at three levels: At the level of the strong formulation (shell theory), at the level of the weak formulation (variational principle), and at the level of the finite element discretization.

The remainder of the paper herein is organised as follows. Section 2 elaborates on different approaches for deriving shell models but more importantly illustrates the main differences of these approaches. Further, in Section 3 three methods for the incorporation of the drill degree of freedom for two-dimensional shell models are discussed. The relation of the drill RC formulations with 6 dof shell models is outlined. Then a brief discussion on the corresponding finite element approximations is provided in Section 4. Further, in Section 5 two alternative implementations of the drill RC are described. The various implementations are tested through a multitude of benchmark examples and the results are presented in Section 6. Closing, a discussion based on the numerical results is provided in Section 7.

2 Shell Theories

One of the first fruitful attempts to derive a consistent theory of thin shells is due to Koiter [20]. Since then, tremendous efforts have been devoted to this topic. A preliminary conclusion of this development has been given in the monograph of Libai and Simmonds [21]. During the past years the number of new contribution has decreased significantly. Recent trends rather deal with smart shell structures [6], micropolar theories [1], and thermodynamics [13].

One way of deriving various shell mathematical models is through the use of 3D Continuum mechanics. Here the two most commonly applied categories of shell theories are considered: The first one can be traced back to Green and Zerna [15] who gave the first general linear theory of shells applying tensor calculus. Its non-linear counterpart has been derived by Simo et al. [32], [33] , [34]. The theory employs symmetric membrane and bending strain measures derived from the Green-Lagrange strain tensor. Recently, this theory has been revisited by Wagner and Gruttmann [36] to obtain a high-end finite element implementation. Commonly this approach is attributed as *geometrically exact shell theory*. An alternative point of departure has been provided by Reissner [26]. It has been recast by Ibrahimbegovic [17]. Via the application of a non-symmetric Biot-type strain measures the drilling dof enters the virtual work principle. Here these type of shell theories are called *micropolar*. It is emphasized that the micro-rotations enter at the level of dimensionaly reduced shell theory rather than at the level of 3D continuum mechanics.

A third type of shell theories is due to Simmonds [28] and has been elaborated by Chro-

scieliewsky et. al. [9]. Unlike in the other approaches, no kinematic assumptions (such as the Reissner-Mindlin assumption) are considered. It leads to almost the same expressions as the aforementioned approach, except for the appearance of drill moments, non-symmetric bending moments, and—in principle—to a non-local constitutive relation. Since the shell balance equations are exact consequences of the principle of virtual work, this type of theory is sometimes called *statically exact*. All approximations are transferred to the constitutive relation.

2.1 Description of the Shell Geometry and Kinematics

The position vector of any material point of the undeformed shell body $\mathcal{B}_0$ is written as

$$\mathbf{X}(\theta^i) = \mathbf{\Phi}(\theta^\alpha) + \theta \mathbf{D}(\theta^\alpha), \tag{1}$$

where $|\mathbf{D}(\theta^\alpha)| = 1$ and $h_- \leq \theta \leq h_+$ with $h = h_+ - h_-$ being the shell thickness. Further, $\mathbf{\Phi}(\theta^\alpha)$ denotes the position vector of the shell mid surface and (θ^i) is a curvilinear co-ordinate system parametrizing the shell body. Greek indices $\alpha, \beta, \ldots$ always range from 1 to 2. Latin indices $i, j, \ldots$ range from 1 to 3. Einstein's summation convention is used in the sequel. Introducing a kinematic assumption according to the well-known Reissner-Mindlin plate theory, the position vector of the material point in the deformed configuration $\mathcal{B}_\kappa$ is,

$$\mathbf{x}(\theta^i) = \boldsymbol{\phi}(\theta^\alpha) + \theta \mathbf{d}(\theta^\alpha), \tag{2}$$

where $|\mathbf{d}(\theta^\alpha)| = |\mathbf{D}(\theta^\alpha)| = 1$.

The covariant tangent vectors in the reference configuration read,

$$\mathbf{G}_\alpha = \mathbf{X}_{,\alpha} = \mathbf{\Phi}_{,\alpha} + \theta \mathbf{D}_{,\alpha}, \qquad \mathbf{G}_3 = \mathbf{X}_{,3} = \mathbf{D} \tag{3}$$

where $(\cdot)_{,\alpha} = \partial(\cdot)/\partial\theta^\alpha$ denotes the partial derivative. According to Eq. (2), the basis vectors of the current configuration are

$$\mathbf{g}_\alpha = \mathbf{x}_{,\alpha} = \boldsymbol{\phi}_\alpha + \theta \mathbf{d}_{,\alpha} \qquad \mathbf{g}_3 = \mathbf{x}_{,3} = \mathbf{d} \tag{4}$$

The corresponding basis vectors at the reference surface are denoted by $\mathbf{A}_\alpha = \mathbf{\Phi}_{,\alpha}$ and $\mathbf{a}_\alpha = \boldsymbol{\phi}_{,\alpha}$, respectively. Further, the contravariant basis vectors $\mathbf{G}^i$, $\mathbf{g}^i$, $\mathbf{A}^i$, and $\mathbf{a}^i$ are defined via $\mathbf{G}^i \cdot \mathbf{G}_j = \delta^i_j$, $\mathbf{A}^\alpha \cdot \mathbf{A}_\beta = \delta^\alpha_\beta$, etc. with δ^i_j being the Kronecker delta. The area and volume element in the undeformed configuration are

$$dA_0 = \sqrt{A}\, d\theta^1 d\theta^2, \qquad dV_0 = \sqrt{G}\, d\theta^1 d\theta^2 d\theta, \tag{5}$$

where $\sqrt{A} = \|\mathbf{A}_1 \times \mathbf{A}_2\|$ and $\sqrt{G} = \|\mathbf{G}_1 \times \mathbf{G}_2\|$. The elements dA and dV for the deformed configuration are defined analogously. Spatial and surface base vectors are related via $\mathbf{G}_\alpha = Z^\gamma_\alpha \mathbf{A}_\gamma$, where Z^γ_α are the components of the *shifter tensor* $\mathbf{Z} = Z^\gamma_\alpha \mathbf{Z}_\gamma \otimes \mathbf{A}^\alpha = \mathbf{G}_\alpha \otimes \mathbf{A}^\alpha$. Its determinant $Z = |Z^\gamma_\alpha| = \sqrt{G/A}$ is used to relate dA_0 and dV_0,

$$dV_0 = Z\, d\theta\, dA_0 \tag{6}$$

Analogously relations hold for the deformed configuration.

2.2 Shell Kinematics Based on Symmetric Green-Lagrange Strains

In the context of convective coordinates, the Green-Lagrange strain tensor reads

$$\mathbf{E} = E_{ij}\mathbf{G}^i \otimes \mathbf{G}^j, \quad E_{ij} = \frac{1}{2}(g_{ij} - G_{ij}), \tag{7}$$

Its components with respect to the basis $(\mathbf{G}^i)$ are

$$\begin{aligned} E_{\alpha\beta} &= \frac{1}{2}(\mathbf{g}_\alpha \cdot \mathbf{g}_\beta - \mathbf{G}_\alpha \cdot \mathbf{G}_\beta) = \varepsilon_{\alpha\beta} + \theta\kappa_{\alpha\beta} + \theta^2\rho_{\alpha\beta} \\ E_{\alpha 3} &= \frac{1}{2}(\mathbf{g}_\alpha \cdot \mathbf{g}_3 - \mathbf{G}_\alpha \cdot \mathbf{G}_3) = \gamma_\alpha \\ E_{33} &= \frac{1}{2}(\mathbf{g}_3 \cdot \mathbf{g}_3 - \mathbf{G}_3 \cdot \mathbf{G}_3) = 0 \end{aligned} \tag{8}$$

where

$$\begin{aligned} \varepsilon_{\alpha\beta} &= \frac{1}{2}(\mathbf{a}_\alpha \cdot \mathbf{a}_\beta - \mathbf{A}_\alpha \cdot \mathbf{A}_\beta) \\ \kappa_{\alpha\beta} &= \frac{1}{2}(\mathbf{a}_\alpha \cdot \mathbf{d}_{,\beta} + \mathbf{a}_\beta \cdot \mathbf{d}_{,\alpha} - \mathbf{A}_\alpha \cdot \mathbf{D}_{,\beta} - \mathbf{A}_\beta \cdot \mathbf{D}_{,\alpha}) \\ \gamma_\alpha &= \frac{1}{2}(\mathbf{a}_\alpha \cdot \mathbf{d} - \mathbf{A}_\alpha \cdot \mathbf{D}), \end{aligned} \tag{9}$$

are the membrane, bending, and transverse shear strains, respectively. It turns out that the second order strains $\rho_{\alpha\beta}$ do not enter the virtual work expression. (Note that this in *not* an approximation!) With the variation of the Green-Lagrange strain components

$$\delta E_{ij} = \frac{1}{2}(\delta\mathbf{g}_i \cdot \mathbf{g}_j + \mathbf{g}_i \cdot \delta\mathbf{g}_j) \tag{10}$$

the virtual strain energy for a deformable body reads

$$\delta W_{int} = \int_{\mathcal{B}_0} \mathbf{S} : \delta\mathbf{E}\, dV_0 \tag{11}$$

In terms of convective coordinates, the 2nd Piola-Kirchoff (PK2) stress tensor is given by

$$\mathbf{S} = j\sigma^{ij}\mathbf{G}_i \otimes \mathbf{G}_j \tag{12}$$

where $j = dV/dV_0 = \sqrt{g/G}$ and σ^{ij} are the components of the Cauchy stress tensor. Substituting the kinematic assumption Eq. (4) into Eq. (10) and the latter into Eq. (11) straight-forward calculations yield

$$\delta W_{int} = \int \left[\mathbf{n}^\alpha \cdot \delta\mathbf{a}_\alpha + \mathbf{m}^\alpha \cdot \delta\mathbf{d}_{,\alpha} + \mathbf{q} \cdot \delta\mathbf{d}\right] \sqrt{a}\, d\theta^1 d\theta^2 \tag{13}$$

Therein, the stress resultants

$$\mathbf{n}^\alpha = \int_{h_-}^{h_+} \sigma^{j\alpha}\mathbf{g}_j z\; d\theta, \quad \mathbf{m}^\alpha = \int_{h_-}^{h_+} \sigma^{j\alpha}\mathbf{g}_j \theta z\; d\theta, \quad \mathbf{q} = \int_{h_-}^{h_+} \sigma^{j3}\mathbf{g}_j z\; d\theta \tag{14}$$

have been defined. Again, applying Eq. (4), the membrane force vectors, bending moments, and transverse shear forces can be written as

$$\mathbf{n}^\alpha = n^{\beta\alpha}\mathbf{a}_\beta + q^\alpha \mathbf{d}, \qquad \mathbf{m}^\alpha = m^{\gamma\alpha}\mathbf{a}_\gamma + m^{3\alpha}\mathbf{d}, \qquad \mathbf{q} = \tilde{q}^\alpha \mathbf{a}_\alpha + q^3\mathbf{d} \tag{15}$$

respectively, where $n^{\beta\alpha} = \widetilde{n}^{\beta\alpha} + b^\beta_\gamma m^{\gamma\alpha}$, $q^\alpha = \tilde{q}^\alpha + b^3_\gamma m^{\gamma\alpha}$, and

$$\widetilde{n}^{\beta\alpha} = \int_{h_-}^{h_+} \sigma^{\beta\alpha} z \, d\theta, \qquad m^{\gamma\alpha} = \int_{h_-}^{h_+} \sigma^{\gamma\alpha}\theta z \, d\theta, \qquad \tilde{q}^\alpha = \int_{h_-}^{h_+} \sigma^{\alpha 3} z \, d\theta \tag{16}$$

Note that the effective membrane force tensor $\widetilde{n}^{\beta\alpha}$ as well as the bending moment tensor $m^{\beta\alpha}$ inherit their symmetry from the Cauchy stress tensor $\sigma^{\beta\alpha}$. Concluding, the virtual work of the internal forces reads

$$\delta W_{int} = \int \left[\delta\varepsilon_{\alpha\beta}\widetilde{n}^{\alpha\beta} + \delta\kappa_{\alpha\beta}m^{\alpha\beta} + \delta\gamma_\alpha\tilde{q}^\alpha\right]\sqrt{a}\, d\theta^1 d\theta^2 \tag{17}$$

Therein, the transverse normal force q^3 and the first order shear force $m^{3\alpha}$,

$$q^3 = \int_{h_-}^{h_+} \sigma^{33}\, z \, d\theta\,, \qquad m^{3\alpha} = \int_{h_-}^{h_+} \sigma^{3\alpha}\,\theta\, z \, d\theta \tag{18}$$

have dropped out. The first on due to $\delta E_{33} = 0$ and the second one due to $\mathbf{d}_{,\alpha}\cdot\mathbf{d} = 0$.

Confining the consideration to the simplest case, a Saint-Venant-Kirchhoff material law is applied. Accordingly, the constitutive equations read

$$\widetilde{n}^{ab} = Q^{abcd}\varepsilon_{cd}, \quad m^{ab} = \frac{t^3}{12}Q^{abcd}\kappa_{cd}, \quad \tilde{q}^a = t\,\alpha_s\,\mu\,\gamma_a \tag{19}$$

With Young's modulus E and Poisson ration ν, the plane-stress elasticity tensor and shear modulus read

$$Q^{abcd} = \frac{tE}{1-\nu^2}\left[\nu\delta^{ab}\delta^{cd} + (1-\nu)\delta^{ac}\delta^{bd}\right], \qquad \mu = \frac{E}{2\,(1+\nu)} \tag{20}$$

respectively. Finally, $\alpha_s = 5/6$ is the well-known shear correction factor.

2.3 Shell Kinematics Based on the Deformation Gradient

An alternative to the above derivations is due to [29], [26], [17]. Although this theory has been derived originally in a different way, here it is deduced analogously to the previous section starting from the variation of the strain energy in terms of the 1st Piola-Kirchhoff (PK1) stress tensor

$$\mathbf{P} = j\sigma^{ij}\mathbf{g}_i \otimes \mathbf{G}_j, \tag{21}$$

That is,

$$\delta W_{int} = \int_{\mathcal{B}_0} \delta\mathbf{F} : \mathbf{P}\, dV_0 \tag{22}$$

Introducing the traction vectors,

$$\mathbf{p}^\alpha = \mathbf{P}\mathbf{G}^\alpha = j\sigma^{i\alpha}\mathbf{g}_i, \qquad \mathbf{p}^3 = \mathbf{P}\mathbf{T}_3 = j\sigma^{i3}\mathbf{g}_i \tag{23}$$

The PK1 stress tensor can be written as,

$$\mathbf{P} = \mathbf{p}^\alpha \otimes \mathbf{G}_\alpha + \mathbf{p}^3 \otimes \mathbf{D} \tag{24}$$

Analogously the variation of the deformation gradient reads

$$\delta\mathbf{F} = \delta\mathbf{g}_\alpha \otimes \mathbf{G}^\alpha + \delta\mathbf{d} \otimes \mathbf{D} \tag{25}$$

According to the Reissner-Mindlin kinematic assumption one can write

$$\mathbf{g}_\alpha = \boldsymbol{\phi}_{,\alpha} + \theta\,\mathbf{d}_{,\alpha}, \quad \mathbf{d} = \mathbf{Q}\mathbf{D} \tag{26}$$

where $\mathbf{Q}$ is a rotation tensor. Consequently,

$$\delta\mathbf{d} = \delta\mathbf{Q}\mathbf{Q}^T\mathbf{d} = \delta\boldsymbol{\vartheta} \times \mathbf{d} \tag{27}$$

Further, writing $\mathbf{D} = \mathbf{Q}_0\mathbf{e}_3$ with $\mathbf{e}_3$ being the global z direction one obtains

$$\mathbf{d}_{,\alpha} = \mathbf{Q}_{,a}\mathbf{D} + \mathbf{Q}\mathbf{D}_{,\alpha} = \mathbf{Q}_{,\alpha}\mathbf{Q}^T\mathbf{d} + \mathbf{Q}\mathbf{Q}_{0,\alpha}\mathbf{Q}_0^T\mathbf{Q}^T\mathbf{d} = (\boldsymbol{\omega}_\alpha + \boldsymbol{\omega}_{0\alpha}) \times \mathbf{d} \tag{28}$$

Note that $\delta\mathbf{Q}\mathbf{Q}^T$, $\mathbf{Q}_{,\alpha}\mathbf{Q}^T$, and $\mathbf{Q}_{0,\alpha}\mathbf{Q}_0^T$ are skew-symmetric tensors. The corresponding axial vectors are denoted by

$$\boldsymbol{\vartheta} = \mathrm{ad}^{-1}\left(\delta\mathbf{Q}\mathbf{Q}^T\right), \quad \boldsymbol{\omega}_\alpha = \mathrm{ad}^{-1}\left(\mathbf{Q}_{,\alpha}\mathbf{Q}^T\right), \quad \boldsymbol{\omega}_{0\alpha} = \mathrm{ad}^{-1}\left(\mathbf{Q}_{0,\alpha}\mathbf{Q}_0^T\right) \tag{29}$$

With the above definitions the variation of the natural basis vectors becomes,

$$\delta\mathbf{g}_\alpha = \delta\boldsymbol{\phi}_{,\alpha} + \theta(\delta\bar{\boldsymbol{\omega}}_\alpha \times \mathbf{d} + \bar{\boldsymbol{\omega}}_\alpha \times \delta\mathbf{d}) \tag{30}$$

From the permutability of the operations of variation and derivative the following expression can be derived,

$$\delta\boldsymbol{\vartheta}_{,\alpha} = \delta\bar{\boldsymbol{\omega}}_\alpha - \delta\boldsymbol{\vartheta} \times \bar{\boldsymbol{\omega}}_\alpha = \delta\boldsymbol{\omega}_\alpha - \delta\boldsymbol{\vartheta} \times \boldsymbol{\omega}_\alpha \tag{31}$$

Putting all together,

$$\delta\mathbf{F} : \mathbf{P} = \delta\mathbf{g}_\alpha \cdot \mathbf{p}^\alpha + \delta\mathbf{d} \cdot \mathbf{p}^3 = \delta\boldsymbol{\phi}_{,\alpha} \cdot \mathbf{p}^\alpha + \underline{\delta\mathbf{d} \cdot \mathbf{p}^3} + \theta(\delta\boldsymbol{\vartheta}_{,\alpha} \times \mathbf{d}) \cdot \mathbf{p}^\alpha \tag{32}$$

The underlined term in Eq. (32) can be reformulated applying the balance of angular momentum, i.e. the symmetry of the σ_{ij} or, equivalently,

$$\mathbf{p}^\alpha \times \mathbf{g}^\alpha + \mathbf{p}^3 \times \mathbf{a}_3 = \mathbf{0} \tag{33}$$

Doing so,

$$\delta\mathbf{F} : \mathbf{P} = (\delta\boldsymbol{\phi}_{,\alpha} - \delta\boldsymbol{\vartheta} \times \boldsymbol{\phi}_{,\alpha})\mathbf{p}^\alpha + \theta\delta\boldsymbol{\vartheta}_{,\alpha} \cdot (\mathbf{d} \times \mathbf{p}^\alpha) \tag{34}$$

Finally,

$$\delta W_{int} = \int \left[\delta\boldsymbol{\varepsilon}_\alpha\mathbf{n}^\alpha + \delta\boldsymbol{\kappa}_\alpha \cdot (\mathbf{d} \times \mathbf{m}^\alpha)\right] \sqrt{A}\, d\theta^1 d\theta^2 \tag{35}$$

where the virtual strains and curvatures vectors,

$$\delta\boldsymbol{\varepsilon}_\alpha = \delta\boldsymbol{\phi}_{,\alpha} - \delta\boldsymbol{\vartheta} \times \boldsymbol{\phi}_{,\alpha}, \qquad \delta\boldsymbol{\kappa}_\alpha = \delta\boldsymbol{\vartheta}_{,\alpha} = \delta\boldsymbol{\omega}_\alpha - \delta\boldsymbol{\vartheta} \times \boldsymbol{\omega}_\alpha \tag{36}$$

and the resultant traction vectors,

$$\mathbf{n}^\alpha = \int_{h_-}^{h_+} \mathbf{p}^\alpha Z \, d\theta = j^0 \int_{h_-}^{h_+} \sigma^{i\alpha} \mathbf{g}_i z \, d\theta \ , \tag{37}$$

$$\mathbf{m}^\alpha = \int_{h_-}^{h_+} \mathbf{p}^\alpha Z \, \theta \, d\theta = j^0 \int_{h_-}^{h_+} \sigma^{i\alpha} \mathbf{g}_i \, z \, \theta \, d\theta \tag{38}$$

have been introduced. Here, $j^0 = \sqrt{a/A}$ and $Z = \sqrt{G/A}$ are the Jacobian determinants of the reference and undeformed shell reference surface, respectively. With the components of the stress resultants,

$$N^{\beta\alpha} = j^0 \int_{h_-}^{h_+} \sigma^{\gamma\alpha} z_\gamma^\beta \, z \, d\theta \, , \qquad Q^\alpha = j^0 \int_{h_-}^{h_+} \sigma^{3\alpha} \, z \, d\theta \ , \tag{39}$$

$$M^{\beta\alpha} = j^0 \int_{h_-}^{h_+} \sigma^{\gamma\alpha} z_\gamma^\beta \, z \, \theta \, d\theta \, , \qquad M^{3\alpha} = j^0 \int_{h_-}^{h_+} \sigma^{3\alpha} z \, \theta \ d\theta \tag{40}$$

one has

$$\mathbf{n}^\alpha = N^{\beta\alpha} \mathbf{a}_\beta + Q^\alpha \mathbf{d} \ , \qquad \mathbf{m}^\alpha = M^{\beta\alpha} \mathbf{a}_\beta + M^{3\alpha} \mathbf{d} \tag{41}$$

Defining the components of the virtual strain vectors with respect to the convected basis,

$$\delta \boldsymbol{\varepsilon}_\alpha = \delta \varepsilon_{\beta\alpha} \mathbf{a}^\beta + \delta \gamma_\alpha \mathbf{d}, \quad \delta \boldsymbol{\kappa}_\alpha = \delta \kappa_{\beta\alpha} \mathbf{d} \times \mathbf{a}^\beta + \delta \kappa_{3\alpha} \mathbf{d} \tag{42}$$

the virtual strain energy reads,

$$\delta W_{int} = \int [\delta \varepsilon_{\beta\alpha} N^{\beta\alpha} + \delta \kappa_{\beta\alpha} M^{\beta\alpha} + \delta \gamma_\alpha Q^\alpha] \sqrt{A} \, d\theta^1 d\theta^2 \tag{43}$$

Apparently, $M^{3\alpha}$ and $\kappa_{3\alpha}$ do not appear in the above expression.

2.4 Shell Kinematics Based on a Non-symmetric Biot-type Strain

It is instructive to consider a backward rotation to the undeformed configuration. Therefore two Cartesian coordinate systems are introduced; $\mathbf{T}_i = \mathbf{Q}_0 \mathbf{e}_i$ and $\mathbf{t}_i = \mathbf{Q}\mathbf{T}_i$ related to the undeformed and deformed configurations respectively. Further a Biot-type stress tensor is defined,

$$\mathbf{B} = \mathbf{Q}^T \mathbf{P} = B^{ij} \mathbf{T}_i \otimes \mathbf{T}_j \tag{44}$$

Accordingly the backward rotated resultant traction vectors are used

$$\mathbf{N}^a = \mathbf{Q}^T \mathbf{n}^a \quad \mathbf{M}^a = \mathbf{Q}^T \mathbf{m}^a \tag{45}$$

In this text, indices $a, b, \cdots$ generally denote Components with respect to a Cartesian basis and take the values 1 and 2 (to be distinguished from $\alpha, \beta, \ldots$ which denote components with respect to a curvilinear (natural) basis. The corresponding strain and curvature vectors are,

$$\mathbf{E}_a = \mathbf{Q}^T \boldsymbol{\phi}_{,a} - \mathbf{T}_a \ , \qquad \mathbf{K}_a = \mathbf{Q}^T \boldsymbol{\kappa}_a \tag{46}$$

It can be easily determined that

$$\delta \mathbf{E}_a = \mathbf{Q}^T \delta \boldsymbol{\varepsilon}_a \ , \qquad \delta \mathbf{K}^a = \mathbf{Q}^T \delta \boldsymbol{\kappa}_a \tag{47}$$

In view of the above the work of internal forces can be written as,

$$W_{int} = \int [\, \delta\mathbf{E}_a \cdot \mathbf{N}^a + \delta\mathbf{K}^a \cdot (\mathbf{D} \times \mathbf{M}^a) \,] \sqrt{A}\, d\theta^1 d\theta^2 \tag{48}$$

Introducing the strain and curvature components,

$$E_{ba} = \mathbf{T}_b \cdot \mathbf{E}_a = \mathbf{t}_b \cdot \boldsymbol{\phi}_{,a} - \delta_{ab}, \qquad E_{3a} = \mathbf{T}_3 \cdot \mathbf{E}_a = \mathbf{t}_3 \cdot \boldsymbol{\phi}_{,a} \tag{49}$$

$$K_{ba} = (\mathbf{T}_3 \times \mathbf{T}_b) = (\boldsymbol{\omega}_a \times \mathbf{t}_3) \cdot \mathbf{t}_b, \quad K_{3a} = (\mathbf{T}_3 \times \mathbf{K}_a) = \mathbf{t}_3 \cdot \boldsymbol{\omega}_a \tag{50}$$

the strain energy becomes

$$W_{int} = \int (E_{ba} N^{ba} + K_{ba} M^{ba} + E_{3a} Q^a) \sqrt{A}\, d\theta^1 d\theta^2 \tag{51}$$

Again, applying the plane-stress stiffness tensor,

$$Q^{abcd} = \frac{tE}{1-\nu^2} \left[\nu \delta^{ab} \delta^{cd} + (1-\nu) \delta^{ac} \delta^{bd} \right] \tag{52}$$

the constitutive relations are postulated according to

$$N^{(ab)} = Q^{abcd} E_{(cd)}, \quad N^{[ab]} = t\mu E_{[ab]} \tag{53}$$

$$M^{(ab)} = \frac{t^3}{12} Q^{abcd} K_{(cd)}, \quad M^{[ab]} = 0 \tag{54}$$

$$Q^a = t\, \alpha_s\, \mu\, E_{3a} \tag{55}$$

As usual, $(\cdot)_{(ab)}$ and $(\cdot)_{[ab]}$ denote symmetrization and anti-symmetrization, respectively.

2.5 Statically exact shell model

The shell theories described so far are 'non-exact' in the sense that they rely on a kinematic assumption, namely Eq. (4). It leads to the fact that higher-order moments do not enter the virtual strain expression. According to an alternative approach due to [26], the stress resultants are defined as given in Eq. (38). The shell balance equation are then *exact* consequences of the three-dimensional balance equations. By postulating the existence of a strain energy

$$\delta W_{int} = \int [\delta\varepsilon_{\alpha\beta} N^{\beta\alpha} + \delta\kappa_{\beta\alpha} M^{\beta\alpha} + \delta\gamma_\alpha Q^\alpha + \delta\kappa_{3\alpha} M^{3\alpha}] \sqrt{A}\, d\theta^1 d\theta^2 \tag{56}$$

and applying the virtual work principle, the work-conjugate virtual strain measures are found again to be given by Eq. (36) and Eq. (42).

Due to the absence of kinematic restrictions, the $M^{3\alpha}$ enter the strain energy and $M^{\beta\alpha}$ is not necessarily symmetric. Therefore, a modified constitutive law is postulated [10],

$$N^{(ab)} = Q^{abcd} E_{(cd)}, \quad N^{[ab]} = t\mu E_{[ab]}, \tag{57}$$

$$M^{(ab)} = \frac{t^3}{12} Q^{abcd} K_{(cd)}, \quad M^{[ab]} = \frac{t^3}{12} \mu E_{[ab]} \tag{58}$$

$$Q^a = t\, \alpha_s\, \mu\, E_{3a}, \quad M^{3a} = \frac{t^3}{12}\, \beta_w\, \mu\, K_{3a} \tag{59}$$

$$\tag{60}$$

or, equivalantly,

$$N^{ab} = \bar{Q}^{abcd} E_{cd}, \qquad M^{ab} = \frac{t^3}{12}\, \bar{Q}^{abcd} K_{cd} \tag{61}$$

Both, α_s and β_w play the role of shear correction factors.

3 Methods for defining the drill rotation

In the turn of shell finite element formulations, two types of rotational quantities appear. The rotation tensor $\mathbf{R}$ represents the rigid body rotation according to the translational deformation field. Based on polar decomposition theorem, $\mathbf{F} = \mathbf{R}\mathbf{U}$ where $\mathbf{F}$ is the deformation gradient and $\mathbf{U}$ the right stretch tensor. In addition, it can be advantageous to introduce an *a priori* independent rotation tensor field $\mathbf{Q}$. When dealing with non polar materials, the two tensor fields $\mathbf{Q}$ and $\mathbf{R}$ are expected to coincide, $\mathbf{Q} = \mathbf{R}$. Since $\mathbf{U}$ is symmetric the following relation holds

$$\mathrm{skew}(\mathbf{Q}^T\mathbf{F}) = \mathbf{0} \qquad \text{or, equivalently,} \qquad \epsilon^{ijk}\mathbf{A}_j^T\mathbf{Q}^T\mathbf{F}\mathbf{A}_k = 0 \tag{62}$$

where the deformation gradient of the shell mid surface reads,

$$\mathbf{F} = \mathbf{a}_\alpha \otimes \mathbf{A}^\alpha + \mathbf{d} \otimes \mathbf{T}_3 + \mathcal{O}(\theta) \tag{63}$$

In Eq. (62), use has been made of the totally antisymmetric tensor ϵ^{ijk}. In the sequel three different approaches to deal with the indeterminacy of the drill rotation are presented. All of these methods provide a definition for the drill rotation through an appropriately defined constraint equation. Their differences as well as similarities are discussed.

3.1 The Approach of Fox & Simo [14]

Let the following modified quantities be defined as,

$$\widetilde{\mathbf{F}} = \widetilde{\mathbf{a}}_\alpha \otimes \mathbf{A}^\alpha + \mathbf{d} \otimes \mathbf{T}_3, \quad \widetilde{\mathbf{Q}} = \exp(\omega\widetilde{\mathbf{T}}_3) \tag{64}$$

where $\widetilde{\mathbf{T}}_3$ defines the skew-symmetric matrix and

$$\widetilde{\mathbf{a}}_\alpha = \widetilde{\mathbf{F}}\mathbf{A}_\alpha = \mathbf{a}_\alpha - (\mathbf{T}_3 \cdot \mathbf{a}_\alpha)\mathbf{T}_3 \tag{65}$$

is the projection of $\mathbf{a}_\beta$ onto the plane normal to the vector $\mathbf{T}_3$. In view of the above the drill rotation constraint equation can be written as follows,

$$\epsilon^{\alpha\beta}\mathbf{A}_\alpha\widetilde{\mathbf{Q}}^T\widetilde{\mathbf{F}}\mathbf{A}_\beta = \epsilon^{\alpha\beta}\mathbf{A}_\alpha\widetilde{\mathbf{Q}}^T\widetilde{\mathbf{a}}_\beta \stackrel{!}{=} 0 \tag{66}$$

where $\epsilon^{\alpha\beta}$ is the antisymmetric permutation symbol. With

$$\widetilde{\mathbf{u}}_{,\alpha} = \mathbf{u}_{,\alpha} - (\mathbf{T}_3 \cdot \mathbf{u}_{,\alpha})\mathbf{T}_3 = u^\beta|_\alpha\mathbf{A}_\beta \tag{67}$$

one obtains

$$\widetilde{\mathbf{a}}_\alpha = \mathbf{A}_\alpha + \widetilde{\mathbf{u}}_{,\alpha} = \mathbf{A}_\beta(\delta_\alpha^\beta + u^\beta|_\alpha) \tag{68}$$

Substituting Eq. (68) into Eq. (66),

$$\epsilon^{\alpha\beta}\mathbf{A}_\alpha\widetilde{\mathbf{Q}}^T\mathbf{A}_\gamma(\delta_\alpha^\gamma + u^\gamma|_\alpha) \stackrel{!}{=} 0 \tag{69}$$

The product $\mathbf{A}_\alpha\widetilde{\mathbf{Q}}^T\mathbf{A}_\gamma$ defines the components of the rotation tensor $\widetilde{Q}_{\gamma\alpha}$ which can be written as,

$$\widetilde{Q}_{\gamma\alpha} = A_{\alpha\gamma}\cos\omega + \sqrt{A}\,\epsilon_{\alpha\gamma}\sin\omega \tag{70}$$

Substituting Eq. (70) into Eq. (66) and after some simplifications the drill rotation constraint is derived to be,

$$(u_1|_2 - u_2|_1)\cos\omega + \sqrt{A}(2 + u^1|_1 + u^2|_2)\sin\omega \stackrel{!}{=} 0 \tag{71}$$

A drawback of the above described method is that the left hand side of Eq. (71) depends on the length of the tangent vectors. In the context of finite element implementation it therefore leads to an unacceptable mesh dependence. As a remedy, one should use the Cartesian basis vectors $\mathbf{T}_\alpha$ instead of the natural basis vectors $\mathbf{A}_\alpha$. The corresponding drill rotation constraint equation is then derived from

$$\epsilon^{ab}\mathbf{T}_a\widetilde{\mathbf{Q}}^T\widetilde{\mathbf{F}}\mathbf{T}_b = 0 \tag{72}$$

3.2 The Modification of Rebel [25]

In an attempt to alleviate the problem with the length dependencies of the covariant derivatives, a modification of the above method was proposed by Rebel [25]. Here, the rotation tensor $\mathbf{Q}$ is split into two parts; the first corresponds to out of plane rotation while the second to in-plane rotation. Thus one can write,

$$\mathbf{Q} = \mathbf{Q}_d\mathbf{\Lambda}, \qquad \mathbf{Q}_d = \exp(\omega\widetilde{\mathbf{d}}), \qquad \mathbf{d} = \mathbf{\Lambda}\mathbf{T}_3 \tag{73}$$

With the above the rotation constraint equation reads,

$$\epsilon^{\alpha\beta}\mathbf{A}_\alpha^T\mathbf{Q}^T\widetilde{\mathbf{F}}\mathbf{A}_\beta = \epsilon^{\alpha\beta}\mathbf{A}_\alpha^T\mathbf{\Lambda}^T\mathbf{Q}_d^T\widetilde{\mathbf{a}}_\beta \stackrel{!}{=} 0\ , \tag{74}$$

where

$$\widetilde{\mathbf{a}}_\beta = \mathbf{a}_\beta - (\mathbf{d}\cdot\mathbf{a}_\beta)\,\mathbf{d} \tag{75}$$

is the projection of $\mathbf{a}_\beta$ onto the plane normal to the director $\mathbf{d}$ and the modified deformation gradient $\widetilde{\mathbf{F}}$ is defined accordingly. Note that the rotated basis vectors $\mathbf{\Lambda}\mathbf{A}_\alpha$ lie in the same plane. In order to avoid the dependency on the length of the deformed basis, as discussed previously, a normalization is introduced. That is,

$$\epsilon^{\alpha\beta}\frac{\mathbf{A}_\alpha^T}{\|\mathbf{A}_\alpha^T\|}\mathbf{Q}^T\frac{\widetilde{\mathbf{a}}_\beta}{\|\widetilde{\mathbf{a}}_\beta\|} = 0 \tag{76}$$

Alternatively, to avoid this artificial normalization procedure it would be beneficial to use the Cartesian basis vectors $\mathbf{T}_a$ instead of $\mathbf{A}_\alpha$. Note that here the drill rotation is defined as the rotation about the director vector unlike the approach of [14] where the drill rotation is defined as the rotation around the direction normal to the tangent plane of the mid surface.

3.3 The approach of Wisniewski & Turska [38]

The third and last alternative to treat the drill rotation in the context of shells, is due to Wiesniwski & Turska [38]. Here the rotation tensor is split into two parts,

$$\mathbf{Q}_0 = \mathbf{Q}_t\mathbf{\Lambda}_0, \qquad \mathbf{Q}_t = \exp(\omega\widetilde{\mathbf{t}}_3), \qquad \mathbf{t}_3 = \mathbf{\Lambda}_0\mathbf{T}_3 \tag{77}$$

where $\mathbf{Q}_t$ corresponds to the drill rotation and $\mathbf{\Lambda}_0$ is the director orienting rotation. With the above definitions the rotation constraint equation becomes,

$$\epsilon^{ab}\mathbf{T}_a^T\mathbf{Q}_0^T\mathbf{F}_0\mathbf{T}_b^T = \epsilon^{ab}\mathbf{T}_a^T\mathbf{\Lambda}_0^T\mathbf{Q}_t^T\boldsymbol{\phi}_{,b} = 0 \tag{78}$$

Since $\boldsymbol{\phi}_{,b}$ lies in the pane spanned by the vectors $\breve{\mathbf{T}}_a$, one can substitute in the above the following expression, $\boldsymbol{\phi}_{,b} = \breve{\mathbf{T}}_c^T\boldsymbol{\phi}_{,b}\breve{\mathbf{T}}_c$. Then,

$$\epsilon^{ab}\breve{\mathbf{T}}_a\mathbf{Q}_t^T\breve{\mathbf{T}}_c^T\boldsymbol{\phi}_{,b}\breve{\mathbf{T}}_c = \epsilon^{ab}Q_{ac}\boldsymbol{\phi}_{,b}\breve{\mathbf{T}}_c = 0 \tag{79}$$

Using $\mathbf{t}_\alpha = Q_{ac}\breve{\mathbf{T}}_c$ the above is written,

$$\epsilon^{ab}(\mathbf{t}_a \cdot \boldsymbol{\phi}_{,b}) = 0 \tag{80}$$

The above can be written with respect to the displacement gradients as follows,

$$(u_{1,2} - u_{2,1})\cos\omega + (2 + u_{1,1} + u_{2,2})\sin\omega = 0 \tag{81}$$

3.4 Comparison of RC with Micropolar Approach

All the above described methods attempt to provide a definition for the drill rotation in terms of a constraint equation. In the following the relation of the above methods with the definition of the drill rotation obtained from the shell theory in Subsection 2.4 is presented here.

$$\delta\mathbf{F} : \mathbf{P} = (\mathbf{Q}^T\delta\mathbf{F}) : (\mathbf{Q}^T\mathbf{P}) = \delta(\mathbf{Q}^T\mathbf{F}) : \mathbf{Q}^T\mathbf{P} - \underline{(\mathbf{Q}^T\mathbf{F}) : (\mathbf{Q}^T\mathbf{P})} \tag{82}$$

The underlined term vanishes since,

$$(\delta\mathbf{Q}^T\mathbf{F}) : (\mathbf{Q}^T\mathbf{P}) = \mathbf{F} : (\delta\mathbf{Q}\mathbf{Q}^T\mathbf{P}) = \mathrm{tr}(\delta\tilde{\boldsymbol{\vartheta}}\mathbf{P}\mathbf{F}^T) = 0 \tag{83}$$

With the help of the above and introducing the Biot-type strain tensor $\mathbf{H} = \mathbf{Q}^T\mathbf{F}$ Eq. (82) can be written as follows

$$\delta\mathbf{F} : \mathbf{P} = \delta\mathbf{H} : \mathbf{B} = \delta\mathbf{H}_s : \mathbf{B}_s + \delta\,\mathrm{skew}(\mathbf{Q}^T\mathbf{F}) : \mathbf{B}_a\,, \tag{84}$$

where the subscripts s, a denote the symmetric and antisymetric part. The integral of the anti-symmetric part is calculated to be,

$$\int \delta\,\mathrm{skew}(\mathbf{Q}^T\mathbf{F}) : \mathbf{B}\; dV = \int [\delta E_{[ab]}N_{[ab]} + \delta K_{[ab]}M_{[ab]} + \delta E_{3a}N_{3a}]\; dA \tag{85}$$

$$= \int [t\mu(\delta E_{[ab]}E_{[ab]}) + \frac{t^3}{12}\mu\delta K_{[ab]}K_{[ab]} + t\mu\delta E_{3a}E_{3a}]\; dA\,, \tag{86}$$

where

$$E_{[ab]} = \mathbf{t}_1 \cdot \boldsymbol{\phi}_{,2} - \mathbf{t}_2 \cdot \boldsymbol{\phi}_{,1} = (\mathbf{t}_1 \cdot \boldsymbol{\phi}_{,1} + \mathbf{t}_2 \cdot \boldsymbol{\phi}_{,2})\tan\omega_d \tag{87}$$

$$K_{[ab]} = \mathbf{t}_1 \cdot \mathbf{t}_{3,2} - \mathbf{t}_2 \cdot \mathbf{t}_{3,1} = (\mathbf{t}_1 \cdot \boldsymbol{\phi}_{,1} + \mathbf{t}_2 \cdot \boldsymbol{\phi}_{,2})\tan\omega_t \tag{88}$$

It turns out that this result corresponds to a linear dependence of the drill rotation on the thickness coordinate θ. That is [38],

$$\omega = \omega_d + \theta\omega_t, \tag{89}$$

where ω_t is the in-plane twist rotation and ω_d is the drill rotation at the reference surface. Enforcing the rotation constraint equation Eq. (80) according to Section 5, a direct equivalence with the first term in Eq. (86) can be established.

4 Finite Element Discretizations

4.1 Linear Flat 4-Node Shell (LFS4) Element

The position vector $\mathbf{\Phi}$ and director vector $\mathbf{D}$ on the flattened shell reference surface at the initial configuration can be interpolated as follows,

$$\mathbf{\Phi}^h = \sum_{I=1}^{4} N_I \mathbf{\Phi}_I, \quad \mathbf{D}^h = \sum_{I=1}^{4} N_I \mathbf{D}_I, \tag{90}$$

where $\mathbf{\Phi}_I$ are the position vectors of the nodes in the global coordinate system and $\mathbf{D}_I = \mathbf{a}_{3I}$ are the nodal shell directors which are perpendicular to the shell mid surface. Note that the superscript h denotes the finite element approximation and shall be used throughout this thesis. The bi-linear shape functions N_I are defined by,

$$N_I = \frac{1}{4}(1 + \xi_I^1 \xi^1)(1 + \xi_I^2 \xi^2), \tag{91}$$

where the ξ^α are the *natural coordinates* which belong to the bi-unit square domain spanned by the coordinates of the corner nodes $\{\xi_I^1, \xi_I^2\} = \{\mp 1, \mp 1\}$. For computational reasons, an additional orthogonal basis $(\mathbf{T}_i)$ is introduced at the element center. The length-measuring coordinates associated with the in-plane basis vectors $\mathbf{T}_a$ are denoted with s^a, where $a = 1, 2$. The derivatives of the interpolated position vector and director vectors with respect to s_a coordinates can be determined as follows,

$$\mathbf{\Phi}^h_{,a} = \sum_{I=1}^{4} N_{I,a} \mathbf{\Phi}_I, \quad \mathbf{D}^h_{,a} = \mathbf{0} \tag{92}$$

Analogous to Eq. (90), the displacements and rotations can be interpolated using the bi-linear interpolation functions as,

$$\mathbf{u}^h = \sum_{I=1}^{4} N_I \mathbf{u}_I, \quad \boldsymbol{\vartheta}^h = \sum_{I=1}^{4} N_I \boldsymbol{\vartheta}_I \tag{93}$$

where $\mathbf{u}_I = u_{Ik} \mathbf{e}_k$ is the nodal displacement vector and $\boldsymbol{\vartheta}$ the rotation vector. In order to avoid shear locking, the transverse shear components are interpolated according to the concept of Assumed Natural Strains (ANS) [12].

4.2 Enhancement of the membrane strains

The method of *Enhanced Assumed Strains* (EAS) was initiated by [35] for the linear case and by [30], [31] for the non linear case. The theory is based on the idea to add to the compatible strain field an incompatible enhanced strain field as,

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^h + \tilde{\boldsymbol{\varepsilon}}, \tag{94}$$

where the compatible strains are calculated by,

$$\boldsymbol{\varepsilon}^h = \sum_{I=1}^{4} \mathbf{B}_I \hat{\mathbf{u}}_I, \quad \hat{\mathbf{u}}_I = [\mathbf{u}_I, \boldsymbol{\vartheta}_I]^T, \tag{95}$$

with $\mathbf{B}$ being the strain-displacement matrix. In addition, the enhanced strains components are determined as [35],

$$\tilde{\boldsymbol{\varepsilon}} = \frac{j_0}{j} \mathbf{J}_0 \mathbf{M}\, \boldsymbol{\alpha}_e, \tag{96}$$

Here, the membrane strain components ε_{ab} are organized in a vector format,

$$\boldsymbol{\varepsilon} = \left[\varepsilon_{11}\ \varepsilon_{22}\ 2\varepsilon_{12}\right]^T \tag{97}$$

The 3×3 matrix $\mathbf{J}_0$ transforms the strain components from the skew basis $(\mathbf{A}^\alpha)$ at the element center to the Cartesian basis $(\mathbf{T}^a)$ according to

$$\tilde{\varepsilon}_{ab} = J_a{}^\alpha\, J_b{}^\beta\, \tilde{\varepsilon}_{\alpha\beta}, \quad \Leftrightarrow \quad \tilde{\boldsymbol{\varepsilon}}^{(T)} = \mathbf{J}_0\, \tilde{\boldsymbol{\varepsilon}}^{(A)} \tag{98}$$

where

$$\mathbf{J}_0 = \begin{bmatrix} (J_1{}^1)^2 & (J_1{}^2)^2 & J_1{}^1\, J_1{}^2 \\ (J_2{}^1)^2 & (J_2{}^2)^2 & J_2{}^1\, J_2{}^2 \\ 2\, J_1{}^1\, J_2{}^1 & 2\, J_1{}^2\, J_2{}^2 & J_1{}^1\, J_2{}^2 + J_2{}^1\, J_1{}^2 \end{bmatrix}, \qquad J_a{}^\alpha = \mathbf{T}_a \cdot \mathbf{A}^\alpha \tag{99}$$

The matrix $\mathbf{M}$ defines the enhanced strain modes. According to [2],

$$\mathbf{M}^{(7)} = \begin{bmatrix} \xi^1 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 \\ 0 & \xi^2 & 0 & 0 & 0 & \xi^1\xi^2 & 0 \\ 0 & 0 & \xi^1 & \xi^2 & 0 & 0 & \xi^1\xi^2 \end{bmatrix} \tag{100}$$

The amplitudes of these 7 modes are stored in the vector $\boldsymbol{\alpha}_e$.

In the case of the micropolar formulation, the strain tensor is nonsymmetric, thus

$$\boldsymbol{\varepsilon} = \left[\varepsilon_{11}\ \varepsilon_{22}\ \varepsilon_{12}\ \varepsilon_{21}\right]^T \tag{101}$$

For the transformation, a 4×4 matrix $\mathbf{J}_0^{4\times4}$ instead of the $\mathbf{J}_0$ has to be applied,

$$\mathbf{J}_0^{(4\times4)} = \begin{bmatrix} (J_1{}^1)^2 & (J_1{}^2)^2 & J_1{}^1\, J_1{}^2 & J_1{}^2\, J_1{}^1 \\ (J_2{}^1)^2 & (J_2{}^2)^2 & J_2{}^1\, J_2{}^2 & J_2{}^2\, J_2{}^1 \\ J_1{}^1\, J_2{}^1 & J_1{}^2\, J_2{}^2 & J_1{}^1\, J_2{}^2 & J_1{}^2\, J_2{}^1 \\ J_2{}^1\, J_1{}^1 & J_2{}^2\, J_1{}^2 & J_2{}^1\, J_1{}^2 & J_2{}^2\, J_1{}^1 \end{bmatrix} \tag{102}$$

The 8 strain modes are defined according to [10],

$$\mathbf{M}^{(8)} = \begin{bmatrix} \xi^1 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 & 0 \\ 0 & \xi^2 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 \\ 0 & 0 & \xi^1 & 0 & 0 & 0 & \xi^1\xi^2 & 0 \\ 0 & 0 & 0 & \xi^2 & 0 & 0 & 0 & \xi^1\xi^2 \end{bmatrix} \tag{103}$$

4.3 Enhancement of the drill rotation constraint

The implementations based on a splitting of the strain tensor in a symmetric and an antisymmetric part, $\varepsilon_{(ab)}$ and $\varepsilon_{[ab]}$, respectively, lead to the method of the *enhanced rotation constraint*. Therefore, the strain vector format

$$\tilde{\underline{\varepsilon}} = \left[\varepsilon_{11}\ \varepsilon_{22}\ 2\varepsilon_{(12)}\ 2\varepsilon_{[12]}\right]^T \tag{104}$$

is adopted. The correspoinding transformation matrix is denoted by $\bar{\mathbf{J}}_0^{(4\times4)}$. Two different sets of enhanced strain modes are considered,

$$\mathbf{M}^{(8_1)} = \begin{bmatrix} \xi^1 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 & 0 \\ 0 & \xi^2 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 \\ 0 & 0 & \xi^1 & \xi^2 & 0 & 0 & \xi^1\xi^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^1\xi^2 \end{bmatrix} \tag{105}$$

and

$$\mathbf{M}^{(8_2)} = \begin{bmatrix} \xi^1 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 & 0 \\ 0 & \xi^2 & 0 & 0 & 0 & \xi^1\xi^2 & 0 & 0 \\ 0 & 0 & \xi^1 & \xi^2 & 0 & 0 & \xi^1\xi^2 & 0 \\ 0 & 0 & \xi^1 & -\xi^2 & 0 & 0 & 0 & \xi^1\xi^2 \end{bmatrix} \tag{106}$$

leading to the implementations termed 'ERC$_1$' and 'ERC$_2$'. Finally, the method which turns out to perform best applies the enhanced strain modes according to Eq. (103) and a transformation of the kind

$$\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ 2\varepsilon_{(12)} \\ 2\varepsilon_{[12]} \end{bmatrix}^{(T)} = \bar{\bar{\mathbf{J}}}_0^{(4\times4)} \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{12} \\ \varepsilon_{21} \end{bmatrix}^{(A)} \tag{107}$$

The corresponding implementation is simply termed 'ERC'.

4.4 Variational formulation

Basis for the linear finite element formulation is the modified Hu-Washizu variational principle

$$\int \Big[t\, Q^{abcd} \left(\varepsilon_{ab} + \tilde{\varepsilon}_{ab}\right) \left(\delta\varepsilon_{cd} + \delta\tilde{\varepsilon}_{cd}\right) + \frac{t^3}{12} Q^{abcd} \kappa_{ab} \delta\kappa_{cd} + \alpha_s\, t\, \mu\, \gamma_a \delta\gamma_a + \beta\, t\, \mu\, c\, \delta c \Big] \sqrt{A}\, d\theta^1 d\theta^2 = \delta W_{ext} \tag{108}$$

where $c = E_{[12]}$ refers to the drill rotation constraint, Section 5, and δW_{ext} is the virtual work of the external forces.

For the micropolar formulation,

$$\int \Big[t\, \bar{Q}^{abcd} \left(E_{ab} + \tilde{\varepsilon}_{ab}\right) \left(\delta E_{cd} + \delta\tilde{\varepsilon}_{cd}\right) + \frac{t^3}{12} \bar{Q}^{abcd} K_{ab} \delta K_{cd} + \alpha_s\, t\, \mu\, E_{3a} \delta E_{3a} + \beta_w \frac{t^3}{12} \mu\, K_{3a} \delta K_{3a} \Big] \sqrt{A}\, d\theta^1 d\theta^2 = \delta W_{ext} \tag{109}$$

Finally, for the enhanced rotation constraint

$$\int \Big[t\, Q^{abcd} \left(E_{(ab)} + \tilde{\varepsilon}_{(ab)}\right) \left(\delta E_{(cd)} + \delta\tilde{\varepsilon}_{(cd)}\right) + \frac{t^3}{12} Q^{abcd} K_{(ab)} \delta K_{(cd)} + \beta\, t\, \mu \left(E_{[12]} + \tilde{\varepsilon}_{[12]}\right) \left(\delta E_{[12]} + \delta\tilde{\varepsilon}_{[12]}\right) + \beta_t \frac{t^3}{12} \mu\, K_{[12]} \delta K_{[12]} + \alpha_s\, t\, \mu\, E_{3a} \delta E_{3a} + \beta_w \frac{t^3}{12} \mu\, K_{3a} \delta K_{3a} \Big] \sqrt{A}\, d\theta^1 d\theta^2 = \delta W_{ext} \tag{110}$$

4.5 Allman Shape Functions

Within the context of Allman shape function the discretized displacement vector is modified as follows,

$$\mathbf{u} = \sum_{I=1}^{4} N_I \mathbf{u}_I + \sum_{H=5}^{8} N_H \Delta\mathbf{u}_H \,, \tag{111}$$

where the hierarchical shape functions N_H are defined as follows,

$$N_5 = \frac{1}{2}(1-(\xi^1)^2)(1-\xi^2), \qquad N_7 = \frac{1}{2}(1-(\xi^1)^2)(1+\xi^2) \tag{112}$$

$$N_6 = \frac{1}{2}(1-(\xi^2)^2)(1+\xi^1), \qquad N_8 = \frac{1}{2}(1-(\xi^2)^2)(1-\xi^2) \tag{113}$$

The hierarchical displacement vectors, $\Delta\mathbf{u}_H$, at the mid-points of the element boundaries are given by,

$$\Delta\mathbf{u}_H = -\frac{L_{JK}}{8}(\omega_K - \omega_J)\mathbf{n}_{JK}, \quad H = 5,6,7,8 \tag{114}$$

with $J = H-4$, $K = \mathrm{mod}(H,4)+1$ and L_{JK} is the length of the boundary JK and $\mathbf{n}_{JK}$ is the vector normal to the initial element boundary.

5 Implementation of drill RC equation

The linearisation of the drill RC equation reads,

$$c = \omega + \frac{1}{2}(u_{1,2} - u_{2,1}) = 0 \tag{115}$$

Eq. (115) can be formulated as an integral (i.e. weak) form and therefore be added in any variational principle. Using a penalty method the extended functional Π^* reads,

$$\Pi^*(\mathbf{u}, \boldsymbol{\sigma}, \mathbf{Q}) = \Pi(\mathbf{u}, \boldsymbol{\sigma}) + \Pi^P_{RC}, \tag{116}$$

where the penalty functional Π_{RC} is given as [37],

$$\Pi^P_{RC} = \int_{\mathcal{B}} F^P_{RC}(\mathbf{u}, \mathbf{Q})\, dV, \quad F^P_{RC}(\mathbf{u}, \mathbf{Q}) = \frac{\gamma}{2}c^2, \tag{117}$$

The linearisation of the penalty functional Π_{RC} yields to the following equations

$$\left(\mathbf{K} + \gamma \begin{bmatrix} \mathbf{K}^{RC}_{uu} & \mathbf{K}^{RC}_{u\vartheta} \\ \mathbf{K}^{RC}_{\vartheta u} & \mathbf{K}^{RC}_{\vartheta\vartheta} \end{bmatrix}\right) \begin{bmatrix} \Delta\mathbf{u} \\ \Delta\boldsymbol{\vartheta} \end{bmatrix} = - \begin{Bmatrix} \mathbf{r} + \gamma\mathbf{r}_u \\ \gamma\mathbf{r}_\vartheta \end{Bmatrix}, \tag{118}$$

where $\mathbf{r}_u, \mathbf{r}_\vartheta$ are the residuals. Here, $\mathbf{u} = [u^i_I] = [u^1_1\ u^2_1\ u^3_1\ \cdots\ u^1_4\ u^2_4\ u^3_4]$ is the element displacement vector and $\boldsymbol{\vartheta} = [\vartheta^i_I] = [\vartheta^1_1\ \vartheta^2_1\ \vartheta^3_1\ \cdots\ \vartheta^1_4\ \vartheta^2_4\ \vartheta^3_4]$ is the rotation vector while $\Delta\mathbf{u}, \Delta\boldsymbol{\vartheta}$ are the incremental vectors. Note that I, i denote the node number and component with respect to the global coordinate system $\mathbf{e}_i$ respectively. When $\mathbf{u} = \mathbf{0}$ and $\boldsymbol{\vartheta} = \mathbf{0}$, it follows that $\mathbf{r}_u = \mathbf{0}, \mathbf{r}_\vartheta = \mathbf{0}$ and $\mathbf{r} = -\mathbf{f}$. Hence Eq. (118) reduces to,

$$\left(\mathbf{K} + \gamma \begin{bmatrix} \mathbf{K}^{RC}_{uu} & \mathbf{K}^{RC}_{u\vartheta} \\ \mathbf{K}^{RC}_{\vartheta u} & \mathbf{K}^{RC}_{\vartheta\vartheta} \end{bmatrix}\right) \begin{bmatrix} \Delta\mathbf{u} \\ \Delta\boldsymbol{\vartheta} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}, \tag{119}$$

where $\mathbf{f}$ is the vector of external loads. Introducing the following notation,

$$\frac{\partial(\cdot)}{\partial \mathbf{u}} = \left[\frac{\partial(\cdot)}{\partial u_I^i}\right]_{I,i=1,2,3}, \qquad \frac{\partial(\cdot)}{\partial \boldsymbol{\vartheta}} = \left[\frac{\partial(\cdot)}{\partial \vartheta_I^i}\right]_{I,i=1,2,3} \tag{120}$$

the residuals of Eq. (118) are defined by,

$$\mathbf{r}_u = \frac{\partial \Pi_{RC}^P}{\partial \mathbf{u}}, \ \frac{\partial \Pi_{RC}^P}{\partial u_I^i} = \int_{\mathcal{B}} \frac{\partial F_{RC}^P}{\partial u_I^i}\, dV, \quad \mathbf{r}_\vartheta = \frac{\partial \Pi_{RC}^P}{\partial \boldsymbol{\vartheta}}, \ \frac{\partial \Pi_{RC}^P}{\partial \vartheta_I^i} = \int_{\mathcal{B}} \frac{\partial F_{RC}^P}{\partial \vartheta_I^i}\, dV \tag{121}$$

and the stiffness matrices,

$$\begin{aligned}
\boldsymbol{K}_{uu}^{RC} &= \frac{\partial \mathbf{r}_u}{\partial \mathbf{u}}, \ \frac{\partial \mathbf{r}_u}{\partial u_J^j} = \int_{\mathcal{B}} \frac{\partial^2 F_{RC}^P}{\partial u_J^j \partial u_I^i}\, dV, \quad \boldsymbol{K}_{u\vartheta}^{RC} = \frac{\partial \mathbf{r}_u}{\partial \boldsymbol{\vartheta}}, \ \frac{\partial \mathbf{r}_u}{\partial \vartheta_J^j} = \int_{\mathcal{B}} \frac{\partial^2 F_{RC}^P}{\partial \vartheta_J^j \partial u_I^i}\, dV \\
\boldsymbol{K}_{\vartheta u}^{RC} &= \frac{\partial \mathbf{r}_\vartheta}{\partial \mathbf{u}}, \ \frac{\partial \mathbf{r}_\vartheta}{\partial u_J^j} = \int_{\mathcal{B}} \frac{\partial^2 F_{RC}^P}{\partial u_J^j \partial \vartheta_I^i}\, dV, \quad \boldsymbol{K}_{\vartheta\vartheta}^{RC} = \frac{\partial \boldsymbol{r}_\vartheta}{\partial \boldsymbol{\vartheta}}, \ \frac{\partial \mathbf{r}_\vartheta}{\partial \vartheta_J^j} = \int_{\mathcal{B}} \frac{\partial^2 F_{RC}^P}{\partial \vartheta_J^j \partial \vartheta_I^i}\, dV
\end{aligned} \tag{122}$$

In the sequel two different methods for the implementation of the dril RC constraint in the context of four-node shell elements are presented. In the first the area integration is performed analytically while in the second a two point Gauss quadrature is employed.

5.1 Analytical Integration Based on Taylor Series Expansion

Using Taylor series expansion for Eq. (115) at the element centre (i.e. $\xi^1 = \xi^2 = 0$) yields,

$$\bar{c}(\xi^1, \xi^2) = c|_0 + c_{,\xi^1}|_0 \xi^1 + c_{,\xi^2}|_0 \xi^2, \tag{123}$$

where here $|_0$ denotes quantities evaluated at the element center. Displacements and drill rotations are interpolated using the bilinear shape functions $N_I(\xi^1, \xi^2)$. For the sake of completeness these relations are repeated here,

$$\mathbf{u} = \sum_{I=1}^{4} N_I(\xi^1, \xi^2)\mathbf{u}_I, \quad \boldsymbol{\vartheta} = \sum_{I=1}^{4} N_I(\xi^1, \xi^2)\boldsymbol{\vartheta}_I \tag{124}$$

The derivatives of the displacement components at Eq. (115) are determined as follows

$$u_{b,\alpha} = J_{\alpha\beta}^{-1} \frac{\partial u_b}{\partial \xi^\beta} = J_{\alpha\beta}^{-1} \sum_{I=1}^{4} N_{,\xi^\beta}(\mathbf{t}_b \cdot \mathbf{e}_i) u_I^i \ , \tag{125}$$

where u_I^i are the displacement components i of the node I with respect to the global coordinate system $(\mathbf{e}_i)$. Substituting Eq. (124) and Eq. (125) into Eq. (115) with $\omega_I = \mathbf{t}_3 \cdot \boldsymbol{\vartheta}_I$ yields,

$$c(\xi^1, \xi^2) = \sum_{I=1}^{4} N_I(\mathbf{t}_3 \cdot \boldsymbol{\vartheta}_I) + \frac{1}{2}\left(J_{2\beta}^{-1} \sum_{I=1}^{4} N_{,\xi^\beta}(\mathbf{t}_1 \cdot \mathbf{e}_i) u_I^i - J_{1\beta}^{-1} \sum_{I=1}^{4} N_{,\xi^\beta}(\mathbf{t}_2 \cdot \mathbf{e}_i) u_I^i \right) \tag{126}$$

With the help of Eq. (126), the three terms of Eq. (123) can be determined as follows,

$$
\begin{aligned}
c|_0 &= \frac{1}{4}\sum_{I=1}^{4}(\mathbf{t}_3\cdot\boldsymbol{\vartheta}_I) + \frac{1}{8}\left(J_{2\beta}^{-1}\sum_{I=1}^{4}\xi_I^{\beta}(\mathbf{t}_1\cdot\mathbf{e}_i)u_I^i - J_{1\beta}^{-1}\sum_{I=1}^{4}\xi_I^{\beta}(\mathbf{t}_2\cdot\mathbf{e}_i)u_I^i\right)\\
c_{,\xi^1}|_0 &= \frac{1}{4}\sum_{I=1}^{4}\xi_I^1(\mathbf{t}_3\cdot\boldsymbol{\vartheta}_I) + \frac{1}{8}\left(J_{22}^{-1}\sum_{I=1}^{4}\xi_I^1\xi_I^2(\mathbf{t}_1\cdot\mathbf{e}_i)u_I^i - J_{12}^{-1}\sum_{I=1}^{4}\xi_I^1\xi_I^2(\mathbf{t}_2\cdot\mathbf{e}_i)u_I^i\right)\\
c_{,\xi^2}|_0 &= \frac{1}{4}\sum_{I=1}^{4}\xi_I^2(\mathbf{t}_3\cdot\boldsymbol{\vartheta}_I) + \frac{1}{8}\left(J_{21}^{-1}\sum_{I=1}^{4}\xi_I^1\xi_I^2(\mathbf{t}_1\cdot\mathbf{e}_i)u_I^i - J_{11}^{-1}\sum_{I=1}^{4}\xi_I^1\xi_I^2(\mathbf{t}_2\cdot\mathbf{e}_i)u_I^i\right)
\end{aligned}
\tag{127}
$$

5.1.1 Weak Formulation and Linearisation

Since F_{RC}^P is a polynomial of the natural coordinates ξ^1, ξ^2 the integral of Eq. (117) can be determined analytically. That is,

$$
\Pi_{RC}^P = \int_{h_-}^{h_+}\int_{\square} F_{RC}^P\sqrt{a}\; d\xi^1 d\xi^2\; dh = \int_{h_-}^{h_+}\int_{\square}\bar{c}^2(\xi^1,\xi^2)\sqrt{a}\; d\xi^1 d\xi^2\; dh \approx\; \gamma V\tilde{\Pi}_{RC}^P, \tag{128}
$$

where V is the element volume and

$$
\widetilde{\Pi}_{RC}^P = [2c^0 + {}^2\!/_3((c_{,\xi^1}^0)^2 + (c_{,\xi^2}^0)^2)] \tag{129}
$$

Note that for the following derivations a constant Jacobian is assumed. The force residual vectors and the stiffness matrices in Eq. (121) and Eq. (122) respectively can be determined in a straight forward way. In order to be able to account for warped element geometries a correction has to be introduced. The correction here is introduced like in [16] where the local displacements and rotations are related to their global components via,

$$
\begin{bmatrix}\mathbf{u}_I^l\\ \boldsymbol{\vartheta}_I^l\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \widetilde{\mathbf{W}}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{u}_I^G\\ \boldsymbol{\vartheta}_I^G\end{bmatrix}, \tag{130}
$$

where $\widetilde{\mathbf{W}} = \mathrm{skew}(\mathbf{r}_I)$ is the skew-symmetric matrix is with $\mathbf{r}_I = w_I\mathbf{t}_3$ where w_I is the distance to the middle surface of each node. Analogously forces and moments are transformed as follows,

$$
\begin{bmatrix}\mathbf{f}_I^l\\ \mathbf{m}_I^l\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \widetilde{\mathbf{W}} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{f}_I^G\\ \mathbf{m}_I^G\end{bmatrix} \tag{131}
$$

The warped stiffness matrix of the element is given by,

$$
\mathbf{K} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\widetilde{\mathbf{W}} & \mathbf{I}\end{bmatrix}\mathbf{K}^G\begin{bmatrix}\mathbf{I} & \widetilde{\mathbf{W}}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \tag{132}
$$

5.2 Numerical Integration of Drill RC Equation

For the case where the area integration is performed numerically, the integrands of Eq. (121) can be determined with the help of Eq. (115) and Eq. (117) in a straight forward way. That is,

$$
\frac{\partial F_{RC}^P}{\partial u_I^i} = \gamma c\left[\frac{1}{2}\left(\frac{\partial u_{1,2}}{\partial u_I^i} - \frac{\partial u_{2,1}}{\partial u_I^i}\right)\right], \quad \frac{\partial F_{RC}^P}{\partial \vartheta_I^i} = \gamma\; c\frac{\partial c}{\partial \vartheta_I^i} = \gamma c N_I\, t_3^i \tag{133}
$$

where the derivatives in Eq. $(133)_1$ read,

$$\frac{\partial u_{1,2}}{\partial u_I^i} = N_{I,2}\,(\mathbf{t}_1 \cdot \mathbf{e}_i), \quad \frac{\partial u_{2,1}}{\partial u_I^i} = N_{I,1}\,(\mathbf{t}_2 \cdot \mathbf{e}_i) \tag{134}$$

The integrands of the force residual vectors and the stiffness matrices in Eq. (121) and Eq. (122) can be determined using a 2-point Gauss quadrature integration scheme. For all the above computations the elemental coordinate system $(\mathbf{t}_i)$ at the element center was used. Although for flat shell elements this is sufficient, for warped geometries the rotation constraint does not account for the element warpage. This problem can be addressed by creating new coordinate systems at every integration point. That is,

$$\mathbf{\Phi}_{,\xi^1}^{h(ip)} = \sum_{I=1}^{4} N_{,\xi^1}^{ip} \mathbf{\Phi}_I, \quad \mathbf{\Phi}_{,\xi^2}^{h(ip)} = \sum_{I=1}^{4} N_{I,\xi^2}^{ip} \mathbf{\Phi}_I, \tag{135}$$

where $\mathbf{\Phi}_I$ is the nodal position vector at the reference shell surface and the superscript ip denotes the integration point. With the above the normal direction is determined as follows,

$$\mathbf{t}_3^{ip} = \frac{\mathbf{\Phi}_{,\xi^1}^{ip} \times \mathbf{\Phi}_{,\xi^2}^{ip}}{\left\| \mathbf{\Phi}_{,\xi^1}^{ip} \times \mathbf{\Phi}_{,\xi^2}^{ip} \right\|} \tag{136}$$

and the plane normal is identified as,

$$\mathbf{t}_1^{ip} = \mathbf{t}_3^{ip} \times \mathbf{t}_2, \quad \mathbf{t}_2^{ip} = \mathbf{t}_3^{ip} \times \mathbf{t}_1^{ip} \tag{137}$$

As such the vectors $\mathbf{t}_i$ should be replaced with the vectors $\mathbf{t}_i^{ip}$ which are determined at every integration point.

6 Numerical Examples

In this section, the performance of the above elements is evaluated based on a set of linear benchmark examples. Besides the quality of the computed displacements, the corresponding stresses and their converge rates are examined. A list of the shell elements that will be considered in the following is given in Table 1. Note that all of the elements below are fully integrated (i.e. 2 point Gauss scheme) quadrilaterals.

Table 1: Elements

Elements	Description
LFS4-RCAI	Linear flat with analytically integrated drill RC, ($\beta = 0.01$).
LFS4-RCFI	Linear flat with numerically integrated drill RC, ($\beta = 0.01$).
LFS4-ERC	Linear flat with numerically integrated enhanced drill RC, ($\beta_t = \beta_w = 0$ and $\beta = 1$).
LFS4-MP	Linear flat based on non-symmetric Biot strain approach, ($\beta = \beta_t = 1$ and $\beta_w = 0.01$).
LFS4+Allman	Linear flat with Allman shape functions
LFS4-MP+Allman	Linear flat based on non-symmetric Biot strain approach with Allman shape functions, ($\beta = \beta_t = 1$ and $\beta_w = 0.01$).
LFS4-RCAI+Allman	Linear flat with analytically integrated drill RC ($\beta = 0.01$) and Allman shape functions
QMHS4	Non-Linear mixed/hybrid with 5/6 Dofs, [18]*
CQUAD4	MSC/NASTRAN
S4	ABAQUS

* Own Results with element of [18].

6.1 Cook's Membrane

This problem was proposed by [11] and it is a very popular benchmark example. It involves a trapezoidal membrane which is clamped on one edge and is subjected to a distributed in-plane bending load on the other, see Figure 1 where the geometrical and material properties are defined.

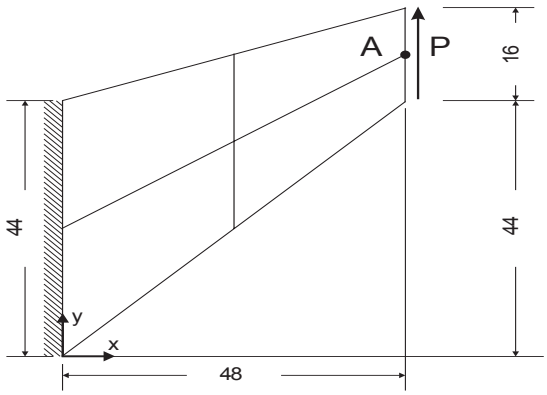


Figure 1: Cook's Membrane: $E = 1$, $\nu = 1/3$, $t = 1$, $P = 1$

In this example, a considerable amount of shear deformation is involved and therefore is excellent for testing the capability of the proposed elements to model membrane dominated situations. Moreover, when quadrilaterals are used for the discretization, the resulting elements are skewed and tapered which makes it even more important since the effects of mesh distortion can also be observed. The results for the vertical displacement and drill rotation at the point A, are listed in Table 2 and Table 3 respectively.

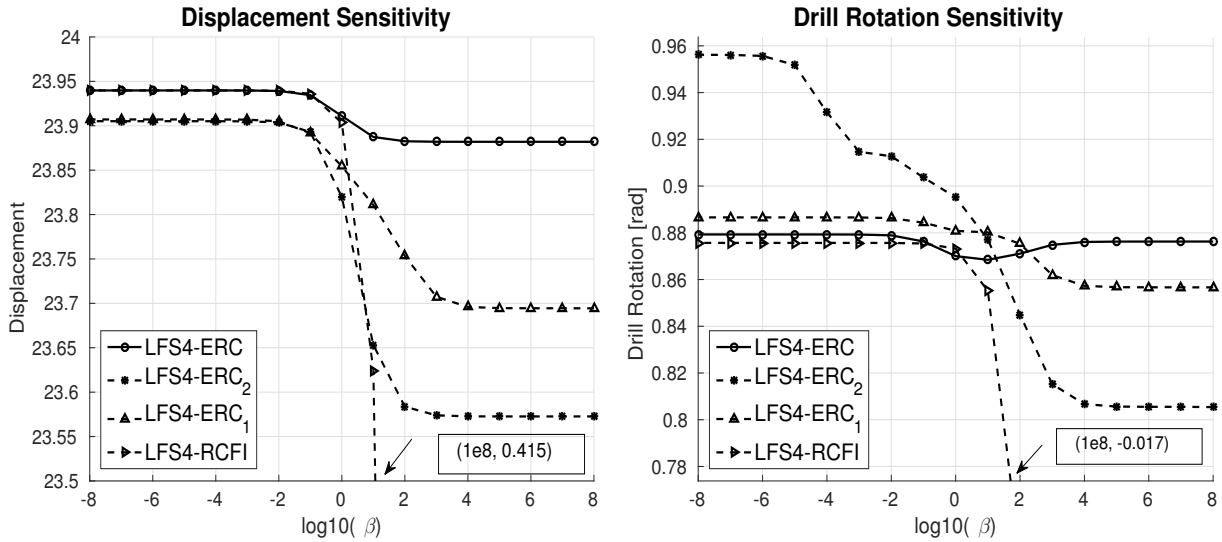
Elements	2×2	4×4	8×8	16×16	32×32
LFS4-RCAI	21.0936	23.0081	23.6850	23.8821	23.9394
LFS4-RCFI	21.1057	23.0135	23.6866	23.8825	23.9395
LFS4-ERC	20.4303	22.7144	23.5491	23.8207	23.9110
LFS4-MP	20.4300	22.7140	23.5486	23.8202	23.9106
LFS4-MP+Allman	22.1472	23.1203	23.6168	23.8296	23.9101
LFS4+Allman	21.2704	23.0663	23.6652	23.8661	23.9312
LFS4-RCAI+Allman	21.1911	23.0361	23.6485	23.8559	23.9260
QMHS4	21.1161	23.0206	23.6891	23.8833	23.9392
S4	20.7115	22.9701	23.6460	23.8588	23.9301
CQUAD4	21.0470	23.0151	23.6881	23.8827	23.9392
Ref.	23.91 [5], 23.936 [39]				

Table 2: Cook's Membrane: Vertical Displacement at point A.

Elements	2×2	4×4	8×8	16×16	32×32
LFS4-RCAI	0.7021	0.7786	0.8431	0.8658	0.8756
LFS4-RCFI	0.7649	0.8254	0.8710	0.8809	0.8838
LFS4-ERC	0.9518	0.7901	0.8464	0.8631	0.8700
LFS4-MP	0.9518	0.7904	0.8502	0.8829	0.8910
LFS4-MP+Allman	0.8662	0.8731	0.8768	0.8860	0.8873
LFS4+Allman	0.8485	0.8906	0.9102	0.9242	0.9268
LFS4-RCAI+Allman	0.8350	0.8842	0.9087	0.9240	0.9197
S4	0.7158	0.7972	0.8504	0.8708	0.8794
CQUAD4	0.7337	0.8102	0.8579	0.8742	0.8805

Table 3: Cook's Membrane: Drill Rotation at Point A.

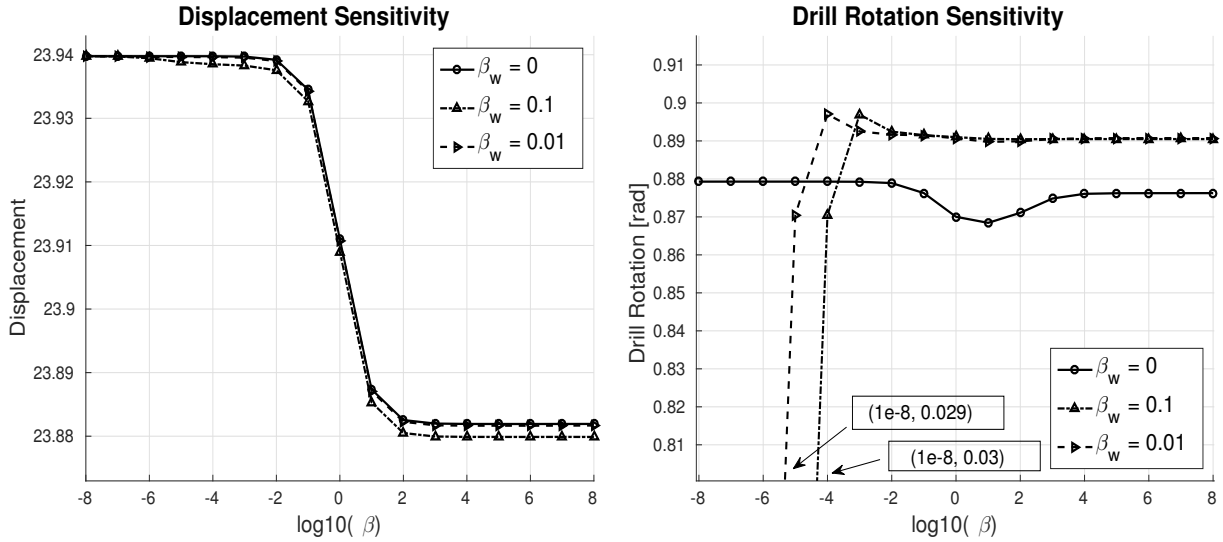
The sensitivity of the converged solution with respect to the penalty parameter β is shown in Figure 2. The performance of the different methods for the incorporation of the enhancement of strains as described in Section **??** is shown.



(a) Vertical Displacement (b) Drill Rotation

Figure 2: Cook's Membrane: Sensitivity of computed results with respect to penalty coefficient $\beta = \text{range}(10^8, 10^{-8})$.

It is obvious that the LFS4-ERC implementation, in which the enhanced strains are interpolated according to Eq. (103), has the best performance among all and therefore the implementations LFS4-ERC$_1$, LFS4-ERC$_2$ -using interpolations Eq. (**??**), Eq. (**??**) respectively- will not be considered further. In Figure 3, the effect of the warping parameter β_w is shown.



(a) Vertical Displacement (b) Drill Rotation

Figure 3: Cook's Membrane: Sensitivity of LFS4-ERC with respect to penalty parameter β for various warping parameters.

Note that since this is a plane problem the solution shall not depend on the twist parameter β_t which is the case. The convergence of the normal stress resultant $\mathbf{n}^{11}$ for various implementations is plotted in Figure 4. Note that the mesh sizes analysed for the stresses differ from the

ones analysed for the displacements. This is due to the fact that stresses are extracted at the centroid[1] and therefore for a meaningful comparison we had to make sure that the centroid of the element from which the stresses are extracted does not change with the mesh refinement. For this example the mesh sizes used are, 6×6, 10×10, 18×18, 34×34, 66×66, 130×130.

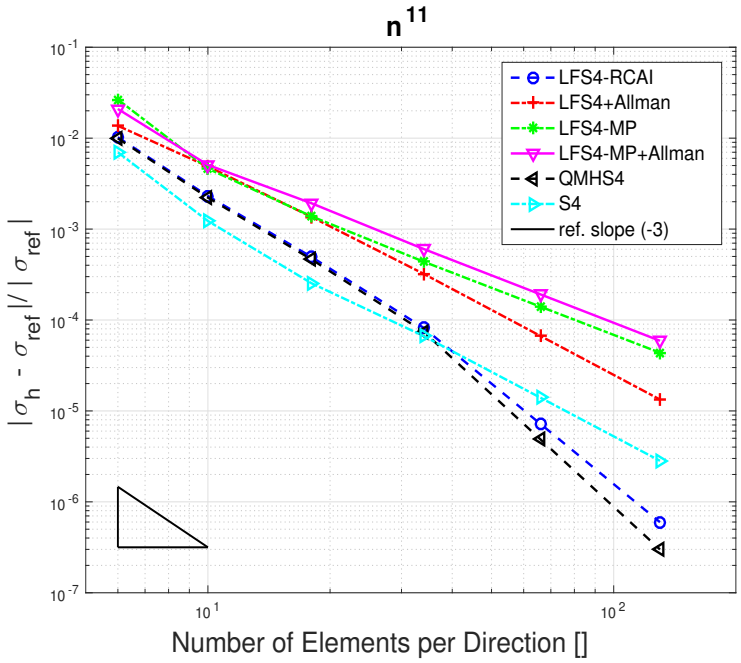


Figure 4: Cook's Membrane: Rate of convergence of stress resultant

The reference value, σ_{ref}, for the error estimation is obtained using Richardson's extrapolation, see [7] and also [8] where a remedy for oscillatory converge is presented. The extrapolated value is computed using the stresses as computed from the last three finer meshes.

6.2 Cantilever Under end Moment

This example involves a straight cantilever beam which is subjected to an in-plane end moment at one end while the other is fully clamped.

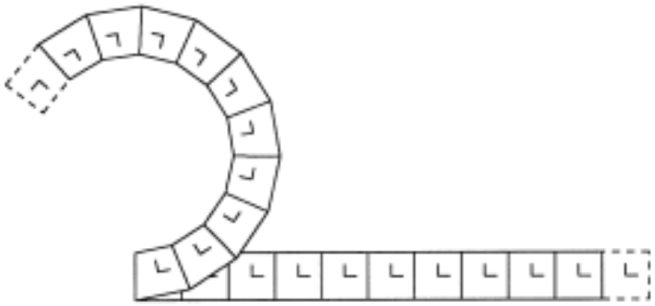


Figure 5: Cantilever under end moment: $E = 1200$, $l = 10$, $h = t = 1$, $\nu = 0.3$, $M = 0.1$

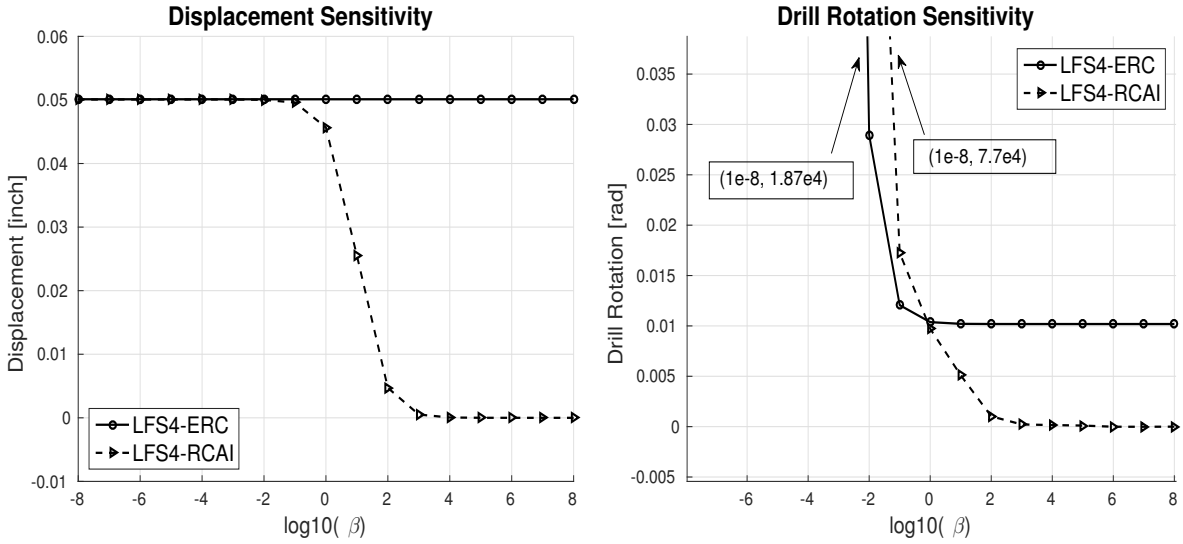
[1] Although the stresses are computed at the integrations points, here it was decided to make the stress comparison based on the centroidal values. For the case of four node quadrilaterals, it can be proved that the centroid is the Barlow point and therefore it is expected to have the highest convergence rate.

This is a very important example since the ability of the proposed element to carry drill moments can be tested. In Table 4, the vertical displacement and drill rotation of the tip of the cantilever are listed.

	10×1	
Elements	Vertical	Rotation
LFS4-RCAI	0.0457	0.0097
LFS4-RCFI	0.0457	0.0097
LFS4-ERC	0.0501	0.0103
LFS4-MP	0.0497	0.0103
LFS4-MP+Allman	0.0497	0.0103
LFS4+Allman	0.0455	0.0351
S4	0.0467	0.0121
Analytical	0.05	0.01

Table 4: Cantilever under tip moment: Vertical displacement and drill rotation at the tip of the beam.

Note that in the above the results corresponding to LFS4-RCAI, LFS4-RCFI implementations are obtained with a penalty parameter $\beta = 1$. As it is already reported in the literature, examples where drill moments are involved require a larger penalty parameter. The main advantage of the proposed implementation LFS4-ERC is that higher penalty parameters can be used without affecting the element performance in other examples. To illustrate this, the sensitivity of the computed values with respect to the penalty parameter is shown in Figure 6.

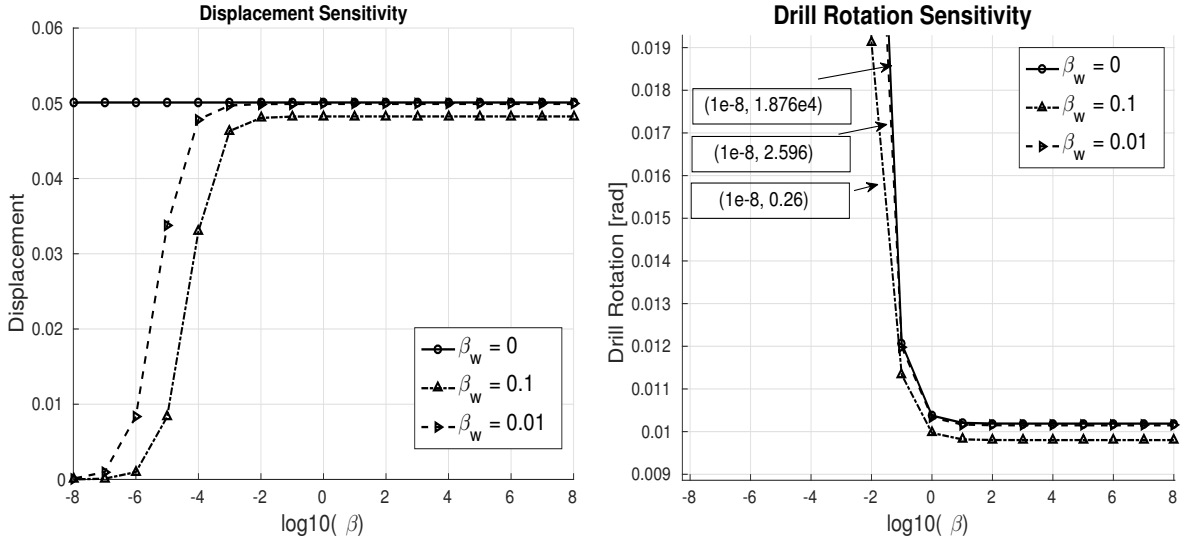


(a) Tip Displacement (b) Tip Drill Rotation

Figure 6: Cantilever under tip moment: Sensitivity of computed results with respect to penalty coefficient $\beta = \text{range}(10^8, 10^{-8})$

Clearly, the performance of the proposed element is superior to LFS4-RCAI which for larger penalty parameters becomes unacceptably stiff. Of course for small penalty parameters the drill

rotation for both formulations is unbounded. The effect of the warping parameter β_w on the computed solutions for the LFS4-ERC is shown in Figure 7.

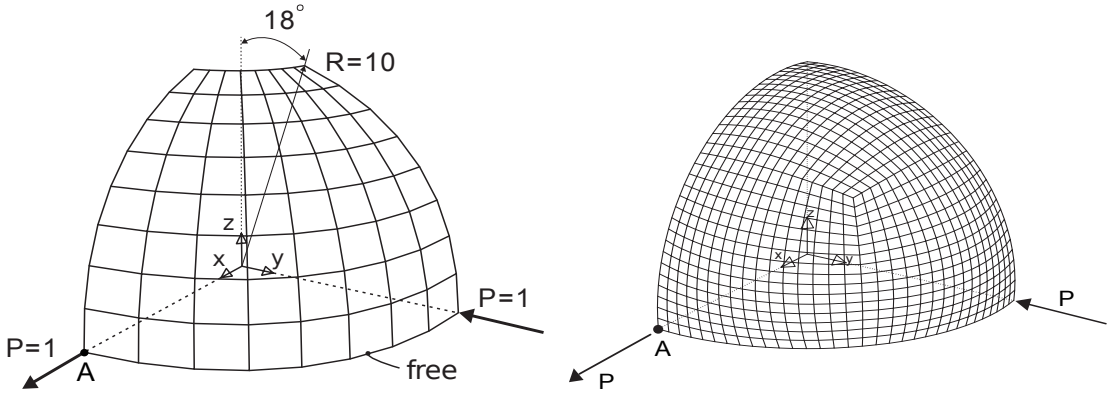


(a) Tip Displacement (b) Tip Drill Rotation

Figure 7: Cantilever under tip moment: Sensitivity of LFS4-ERC with respect to penalty parameter β and warping parameter β_w.

6.3 Pinched Hemispherical Shell

The third example involves a hemispherical shell, subjected to two inward and outward point loads $90°$ apart. Due to the double symmetry, only one quadrant can be modelled and symmetry boundary conditions can be applied. This example is popular for multiple reasons, firstly it involves a double curvature structure and secondly the shell undergoes an almost inextensional bending deformation.



(a) Pinched Hemispherical Shell with hole (b) Pinched Hemispherical Shell without hole

Figure 8: Pinched Hemispherical Shells. $E = 6.825 \times 10^7$, $\nu = 0.3$, $h = 0.04$, $P = 1$

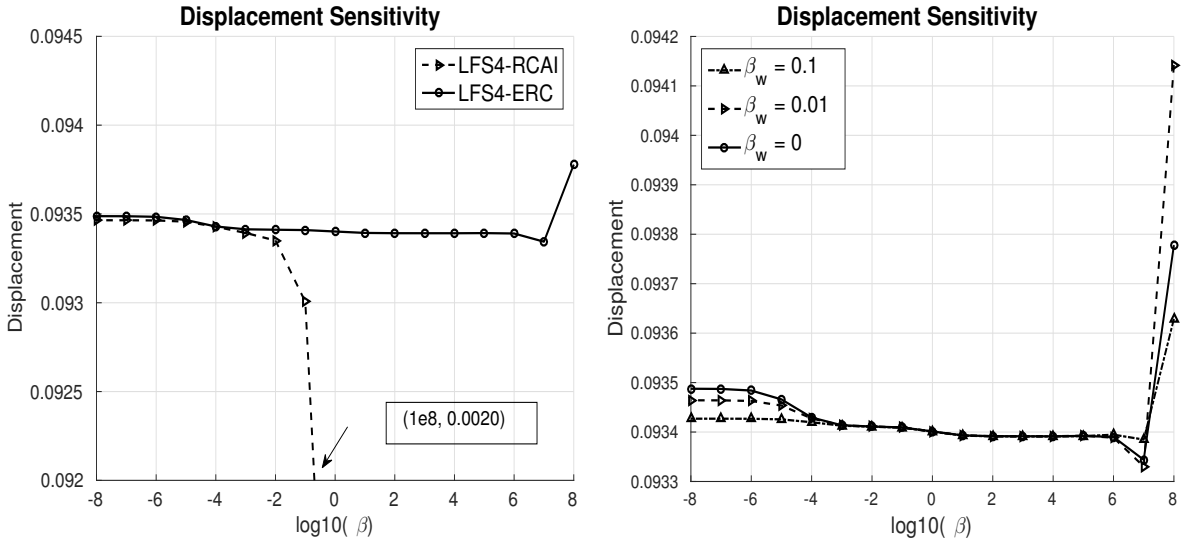
Two variants of this test exist in the literature. The first was proposed by [23] where a hole of $18°$ is introduced at the top in order to avoid the use of triangles near the axis of revolution,

see Figure 8. In this case, the resulted finite elements are planar and therefore the effect of warp is not present. The computed displacements for the point A are listed in Table 5. Note that the displacement of both loaded points should be identical for the case of linear static analysis due to symmetry.

Elements	4×4	8×8	16×16	32×32
LFS4-RCAI	9.0662	9.2295	9.2943	9.3351
LFS4-RCFI	8.6881	9.2457	9.3067	9.3387
LFS4-ERC	9.1400	9.3010	9.3137	9.3401
LFS4-MP	9.0603	9.2714	9.3043	9.3373
LFS4-MP+Allman	3.7017	8.9415	9.3086	9.3431
LFS4+Allman	3.4786	8.9246	9.3078	9.3427
LFS4-RCAI+Allman	3.4361	8.8675	9.2927	9.3388
QMHS4	9.6631	9.3781	9.3333	9.3454
S4	9.1966	9.2562	9.2997	9.3364
CQUAD4	9.5071	9.4478	9.3661	9.3574
Ref.		9.4 [23], 9.3 [33]		

Table 5: Pinched Hemispherical Shell with Hole: Displacement $-u_1 \times 100$

In Figure 9a, the sensitivity with respect to the penalty parameter β is plotted. Mooreover, in Figure 9b, the effect of the warping parameter for the LFS4-ERC implementation is shown.



(a) Comparison with LFS4-RCAI

(b) Effect of warping parameter on LFS4-ERC

Figure 9: Pinched Hemispherical Shell with Hole: Sensitivity of computed results with respect to penalty coefficient $\beta = \mathrm{range}(10^8, 10^{-8})$ and warping parameter β_w.

The advantage of the enhanced drill rotation constraint is apparent since even for doubly curved problems the computed solutions are almost insensitive to the penalty parameter. Unlike the previous examples, being all plane, here it is instructive to study the effect of the twist parameter β_t for the LFS4-ERC element. In Figure 10 the sensitivity of the computed displacement is plotted.

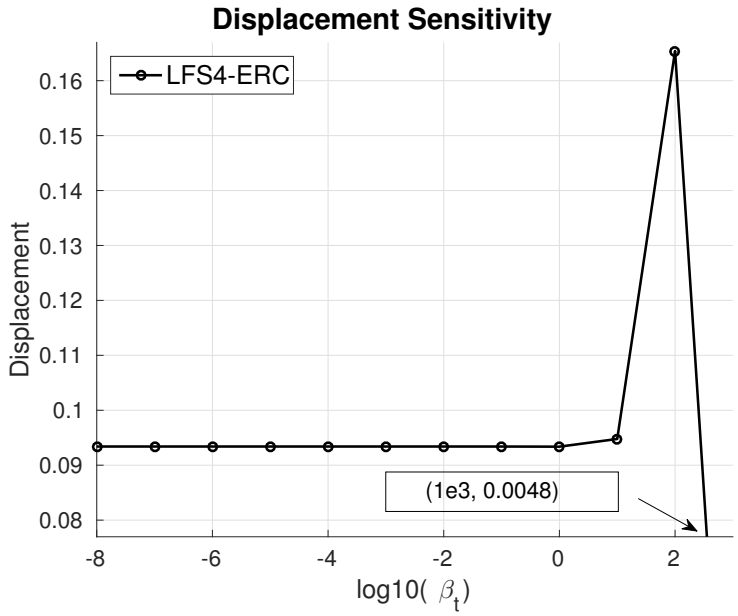


Figure 10: Pinched Hemispherical Shell with Hole: Sensitivity of LFS4-ERC with respect to the twist parameter, $\beta_t = \mathrm{range}(10^3, 10^{-8})$.

The above figure clearly dictates that the incorporation of the twist parameter impairs the solution while for $\beta_t < 1$ the result almost remains constant. The selection of $\beta_t = 0$ can be justified. The second variant of the pinched hemispherical problem, see Figure 8b does not involve a hole but the resulted elements are highly skewed and warped especially for coarse discretizations. The computed displacements for this case are listed in Table 6.

Elements	4×4	8×8	16×16	32×32
LFS4-RCAI	8.3951	8.9468	9.1254	9.2049
LFS4-RCFI	7.8903	9.0457	9.1698	9.2177
LFS4-ERC	8.7632	9.1741	9.1889	9.2219
LFS4-MP	8.6820	9.1428	9.1789	9.2188
LFS4-MP+Allman	3.0762	8.6649	9.1755	9.2270
LFS4+Allman	2.8801	8.6630	9.1749	9.2261
LFS4-RCAI+Allman	2.7853	8.4383	9.1101	9.2095
QMHS4	5.2690	9.0300	9.1983	9.2264
S4	4.7672	8.8644	9.1628	9.2193
CQUAD4	9.0688	9.3113	9.2479	9.2425
Ref.	9.24 [33]			

Table 6: Pinched Hemispherical Shell without hole: Displacement $-u_1 \times 100$

Closing this example, the convergence of the torsional moment for the hemispherical shell with the hole is shown in Figure 11. For this example the mesh sizes used are, 10×10, 18×18, 34×34, 66×66 and correspond to the one quadrant of the hemisphere.

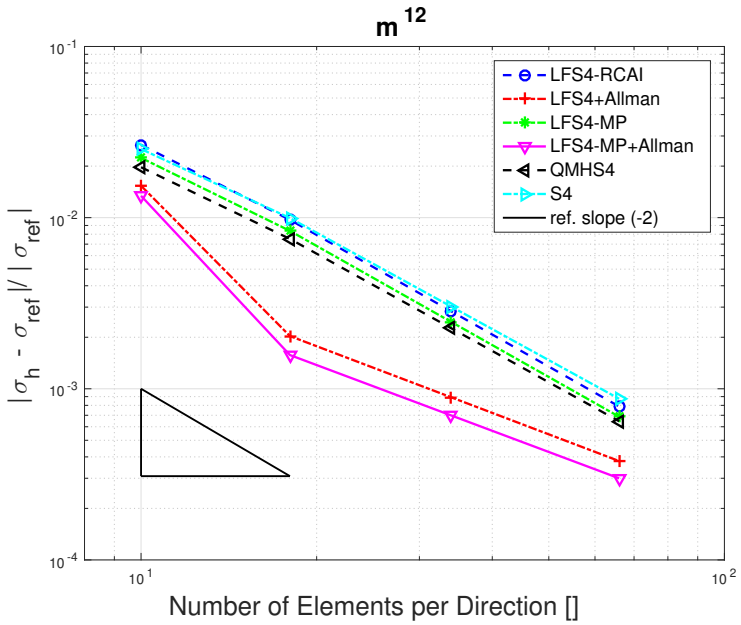


Figure 11: Pinched Hemispherical Shell with Hole: Rate of convergence of resultant torsional moment.

6.4 Raasch Hook

The Raasch hook or Raasch challenge as it is usually referred to involves a thick curved strip - essentially a hook like geometry - which is clamped on one edge and is subjected to a distributed unit tip in-plane shear load on the other, see Figure 12. The hook is composed out of two cylindrical segments of radius R_1, R_2 which span an opening angle of $60°$and $150°$respectively. The geometric as well as the material properties are shown in Figure 12. This example was first reported by I. Raasch of BMW in 1990 and revealed an erroneous behaviour (i.e. solution failed to converge with the mesh refinement) of the QUAD4 element of the MSC/NASTRAN FE code.

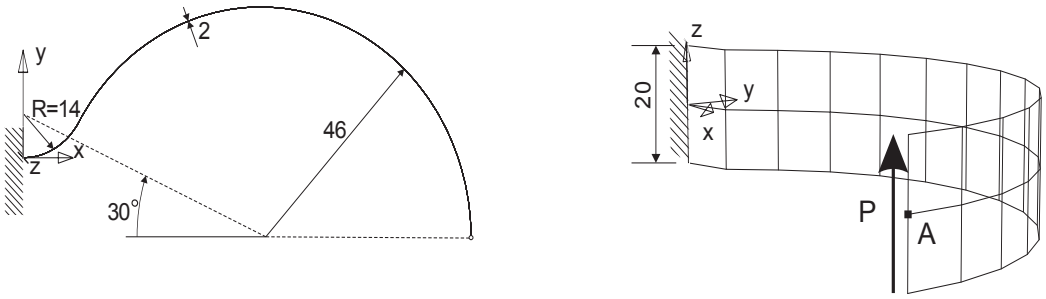


Figure 12: Raasch's Hook: Initial geometry, $E = 3300,\ \nu = 0.35,\ h = 2,\ P = 1$

In [24], it was found that the problem was caused by the treatment of the shell normals and resulted in an inability to transfer correctly the twisting moments. This failure introduced significant transverse shears whose effect was greater with the mesh refinement since the drill rotation was less and less restrained by the bending stiffness of an adjacent element. In Table 7 and Table 8 the displacements and drill rotations at the point A of the loaded edge are listed

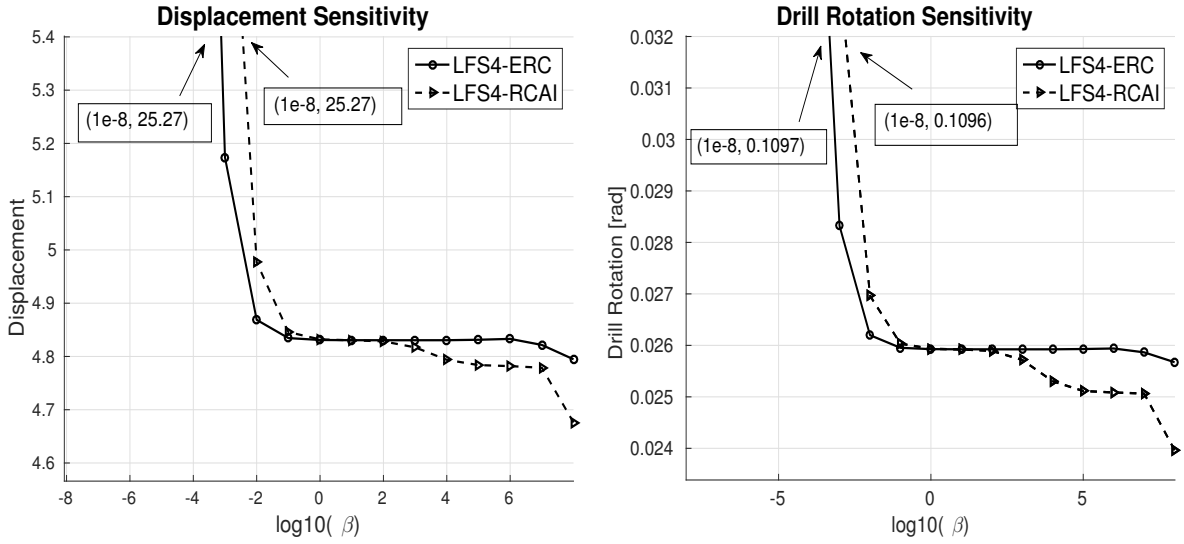
Elements	9×1	18×3	36×5	72×10	144×20
LFS4-RCAI	4.4376	4.6786	4.7987	4.9103	4.9769
LFS4-RCFI	4.4376	4.6786	4.7987	4.9103	4.9769
LFS4-ERC	4.4297	4.6085	4.6904	4.7761	4.8310
LFS4-MP	4.3799	4.5912	4.6731	4.7316	4.7581
LFS4-MP+Allman	4.3799	4.5890	4.6712	4.7313	4.7580
LFS4+Allman	4.4136	4.6917	5.0562	6.2416	10.528
LFS4-RCAI+Allman	4.4048	4.6293	4.7397	4.8599	4.9458
QMHS4	4.5462	4.6668	4.7109	4.7850	4.8368
S4	4.8531	4.8799	4.9541	5.0107	5.0335
CQUAD4	4.592	4.7105	4.7308	4.7252	4.7194
Ref.	4.9352 [19]				

Table 7: Raasch Hook: Vertical Displacement at Point A.

Elements	9×1	18×3	36×5	72×10	144×20
LFS4-RCAI	0.0235	0.0251	0.0259	0.0266	0.0270
LFS4-RCFI	0.0235	0.0251	0.0259	0.0266	0.0270
LFS4-ERC	0.0236	0.0247	0.0252	0.0256	0.0260
LFS4-MP	0.0231	0.0246	0.0251	0.0254	0.0256
LFS4-MP+Allman	0.0231	0.0246	0.0251	0.0254	0.0255
LFS4+Allman	0.0230	0.0240	0.0227	0.0161	-0.0110
LFS4-RCAI+Allman	0.0231	0.0246	0.0254	0.0261	0.0267
QMHS4	0.0134	0.0198	0.0227	0.0244	0.0253
S4	0.0257	0.0261	0.0265	0.0268	0.0269
CQUAD4	0.0247	0.0252	0.0253	0.0253	0.0253

Table 8: Raasch Hook: Drill Rotation at Point A.

In Figure 13, the sensitivity of the converged solutions with respect to the penalty parameter is plotted.

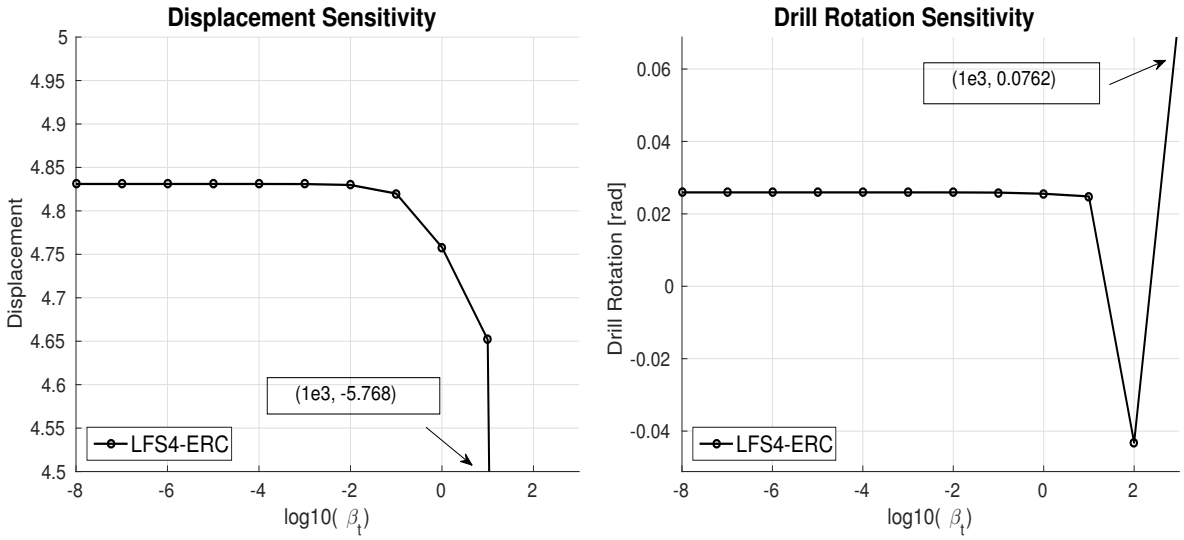


(a) Vertical Displacement

(b) Drill Rotation

Figure 13: Raasch Hook: Sensitivity of computed results with respect to penalty coefficient $\beta = \text{range}(10^8, 10^{-8})$.

Again the sensitivity of the computed results with respect to the twist parameter β_t is illustrated in Figure 14



(a) Vertical Displacement

(b) Drill Rotation

Figure 14: Raasch Hook: Sensitivity of LFS4-ERC with respect to the twist parameter $\beta_t = \text{range}(10^3, 10^{-8})$.

In Figure 15, the convergence of the stress resultant torsional moment is shown. For the Raasch Hook the mesh sizes analysed for the stress comparison are, 46×6, 82×10, 138×18 and 248×34.

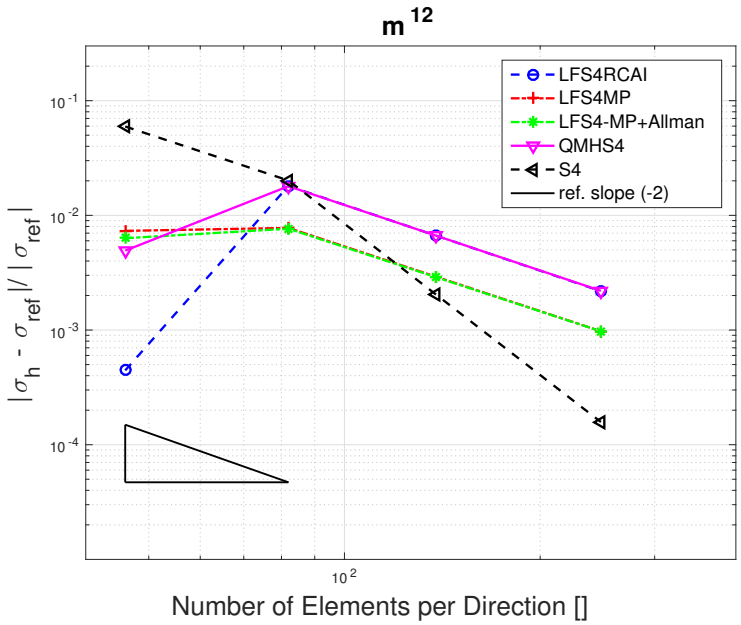


Figure 15: Raasch Hook: Rate of convergence of resultant torsional moment.

6.5 Pre-Twisted Beam

This test was initially proposed by [23] and it is very valuable since the effect of element warpage can be assessed. Although each element is approximately 7.5° warped, this amount can have quite significant effects, see [4]. The initial geometry of the beam is twisted by 90° with this being a stress free configuration. One edge is clamped while the other is subjected separately to three different concentrated loads at the tip. Thus the three load-cases are i) a unit shear load in width direction and ii) a unit shear load in thickness direction and iii) a unit axial load. Note that for the latter the initial twost of the beam is 180°. The material as well as geometrical properties are given in Figure 16.

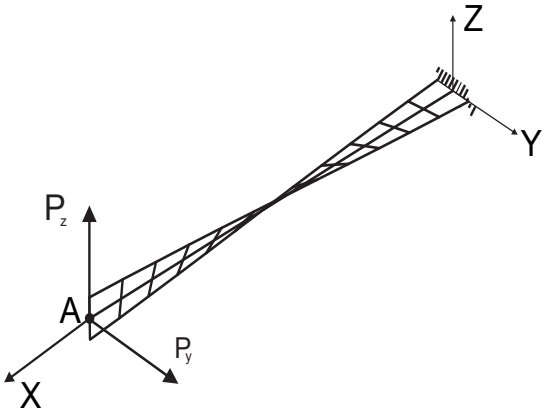


Figure 16: Pre-Twisted Beam Configuration: $E = 2.9 \times 10^7$, $\nu = 0.22$, $L = 12$, $w = 1.1$, twist=90°, $P_y = P_z = 1 \times 10^{-6}$ and thickness $h = 0.0032$

In [23], the thickness is $h = 0.32$ however, in [4] a thinner variant (i.e. $h = 0.0032$) is used in order to invoke membrane locking phenomena that are more significant for thinner shells. In the latter, the significance of accounting for the variation of the Jacobian through the thickness, i.e. account for curvature-displacement coupling, is illustrated. In the absence of these terms, an error as large as 70% can occur. The following results correspond to the thin case and the reference solutions are based on beam theory. The computed displacements for all loadcases are listed in Table 9, Table 10 and Table 11.

Elements	2×12	4×24	8×48
LFS4-RCAI	5.2781	5.2264	5.2421
LFS4-RCFI	5.2027	5.2358	5.2462
LFS4-ERC	5.2801	5.2374	5.2463
LFS4-MP	5.2757	5.2370	5.2460
LFS4-MP+Allman	5.2064	5.2351	5.2463
QMHS4	5.1768	5.2310	5.2462
S4	5.1969	5.2377	5.2482
CQUAD4	5.2674	5.2427	5.2471
Ref.		5.256 [4]	

Table 9: Pre-Twisted Beam: Load Case 1, In plane Displacement $100 \times u_2$

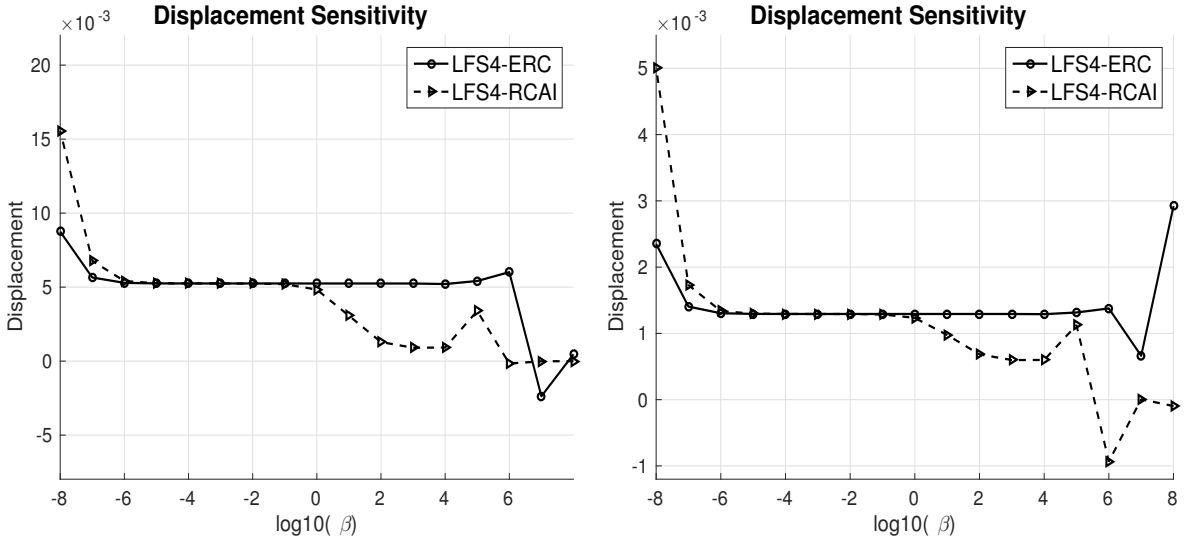
Elements	2×12	4×24	8×48
LFS4-RCAI	1.2990	1.2918	1.2925
LFS4-RCFI	1.2955	1.2932	1.2930
LFS4-ERC	1.2990	1.2936	1.2931
LFS4-MP	1.2989	1.2936	1.2931
LFS4-MP+Allman	1.2954	1.2933	1.2931
QMHS4	1.2691	1.2874	1.2917
S4	1.2748	1.2894	1.2922
CQUAD4	1.2791	1.2888	1.2919
Ref		1.294 [4]	

Table 10: Pre-Twisted Beam: Load Case 2, Out of plane Displacement $100 \times u_3$

Elements	2×12	4×24	8×48
LFS4-RCAI	6.197	4.482	4.479
LFS4-RCFI	4.885	4.812	4.493
LFS4-ERC	8.465	4.848	4.571
LFS4-MP	8.503	4.842	4.570
QMHS4	4.725	4.512	4.404
S4	3.445	4.153	4.290

Table 11: Pre-Twisted Beam: Load Case 3, Axial Displacement $10^5 \times u_3$

For the last load case the results of the LFS4-MP+Allman element are bad and therefore not included. In Figure 17, the sensitivity of the displacement with respect to the penalty parameter β for the first two load cases is plotted.

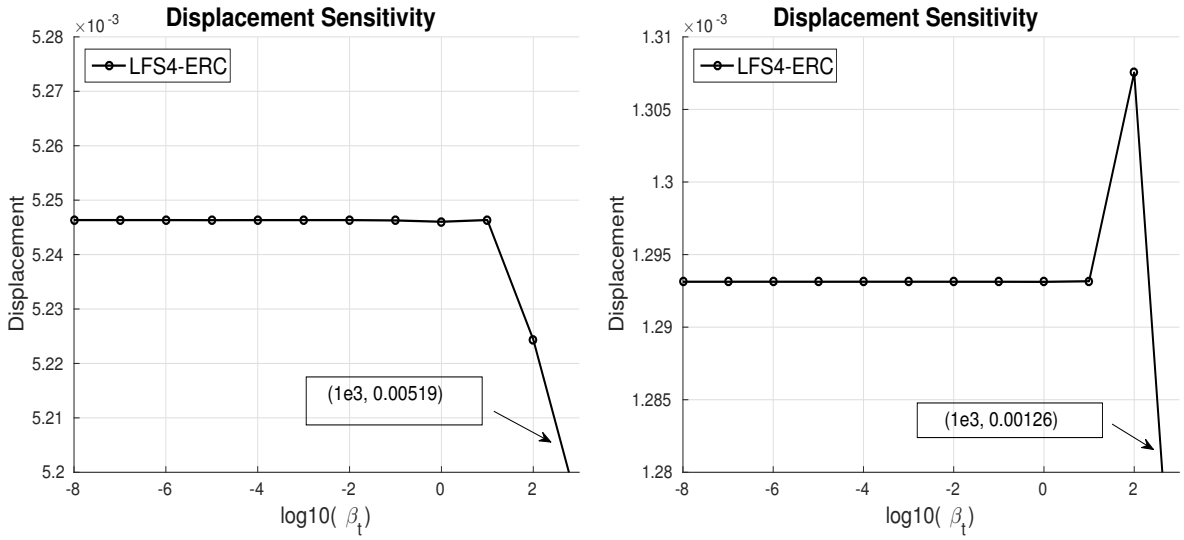


(a) Load Case 1

(b) Load Case 2

Figure 17: Pre-Twisted Beam: Sensitivity of computed results with respect to penalty parameter $\beta = \text{range}(10^8, 10^{-8})$.

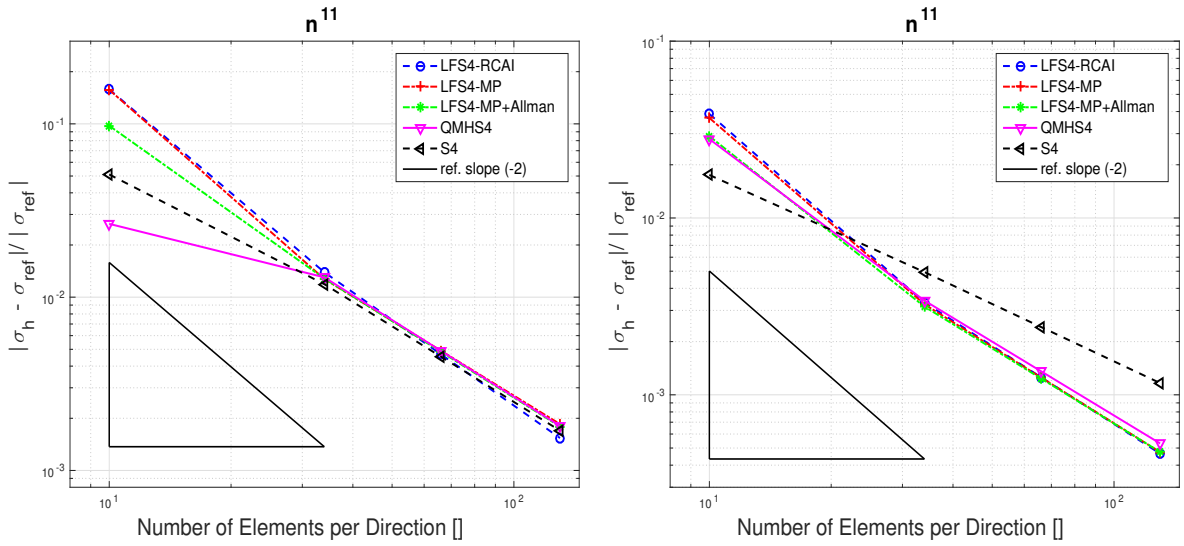
In Figure 18, the sensitivity with respect to the twist parameter for the LFS4-ERC is plotted. Note that only the first two load case are considered.



(a) Load Case 1 (b) Load Case 2

Figure 18: Pre-Twisted Beam: Sensitivity of LFS4-ERC with respect to the twist parameter $\beta_t = \text{range}(10^3, 10^{-8})$.

Again the selection of a zero twist penalty parameter can be justified. Closing this example the convergence of the membrane stress resultant SF1 is shown for both load cases in Figure 19a and Figure 19b. For this examples the mesh sizes used are, 10×2, 34×6, 66×10, 130×18.



(a) Load Case 1 (b) Load Case 2

Figure 19: Pre-Twisted Beam: Rate of convergence of resultant axial force for out-of plane shear load.

6.6 Pinched Cylinder with End Diaphragm

The fifth example involves a cylindrical shell with rigid diaphragms at both ends and is subjected to two opposing forces which are applied at the middle. The geometrical as well

as the material characteristics are shown in Figure 20. Here it is important to note that in the literature exist alternative geometrical and material properties, compare [27] and [33]. Here the latter is used. This is a very popular test since it involves inextensional bending deformation, finite rotations and quite complex stress states. Firstly, it was used by [22] and later by [3] for testing their MITC elements.

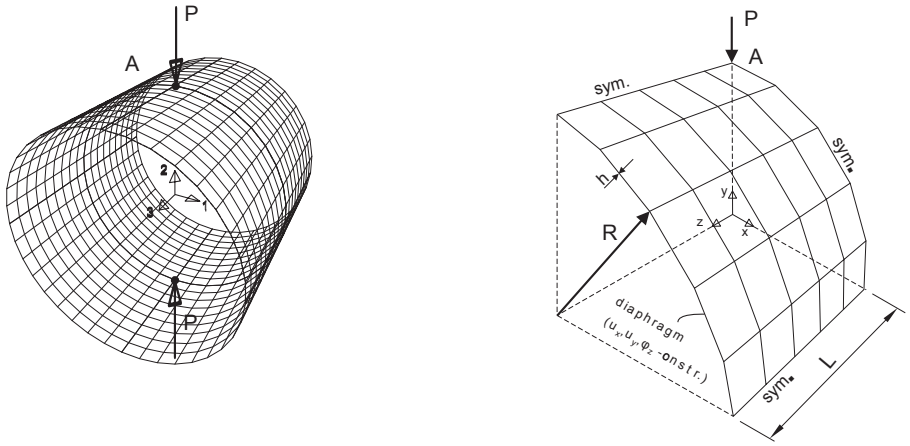


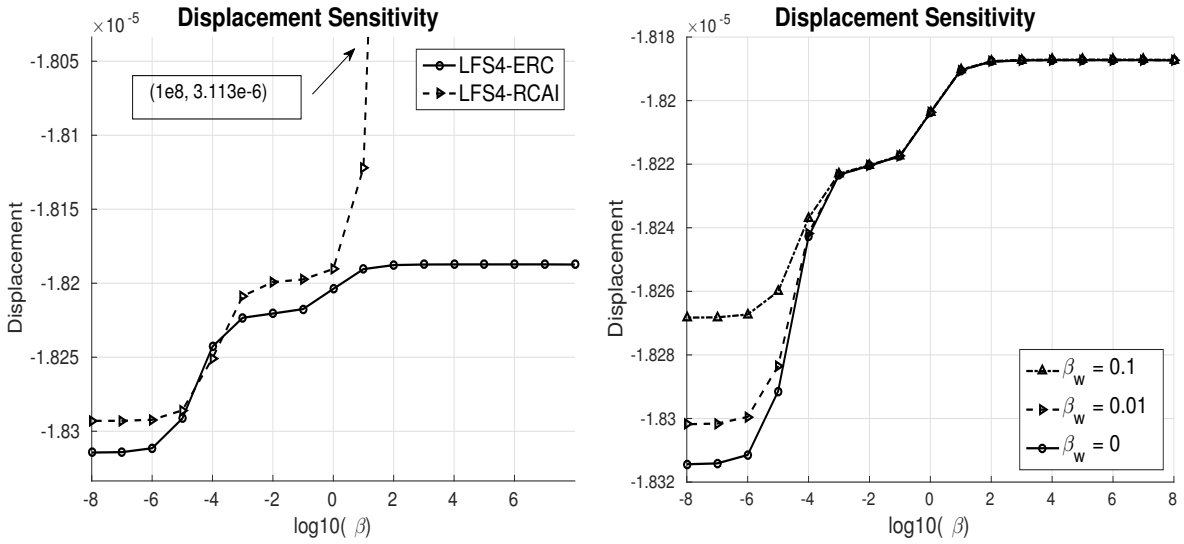
Figure 20: Pinched Cylinder with end diaphragms. Initial Geometry, $E = 3 \times 10^6$, $R = 300$, $L = 300$, $\nu = 0.3$, $h = 3$.

Taking advantage of the symmetry, only one-eighth of the cylinder is modelled. The vertical displacements of point A for a linear static analysis are listed in Table 12.

Elements	10×10	20×20	40×40
LFS4-RCAI	1.517	1.749	1.820
LFS4-RCE7	1.517	1.749	1.820
LFS4-ERC	1.508	1.749	1.820
LFS4-MP	1.497	1.742	1.818
LFS4-MP+Allman	1.488	1.740	1.817
QMHS4	1.532	1.758	-
S4	1.519	1.750	1.821
Ref.	1.8249 [33]		

Table 12: Pinched with end diaphragms: Vertical Displacement at Point A, $-u_3 \times 10^5$

The sensitivity of the computed displacement with respect to the penalty parameter and also the effect of the warping penalty parameter β_w for the LFS4-ERC implementation are shown in Figure 21.



(a) Comparison with LFS4-RCAI (b) Effect of warping parameter on LFS4-ERC

Figure 21: Pinched Cylinder with end diaphragms: Sensitivity of computed results with respect to penalty coefficient $\beta = \text{range}(10^8, 10^{-8})$.

The sensitivity with respect to the twist parameter for the LFS4-ERC is plotted in Figure 22.

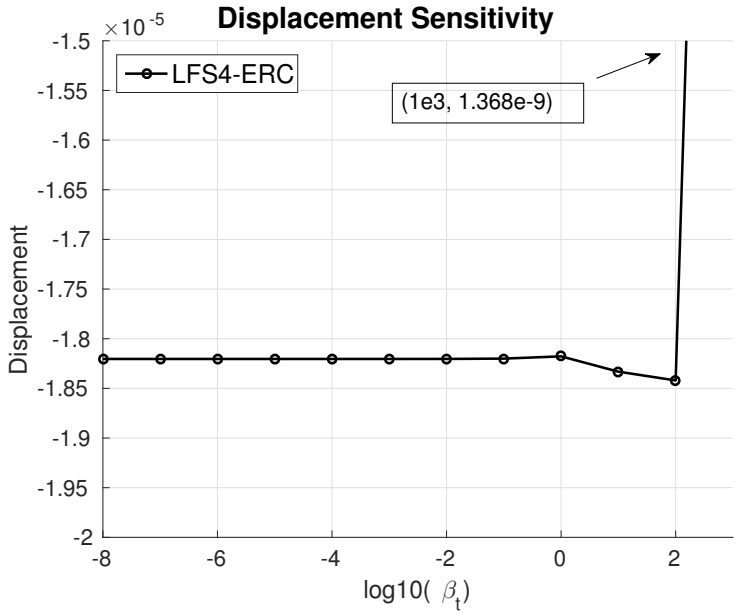


Figure 22: Pinched Cylinder with end diaphragms: Sensitivity of LFS4-ERC with respect to the twist parameter $\beta_t = \text{range}(10^3, 10^{-8})$.

Closing, the convergence of the resultant bending moment is shown in Figure 23. For this example the mesh sizes used are, 10×10, 18×18, 34×34, 66×66.

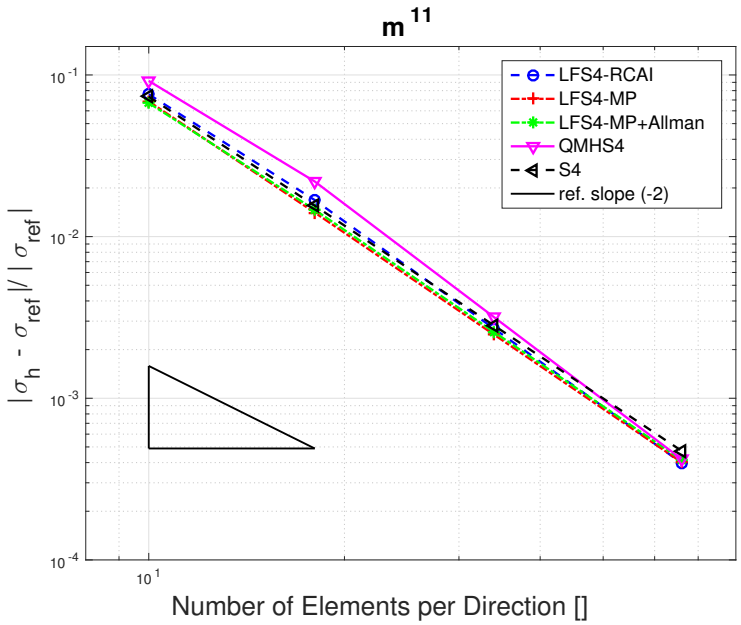


Figure 23: Pinched Cylinder with end Diaphragms: Rate of convergence of resultant bending moment.

6.7 C- Channel Section

The last benchmark example involves a short C-beam which is fully clamped at one end and is loaded by a vertical force at the other. The direction of the applied load is parallel to the web of the C-Section. The geometric as well as material properties are given in Figure 24. This test was proposed in [9] and it is a benchmark test where a drilling degree of freedom is required. This is due to the fact that at the juncture the rotations associated with flexure will induce in-plane twisting in the adjacent elements.

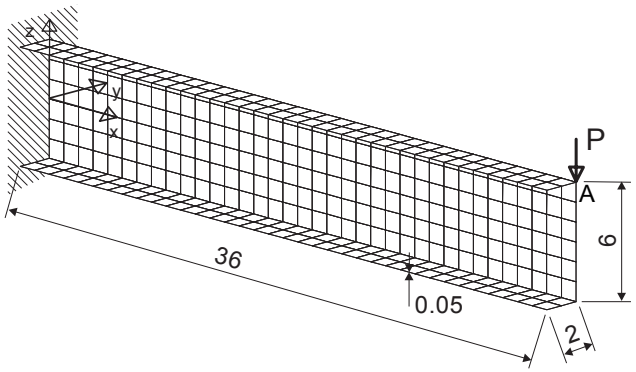


Figure 24: C-Section Channel Beam. $E = 10^7,\ \nu = 0.333,\ h = 0.05,\ P = 1$

The computed displacements as well as the drill rotation of point A are listed in Table 13 and Table 14 respectively.

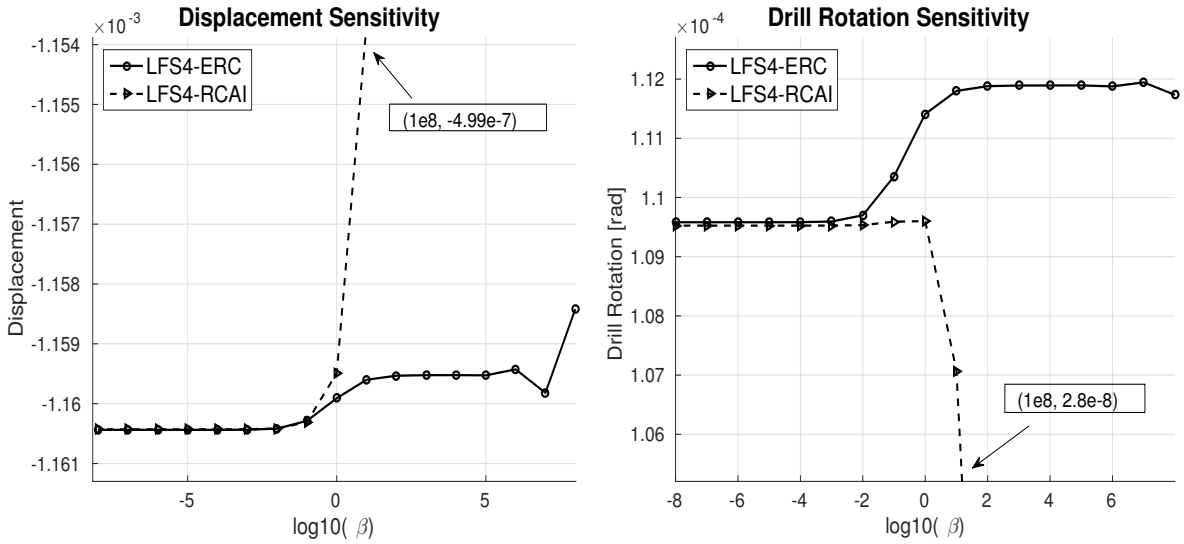
Elements	18×5	36×10	72×20	144×40	288×80
LFS4-RCAI	1.1492	1.1540	1.1576	1.1604	1.1628
LFS4-RCFI	1.1491	1.1540	1.1576	1.1604	1.1628
LFS4-ERC	1.1477	1.1530	1.1569	1.1599	1.1624
LFS4-MP	1.1476	1.1530	1.1569	1.1599	1.1624
LFS4-MP+Allman	1.1493	1.1544	1.1585	1.1616	1.1638
LFS4+Allman	1.1118	1.1451	1.1565	1.1613	1.1639
QMHS4	1.1496	1.1542	1.1577	1.1605	1.1628
S4	1.1488	1.1538	1.1575	1.1603	1.1627
CQUAD4	1.1450	1.1544	1.1578	1.1605	1.1626

Table 13: C-Section Channel Beam: Vertical Displacement $-u_1 \times 1000$

Elements	18×5	36×10	72×20	144×40	288×80
LFS4-RCAI	1.0178	1.0359	1.0585	1.0953	1.1508
LFS4-RCE7	1.0210	1.0348	1.0578	1.0951	1.1508
LFS4-ERC	1.0133	1.0333	1.0650	1.0951	1.1508
LFS4-MP	1.0139	1.0343	1.0650	1.1134	1.1872
LFS4-MP+Allman	1.0491	1.0874	1.1624	1.2912	1.4432
LFS4+Allman	0.9831	1.0684	1.1552	1.2814	1.4135
QMHS4	1.0270	1.0366	1.0586	1.0955	1.1513
S4	1.0239	1.0354	1.0577	1.0970	1.1617
CQUAD4	1.0227	1.0373	1.0565	1.0904	1.1423

Table 14: C-Section Channel Beam: Drill Rotation $u_6 \times 10^4$

The meshes used correspond to $n \times m$ elements where n is the number of elements in length direction and m is the total number of elements through the width and height. The sensitivity of the computed results with respect to the penalty parameter β is shown in Figure 25.

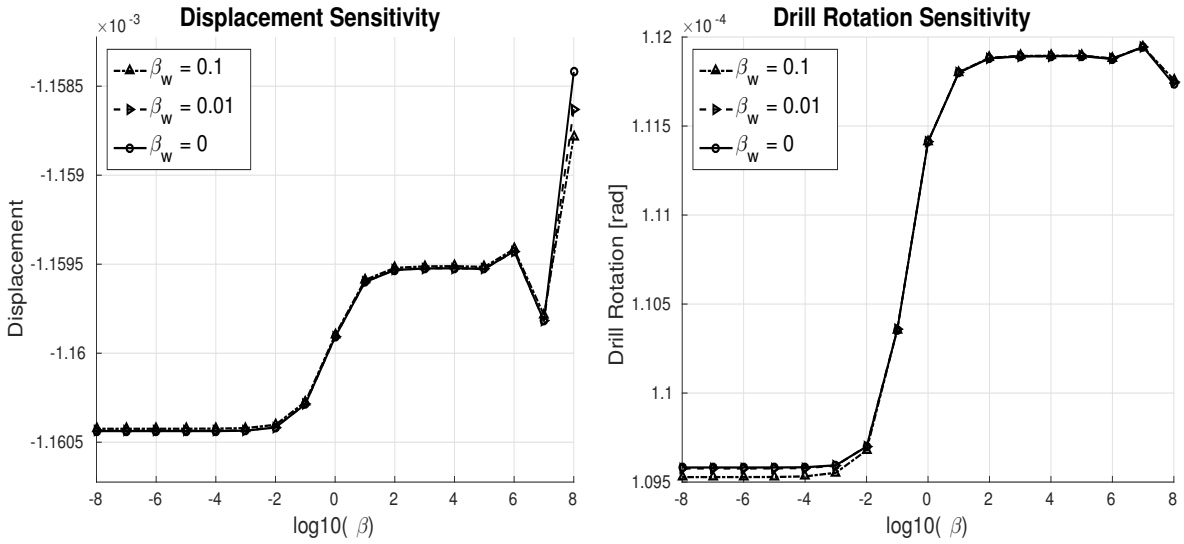


(a) Vertical Displacement

(b) Drill Rotation

Figure 25: C-Section: Sensitivity of computed results with respect to penalty parameter $\beta = \text{range}(10^8, 10^{-8})$.

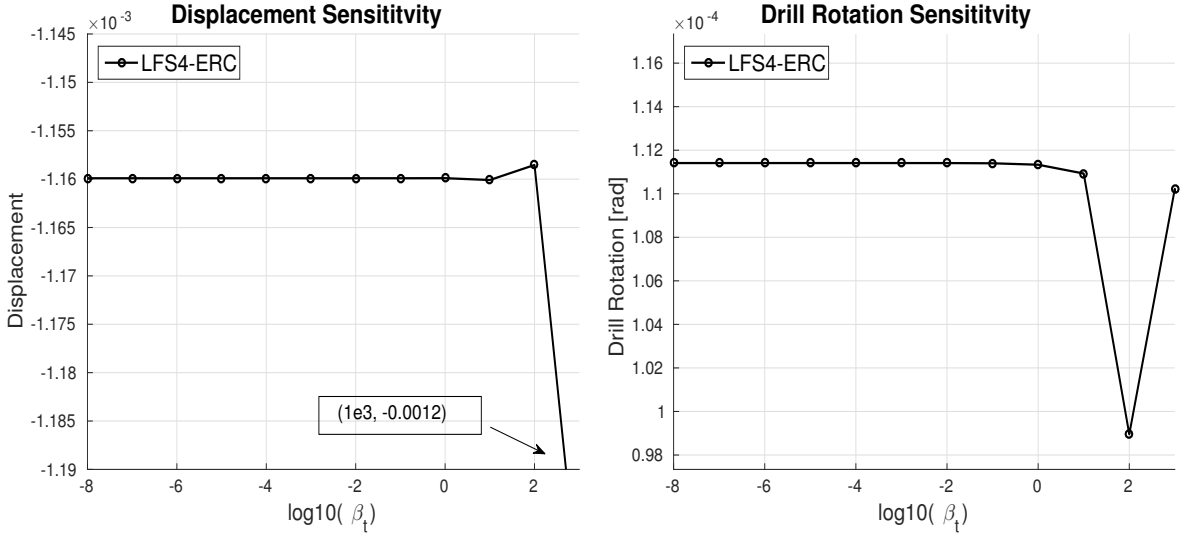
In Figure 26, the behaviour of the LFS4-ERC for various warping parameters is shown.



(a) Vertical Displacement (b) Drill Rotation

Figure 26: C-Section: Sensitivity of LFS4-ERC with respect to penalty parameter β for various warping parameters.

Lastly, the sensitivity of the computed solutions with respect to the twist parameter is plotted in Figure 27.



(a) Vertical Displacement (b) Drill Rotation

Figure 27: C-Section: Sensitivity of displacement with respect to the twist coefficient $\beta_t = \text{range}(10^3, 10^{-8})$.

7 Conclusions

It has been demonstrated that by introducing a proper incompatible enhancement of the drill rotation constraint the sensitivity of displacement and stress results on the penalty parameter $\gamma = \beta\,\mu$, which was present in earlier contributions, can be eliminated. This is valid for a wide range of β, i.e., roughly, $1 \leq \beta \leq 10^6$.

Additionally, it has been shown that micropolar shell element formulations involving additional material parameters, which have to be determined by ad-hoc assumptions, do not yield better results in terms of accuracy and convergence rate. Moreover, it is demonstrated that the stiffness related to the nonsymmetric membrane strain plays the role of a penalty parameter rather than that of a material parameter. None of the numerical examples indicate the necessity of assuming nonsymmetric bending strains.

The new element formulation allows the application of nodal drill moments. In this context, the first-order shear correction factor β_w might deserve further investigation. There are indications that it is related to a non-local characteristic of the material behaviour.

REFERENCES

[1] H. Altenbach and V. A. V. A. Eremeyev. Actual Developments in the Nonlinear Shell Theory - State of the Art and New Applications of the Six-Parameter Shell Theory. In *Shell Structures: Theory and Applications, Vol. 3*. CRC Press/Balkema, 2014.

[2] U. Andelfinger and E. Ramm. EAS-elements for two-dimensional, three-dimensional, plate and shell structures and their equivalence to HR-elements. *Int. J. Numer. Methods Eng.*, 36(8):1311–1337, 1993.

[3] K. J. Bathe and E. N. Dvorkin. A formulation of general shell elements - the use of mixed interpolation of tensorial components. *Int. J. Numer. Methods Eng.*, 22(July 1985):697–722, 1986.

[4] Ted Belytschko, L. B. Wong, and H. Stolarski. Assumed Strain Stabilization Procedure For The 9 Node Lagrange Shell Element. *Int. J. Numer. Methods Eng.*, 28:385–414, 1989.

[5] P. G. Bergan and C.A. Felippa. A triangular membrane element with rotational degrees of freedom. *Comput. Methods Appl. Mech. Eng.*, 50:25–69, 1985.

[6] Erasmo Carrera, Salvatore Brischetto, and Pietro Nali. *Plates and Shells for Smart Structures: Classical and Advanced Theories for Modeling and Analysis*. 2011.

[7] Ismail Celik and Ozgur Karatekin. Numerical Experiments on Application of Richardson Extrapolation With Nonuniform Grids. *J. Fluids Eng.*, 119(3):584, 1997.

[8] Ismail Celik, Jun Li, Gusheng Hu, and Christian Shaffer. Limitations of Richardson Extrapolation and Some Possible Remedies. *J. Fluids Eng.*, 127(5):795–805, 2005.

[9] J. Chróścielewski, J. Makowski, and H Stumpf. Genuinely resultant shell finite elements accounting for geometric and material non-linearity. *Int. J. Numer. Methods Eng.*, 35(October 1991):63–94, 1992.

[10] J. Chróścielewski and Wojciech Witkowski. Four-node semi-EAS element in six-field nonlinear theory of shells. *Int. J. Numer. Methods Eng.*, 68(11):1137–1179, 2006.

[11] Robert D. Cook. Improved two-dimensional finite element. *J. Struct. Div.*, 100(9):1851–1863, 1974.

[12] E. N. Dvorkin and K. J. Bathe. A continuum mechanics based four-node shell element for general non-linear analysis, 1984.

[13] V. A. Eremeyev and W. Pietraszkiewicz. Phase transitions in thermoelastic and thermoviscoelastic shells. *Arch. Mech.*, 61(1):41–67, 2009.

[14] D.D. Fox and J. C. Simo. A drill rotation formulation for geometrically exact shells, 1992.

[15] A E Green and W Zerna. *Theoretical Elasticity*. Phoenix Edition Series. Dover Publications, 1969.

[16] F. Gruttmann and W. Wagner. A linear quadrilateral shell element with fast stiffness computation. *Comput. Methods Appl. Mech. Eng.*, 194:4279–4300, 2005.

[17] A Ibrahimbegovic. Stress resultant geometrically nonlinear shell theory with drilling rotations-Part I. A consistent formulation. *Comput. Methods Appl. Mech. Eng.*, 7825(94):265–284, 1994.

[18] S Klinkel, F. Gruttmann, and W. Wagner. A mixed shell formulation accounting for thickness strains and finite strain 3d material models. *Int. J. Numer. Methods Eng.*, 74(October 2007):945–970, 2008.

[19] N. F. Knight. Raasch challenge for shell elements. *AIAA J.*, 35(2):375–381, 1997.

[20] W T Koiter. A Consistent First approximation in the General Theory of Thin Elastic Shells. In D. G. Ashwell and W. T. Koiter, editors, *IUTAM Symp. Theory Thin Elastic Shells, Delft*, pages 12–33, North Holland, Amsterdam, 1959. Laboratorium voor Toegepaste Mechanica der Technische Hogeschool.

[21] A Libai and J. G. Simmonds. *The Nonlinear Theory of Elastic Shells*. Cambridge University Press, 2nd edition, 1998.

[22] G M Lindberg, M D Olson, and G R Cowper. New Developments in the Finite Element Analysis of Shells. Technical report, Structures and Materials Laboratory National Aeronautical Establishement, Ottawa, 1969.

[23] R. H. MacNeal and R. L. Harder. A proposed standard set of problems to test finite element accuracy. *Finite Elem. Anal. Des.*, 1:3–20, 1985.

[24] R. H. MacNeal, Charles T. Wilson, R. L. Harder, and Claus C. Hoff. The treatment of shell normals in finite element analysis. *Finite Elem. Anal. Des.*, 30:235–242, 1998.

[25] G. Rebel. *Finite Rotation Shell Theory Including Drill Rotations and Its Finite Element Implementations*. Phd dissertation, Delft University of Technology, 1998.

[26] Eric Reissner. Linear and nonlinear theory of shells. *Thin-shell Struct. Theory, Exp. Des.*, pages 29–44, 1974.

[27] Carlo Sansour and Herbert Bednarczyk. The Cosserat surface as a shell model, theory and finite-element formulation. *Comput. Methods Appl. Mech. Eng.*, 120(1-2):1–32, 1995.

[28] J. G. Simmonds. The Nonlinear Thermodynamical Theory of Shells: Descent from 3-Dimensions without Thickness Expansions. In E. L. Axerlrad and F. A. Emmerling, editors, *Flex. Shells, Theory Appl.* Springer-Verlag, 1984.

[29] J. G. Simmonds and Donald A Danielson. Nonlinear shell theory with finite rotation and stress-function vectors. *J. Appl. Mech.*, 39(4):1085–1090, 1972.

[30] J. C. Simo and F Armero. Geometrically non-linear enhanced strain mixed methods and the method of incompatible modes. *Int. J. Numer. Methods Eng.*, 33(2):1413–1449, 1992.

[31] J. C. Simo, F Armero, and R. L. Taylor. Improved versions of assumed enhanced strain tri-linear elements for 3D finite deformation problems. *Comput. Methods Appl. Mech. Eng.*, 110(2):359–386, 1993.

[32] J. C. Simo and D.D. Fox. On a stress resultant geometrically exact shell model. Part I: Formulation and optimal parametrization, 1989.

[33] J. C. Simo, D.D. Fox, and M.S. Rifai. On a stress resultant geometrically exact shell model. Part II: The linear theory; Computational aspects. *Comput. Methods Appl. Mech. Eng.*, 73:53–92, 1989.

[34] J. C. Simo, D.D. Fox, and M.S. Rifai. On a stress resultant geometrically exact shell model. Part III: Computational aspects of the nonlinear theory, 1990.

[35] J. C. Simo and M.S. Rifai. A class of mixed assumed strain methods and the method of incompatible modes. *Int. J. Numer. Methods Eng.*, 29(October 1989):1595–1638, 1990.

[36] W. Wagner and F. Gruttmann. A robust non-linear mixed hybrid quadrilateral shell element. *Int. J. Numer. Methods Eng.*, 64(October 2004):635–666, 2005.

[37] Krzysztof Wiśniewski. *Finite Rotation Shells Basic Equations and Finite Elements for Reissner Kinematics*. Springer, 2010.

[38] Krzysztof Wiśniewski and E Turska. Kinematics of finite rotation shells with in-plane twist parameter. *Comput. Methods Appl. Mech. Eng.*, 190:1117–1135, 2000.

[39] Krzysztof Wiśniewski and Ewa Turska. Recent Improvements in Hu-Washizu Shell Elements with Drilling Rotation. In *Shell-like Struct.*, pages 391–412. Springer, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CURVED GRID GENERATION AND DG COMPUTATION FOR THE DLR-F11 HIGH LIFT CONFIGURATION

R. Hartmann[1], **H. McMorris**[2], **T. Leicht**[1]

[1]DLR (German Aerospace Center)
Institute of Aerodynamics and Flow Technology, Lilienthalplatz 7, 38104 Braunschweig, Germany
e-mail: {ralf.hartmann, tobias.leicht}@dlr.de

[2] CentaurSoft
3409 Executive Center Drive, Suite 241, Austin, TX 78731
e-mail: harlan@centaursoft.com

Keywords: High-order grid generation, high-order Discontinuous Galerkin method, compressible flows, high lift configuration

Abstract. *The DLR-F11 high lift configuration previously considered at the 2nd AIAA High Lift Prediction workshop (HiLiftPW-2) consists of a wing-body configuration with a three-element high lift system including slat tracks and flap track fairings. This test case has been selected a "computational/meshing challenge" in the 4th International Workshop on High-Order CFD Methods due to its geometrical and computational complexity. In this article, we demonstrate the applicability of current grid generation technology (CENTAUR) to generate a quadratic curved mesh for this configuration. Furthermore, we employ a high-order Discontinuous Galerkin discretization for the Reynolds-averaged Navier-Stokes (RANS) equations with the Wilcox-kω turbulence model which is sufficiently stable to be applicable to this test case. Finally, a 3rd-order Discontinuous Galerkin (DG) flow solution is computed on the quadratic (3rd-order) computational mesh. The numerical results are compared to experimental data as well as to computational results published in the HiLiftPW-2.*

1 INTRODUCTION

In recent years there has been significant progress in the development of curved grid generation as well as of high-order aerodynamic flow solvers, cf. the EU-projects ADIGMA [19] and its successor IDIHOM [20], for example. Nevertheless, in terms of geometrical and aerodynamic complexity the capabilities reached were still far from being applicable to "real-world" industrial test cases.

Particularly problematic was the generation of computational meshes sufficiently coarse to be suited to higher order methods and the subsequent curving of these meshes for a higher order approximation of curved wall boundaries, which is known to be particularly difficult in the presence of the highly stretched mesh elements typically used in the boundary layer of high Reynolds number aerodynamic flow simulations. There have been a few efforts to generate high-order meshes for flow field computations as demonstrated in [7], [22], and [28]. The majority of these existing methods start with a relatively coarse linear grid and insert the extra points needed to convert each element to high-order. There have also been methods that start with a curved, high-order surface mesh. This surface is then marched outwards in order to create the volume elements. The robustness of both the conversion method and the marching process is an issue in the vicinity of highly curved convex and concave regions within the geometry. A primary issue for high-order meshes is to ensure validity of the generated elements [17]. Work on creating high order elements only in curved regions of the geometry has also been studied [18].

Also, the high-order flow solvers reached their limits in terms of the stability of the discretizations and the solvers employed. Much effort concentrated on the development and industrialization of high-order CFD methods [19, 20] as well as on comparing the methods in the International Workshops on High-Order CFD Methods [25]. However, the configurations treated were bound to low to medium complexity. More recently, Discontinuous Galerkin discretizations were applied to the Common Research Model (CRM) wing-body configuration [6, 10] and to a simple 3D high lift configuration in [15]. Here, it was particularly important for the stability of the discretization to use a discretization at the wall boundary which resembles as closely as possible the discretization on interior faces, and to use an according discretization of force coefficients to keep adjoint consistency [14].

The DLR-F11 high lift configuration previously considered at the 2nd AIAA High Lift Prediction workshop [1] consists of a wing-body configuration with a three-element high lift system including slat tracks and flap track fairings. This test case has been selected a "computational / meshing challenge" in the 4th International Workshop on High-Order CFD Methods [16] due to its geometrical and computational complexity.

In the current work, we give details on how we create a coarse linear mesh for this configuration using the CENTAUR grid generator [5], and describe the algorithms developed to curve the mesh with piecewise quadratic polynomials for a 3rd-order boundary approximation. Furthermore, we describe the main ingredients of the discretization employed [14, 15] which we found sufficiently stable to be applicable to a test case of this complexity. Finally, the 3rd-order DG flow solution computed on the quadratic (3rd-order) computational mesh is compared to experimental data [1] as well as to numerical results published in the HiLiftPW-2 [1].

2 LINEAR AND QUADRATIC GRID GENERATION

In order to generate curved meshes for high-order CFD simulations, two problems must be solved. First, a coarse linear mesh must be generated that both sufficiently models the geometry

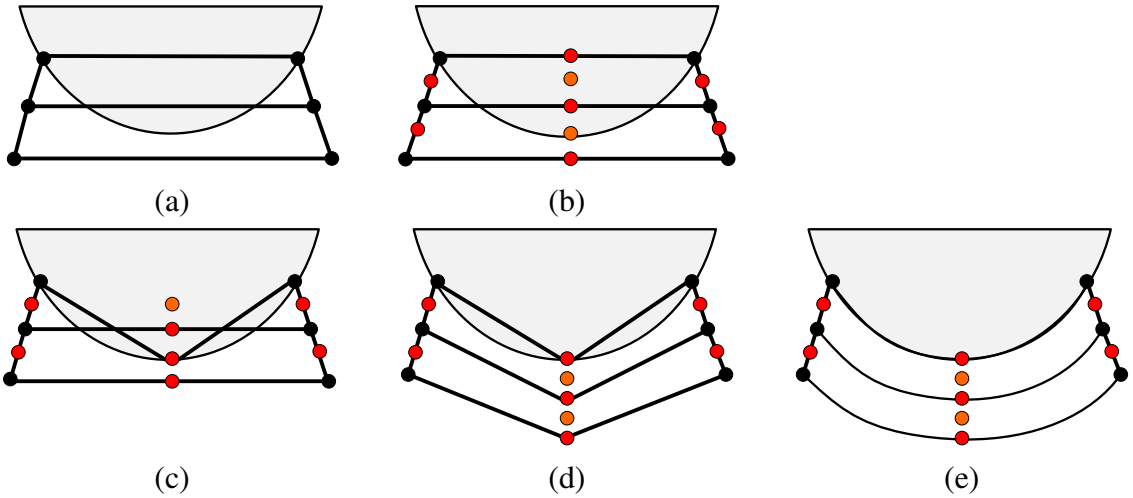


Figure 1: Sequence of steps used by CENTAUR to construct curved, high-order grids.

and ensures that any subsequent high-order mesh is not too big to use for simulations. Second, a robust method to curve the linear mesh is required. This method must be able to work for all types of elements. In particular, the method must be able to handle highly stretched mesh elements used to model the boundary layer region in high Reynolds number aerodynamic flow simulations. The method must also be able to handle highly curved surfaces in both convex and concave regions.

For the purpose of this work, the CENTAUR grid generator [5] is used to create a quadratic curved 3D hybrid grid. The basic algorithm used to create a curved, high-order mesh is as follows:

(a) Generate a coarse, linear hybrid grid.

(b) Insert an additional midpoint for every grid edge and quadrilateral face.

(c) Use the CAD information to map each new boundary point onto the underlying CAD surface.

(d) Adjust the position of the interior points based on the mapped position of the boundary edge midpoints, in order to prevent self-intersecting grid elements and to ensure grid validity.

First, a coarse linear 3D hybrid grid is created. CENTAUR allows for the mesh generation process to be tuned so that fewer elements are created while still preserving the geometry. For the surface mesh generation, a lower analytic curvature clustering is used along with a larger maximum element size. Next, prisms or hexahedra generated in the boundary layer are similarly adjusted to increase the first layer thickness in preparation for the added high-order points. Then, tetrahedra are used to automatically fill the remainder of the domain matching the length scales from both the surface mesh and the boundary layer mesh.

Second, the linear mesh is converted into a high-order mesh. As shown in Figure 1(b), new mid-edge and mid-face points are added to all edges and quadrilateral faces in the linear mesh. Any new point that is located on a boundary surface is then mapped onto the CAD surface using the same geometry routines used during mesh generation to place points on the surface. At this stage, the mesh may be invalid in convex curved regions as seen in Figure 1(c). While in other

curved regions, the mesh point spacing may be valid but distorted after the mapping process is completed. To correct both types of problems, the motion of the boundary points is smoothed into the interior as shown in Figure 1(d).

After this process is completed, the grid quality and validity is evaluated using the volumes of the complete elements and sub-elements formed using the points added in Step (b). To ensure good quality, volume ratios of the elements and sub-elements are also used. If any element violates the quality or validity measures, its points are adjusted to ensure the criteria are met. This process iterates until the mesh meets the various criteria.

3 DG DISCRETIZATION OF THE RANS AND WILCOX k-ω EQUATIONS

In this section, we give details on the Discontinuous Galerkin discretization similar to that previously developed in [14] and already applied to a simple 3D high lift configuration in [15]. In particular, we consider the Reynolds-averaged Navier-Stokes (RANS) equations and the Wilcox k-ω turbulence model equations [26, 27],

$$\nabla \cdot (\mathcal{F}^c(\mathbf{u}) - \mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u})) = \mathcal{S}(\mathbf{u}, \nabla\mathbf{u}) \quad \text{in } \Omega, \tag{1}$$

where $\mathbf{u} = (\rho, \rho v_1, \rho v_2, \rho v_3, \rho E, \rho k, \rho\tilde{\omega})^\top$ is the vector of conservative variables, with ρ, E, k, and $\mathbf{v} = (v_1, v_1, v_3)^\top$, denoting the density, specific total energy, turbulence kinetic energy, and velocity vector, respectively. Similar to Bassi *et al.* [3] the equations are considered in terms of the auxiliary variable $\tilde{\omega} = \ln\omega$ instead of the specific dissipation rate ω for a more moderate near-wall behavior of the variable and for guaranteeing positivity of ω. Additionally, we apply some realizability conditions for the turbulent stresses [3, 11]. For a detailed description of the convective and viscous fluxes, $\mathcal{F}^c(\mathbf{u})$ and $\mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u})$, and the source terms $\mathcal{S}(\mathbf{u}, \nabla\mathbf{u})$ involved in (1), and of the boundary conditions imposed on the boundary $\Gamma = \partial\Omega$, we refer to [9, 13]. In this work, adiabatic no-slip wall boundary conditions, $\mathbf{v} = 0$, $\mathbf{n} \cdot \nabla T = 0$, are imposed on the wall boundary Γ_{W}.

Let the domain Ω be subdivided into a shape-regular mesh $\mathcal{T}_h = \{\kappa\}$ consisting of (possibly curved) elements κ. Furthermore, let $\mathbf{V}_h^p$ be the finite element space consisting of discontinuous vector-valued polynomial functions of degree $p \geq 0$ on $\mathcal{T}_h$, then the Discontinuous Galerkin discretization of (1) employed in this work is given by: Find $\mathbf{u}_h$ in $\mathbf{V}_h^p$ such that

$$\begin{aligned}\int_\Omega \{(-\mathcal{F}^c(\mathbf{u}_h) + \mathcal{F}^v(\mathbf{u}_h, \nabla_h\mathbf{u}_h)) : \nabla_h\mathbf{v}_h - \mathcal{S}(\mathbf{u}_h, \nabla_h\mathbf{u}_h) \cdot \mathbf{v}_h\} \, \mathrm{d}\mathbf{x} \\ + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa} \left(\hat{\mathbf{h}}_h - \underline{\hat{\sigma}}_h\mathbf{n}\right) \cdot \mathbf{v}_h \, \mathrm{d}s + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa} (\hat{\mathbf{u}}_h - \mathbf{u}_h) \otimes \mathbf{n} : \left(G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h\right) \, \mathrm{d}s = 0\end{aligned} \tag{2}$$

for all $\mathbf{v}_h \in \mathbf{V}_h^p$, where $\mathbf{n}|_{\partial\kappa}$ denotes the outward unit normal vector to the boundary $\partial\kappa$ of element κ. Here, the convective and diffusive numerical flux functions, $\hat{\mathbf{h}}_h$ and $\underline{\hat{\sigma}}_h$, are approximations to the normal convective flux $\mathcal{F}^c(\mathbf{u}_h) \cdot \mathbf{n}$ and the viscous flux $\mathcal{F}^v(\mathbf{u}_h, \nabla_h\mathbf{u}_h)$, respectively. On an interior face $\partial\kappa \cap \partial\kappa'$ between two neighboring elements $\kappa, \kappa' \in \mathcal{T}_h$, the numerical flux functions

$$\begin{aligned}\hat{\mathbf{h}}_h &= \hat{\mathbf{h}}(\mathbf{u}_h, \mathbf{n}) = \hat{\mathbf{h}}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}), \\ \underline{\hat{\sigma}}_h &= \underline{\hat{\sigma}}(\mathbf{u}_h, \nabla\mathbf{u}_h) = \underline{\hat{\sigma}}(\mathbf{u}_h^+, \mathbf{u}_h^-, \nabla\mathbf{u}_h^+, \nabla\mathbf{u}_h^-),\end{aligned} \tag{3}$$

connect the interior and the exterior traces, $\mathbf{u}_h^+$ and $\mathbf{u}_h^-$, and their derivatives, $\nabla\mathbf{u}_h^+$ and $\nabla\mathbf{u}_h^-$, of

$\mathbf{u}_h$. On a boundary face $\partial\kappa \cap \Gamma \neq \emptyset$, the numerical boundary flux functions,

$$\begin{aligned} \hat{\mathbf{h}}_h|_\Gamma &= \hat{\mathbf{h}}_{\Gamma,h} = \hat{\mathbf{h}}_\Gamma(\mathbf{u}_h^+, \mathbf{n}), \\ \underline{\hat{\sigma}}_h|_\Gamma &= \underline{\hat{\sigma}}_{\Gamma,h} = \underline{\hat{\sigma}}_\Gamma(\mathbf{u}_h^+, \nabla\mathbf{u}_h^+). \end{aligned} \tag{4}$$

depend on the interior trace $\mathbf{u}_h^+$, directly and/or through the boundary function $\mathbf{u}_\Gamma(\cdot)$ given by

$$\mathbf{u}_\Gamma(\mathbf{u}_h) = (u_{h,1}, 0, 0, 0, u_{h,5}, 0, u_{h,1}\tilde{\omega}_{\text{wall}})^\top \qquad \text{on } \Gamma_{\text{W}}, \tag{5}$$

Here, $\tilde{\omega}_{\text{wall}}$ is determined using a projection of the analytic $\ln(\omega)$ near-wall behavior onto the polynomial ansatz space in the elements at the wall [23]. Similarly, the vector-valued numerical flux function $\hat{\mathbf{u}}_h$ in (2) is an approximation to $\mathbf{u}_h$ and is given by $\hat{\mathbf{u}}_h = \hat{\mathbf{u}}(\mathbf{u}_h) = \hat{\mathbf{u}}(\mathbf{u}_h^+, \mathbf{u}_h^-)$ on interior faces and by $\hat{\mathbf{u}}_h|_\Gamma = \hat{\mathbf{u}}_{\Gamma,h} = \hat{\mathbf{u}}_\Gamma(\mathbf{u}_h^+)$ on boundary faces. Finally, $G(\mathbf{u})$ denotes the homogeneity tensors defined by $\mathbf{f}_k^v(\mathbf{u}, \nabla\mathbf{u}) = G_{kl}(\mathbf{u})\partial\mathbf{u}/\partial x_l$, $k, l = 1, 2, 3$, for $\mathcal{F}^v(\mathbf{u}_h, \nabla_h\mathbf{u}_h) = (\mathbf{f}_1^v, \mathbf{f}_2^v, \mathbf{f}_3^v)$. Assuming that the numerical fluxes $\hat{\mathbf{h}}_h$ and $\underline{\hat{\sigma}}_h$ are consistent, then (2) is a consistent discretization of the flow equations (1), cf. [13, 14] for more details.

3.1 Numerical fluxes on interior faces

In the following, we give details on the numerical fluxes used on interior faces. For the convective numerical flux $\hat{\mathbf{h}}$ a Roe flux is chosen based on the diagonalization of the full Jacobian of the convective flux [24]. Furthermore, an entropy fix is employed similar to that of Harten [8]. Finally, the numerical fluxes for the diffusive terms are those of the BR2 scheme [3],

$$\hat{\mathbf{u}}_h = \{\!\{\mathbf{u}_h\}\!\}, \qquad\qquad \underline{\hat{\sigma}}_h = \{\!\{G(\mathbf{u}_h)\nabla_h\mathbf{u}_h\}\!\} - \underline{\delta}(\mathbf{u}_h), \tag{6}$$

where the penalization term $\underline{\delta}(\mathbf{u}_h)$ is given by

$$\underline{\delta}(\mathbf{u}_h) = C_{\text{BR2}}\{\!\{G(\mathbf{u}_h)\underline{L}_0^e(\mathbf{u}_h)\}\!\}. \tag{7}$$

Here $\{\!\{\underline{\tau}\}\!\} = \frac{1}{2}(\underline{\tau}^+ + \underline{\tau}^-)$ denotes the standard mean used in the DG context, and $\underline{L}_0^e(\mathbf{u}_h)$ denotes the local lifting operator, cf. [3, 14] for more details. Note that the penalization constant C_{BR2} must be sufficiently large for stability [4] and is taken as the number of faces of an element.

3.2 Numerical fluxes on wall boundary faces

The discretization on the boundary is chosen as close as possible to the discretization employed on interior faces. In particular, we use the same numerical flux functions for defining the numerical boundary fluxes as employed on interior faces, i.e., we consider

$$\hat{\mathbf{h}}_\Gamma(\mathbf{u}_h^+, \mathbf{n}) = \hat{\mathbf{h}}(\mathbf{u}_h^+, \mathbf{u}_\Gamma^-(\mathbf{u}_h^+), \mathbf{n}), \;\; \hat{\mathbf{u}}_{\Gamma,h} = \{\!\{\mathbf{u}_h\}\!\}_\Gamma, \;\; \underline{\hat{\sigma}}_{\Gamma,h} = \{\!\{\widetilde{\mathcal{F}}^v(\mathbf{u}_h, \nabla\mathbf{u}_h)\}\!\}_\Gamma - \underline{\tilde{\delta}}_\Gamma(\mathbf{u}_h^+), \tag{8}$$

where the boundary mean value $\{\!\{\cdot\}\!\}_\Gamma$ of $\mathbf{u}_h$ is defined by $\{\!\{\mathbf{u}_h\}\!\}_\Gamma = \frac{1}{2}\left(\mathbf{u}_h^+ + \mathbf{u}_\Gamma^-(\mathbf{u}_h^+)\right)$, and the wall exterior state $\mathbf{u}_\Gamma^- = \mathbf{u}_\Gamma^-(\mathbf{u}_h^+)$ is obtained by mirroring the interior state $\mathbf{u}_h^+$ at the wall boundary state $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$ (cf. (5)) according to $\frac{1}{2}\left(\mathbf{u}_h^+ + \mathbf{u}_\Gamma^-(\mathbf{u}_h^+)\right) = \mathbf{u}_\Gamma(\mathbf{u}_h^+)$. Furthermore, for the BR2 scheme [3] the penalization term $\underline{\tilde{\delta}}_\Gamma$ in (8) is given by

$$\begin{aligned} \underline{\tilde{\delta}}_\Gamma(\mathbf{u}_h^+) &= C_{\text{BR2}}\{\!\{\widetilde{G}(\mathbf{u}_h)\underline{L}_\Gamma^e(\mathbf{u}_h)\}\!\}_\Gamma = \tfrac{C_{\text{BR2}}}{2}\left(\widetilde{G}(\mathbf{u}_h^+)\underline{L}_\Gamma^e(\mathbf{u}_h^+) + \widetilde{G}(\mathbf{u}_\Gamma^-)\underline{L}_\Gamma^e(\mathbf{u}_\Gamma^-)\right) \\ &= C_{\text{BR2}}\{\!\{\widetilde{G}(\mathbf{u}_h)\}\!\}_\Gamma\, \underline{L}_\Gamma^e(\mathbf{u}_h^+), \end{aligned} \tag{9}$$

where $\underline{L}^e_\Gamma(\mathbf{u})$ denotes the local lifting operator on the boundary with $\underline{L}^e_\Gamma(\mathbf{u}^+_h) = \underline{L}^e_\Gamma(\mathbf{u}^-_\Gamma)$, cf. [13, 14] for more details. Finally, the diffusive flux $\widetilde{\mathcal{F}}^v$ in (8) and the corresponding homogeneity tensor $\widetilde{G}$ in (9) is modified on the adiabatic walla boundary Γ_{W} such that no heat flux is added, corresponding to the adiabatic condition $\mathbf{n} \cdot \nabla T = 0$, i. e.,

$$\mathbf{n} \cdot \widetilde{\mathcal{F}}^v(\mathbf{u}, \nabla\mathbf{u}) = \mathbf{n} \cdot \Big(\widetilde{G}(\mathbf{u})\nabla\mathbf{u}\Big) = (0,\ (\underline{\tau}\,\mathbf{n})_1,\ (\underline{\tau}\,\mathbf{n})_2,\ \mathbf{n} \cdot (\underline{\tau}\,\mathbf{v}))^\top \text{on } \Gamma_{\mathrm{W}}. \tag{10}$$

The boundary treatment of fluxes in (8) corresponds to introducing a ghost layer of elements at the wall boundary and evaluating the discretization on the wall boundary like on interior faces. Finally, $(\nabla\mathbf{u})^-_\Gamma$ in $\{\!\{\widetilde{\mathcal{F}}^v(\mathbf{u}_h, \nabla\mathbf{u}_h)\}\!\}_\Gamma = \frac{1}{2}(\widetilde{\mathcal{F}}^v(\mathbf{u}^+_h, \nabla\mathbf{u}^+_h) + \widetilde{\mathcal{F}}^v(\mathbf{u}^-_\Gamma, (\nabla\mathbf{u})^-_\Gamma))$ is the wall exterior gradient which might depend on the interior state and gradient but is chosen as $(\nabla\mathbf{u})^-_\Gamma = \nabla\mathbf{u}^+_h$, for simplicity, which results in $\{\!\{\widetilde{\mathcal{F}}^v(\mathbf{u}_h, \nabla\mathbf{u}_h)\}\!\}_\Gamma = \{\!\{\widetilde{G}(\mathbf{u}_h)\}\!\}_\Gamma \nabla\mathbf{u}^+_h$.

3.3 Evaluation of force coefficients

Let us consider the total drag and lift coefficients, C_{D} and C_{L},

$$J(\mathbf{u}) = \int_{\Gamma_{\mathrm{W}}} (p\,\mathbf{n} - \underline{\tau}\,\mathbf{n}) \cdot \boldsymbol{\psi}\,\mathrm{d}s = \int_{\Gamma_{\mathrm{W}}} (p\,n_i - \tau_{ij} n_j)\,\psi_i\,\mathrm{d}s, \tag{11}$$

where $\tau = (\mu + \mu_t)\left(\nabla\mathbf{v} + (\nabla\mathbf{v})^\top - \frac{2}{3}(\nabla\cdot\mathbf{v})I\right)$ presents the stress tensor at the wall including the viscous and turbulent viscosities μ and μ_t. Furthermore, $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi}_d = \frac{1}{C_\infty}(\cos(\alpha), 0, \sin(\alpha))^\top$ or $\boldsymbol{\psi}_l = \frac{1}{C_\infty}(-\sin(\alpha), 0, \cos(\alpha))^\top$ for the drag and lift coefficient, respectively, and α is the angle of attack. Finally, $C_\infty = q_\infty A$, where $q = \frac{1}{2}\rho|\mathbf{v}|^2$ denotes the dynamic pressure and A denotes a reference area and subscripts $_\infty$ indicate freestream quantities.

Given that the numerical boundary fluxes $\hat{\mathbf{h}}_h|_\Gamma$ and $\underline{\hat{\sigma}}_h|_\Gamma$ are consistent, $J_h(\mathbf{u}_h)$ defined by

$$\begin{aligned} J_h(\mathbf{u}_h) &= \int_{\Gamma_{\mathrm{W}}} \Big(\hat{\mathbf{h}}_{\Gamma,h} - \underline{\hat{\sigma}}_{\Gamma,h}\mathbf{n}\Big) \cdot \tilde{\boldsymbol{\psi}}\,\mathrm{d}s \\ &= \int_{\Gamma_{\mathrm{W}}} \Big(\hat{\mathbf{h}}(\mathbf{u}^+_h, \mathbf{u}^-_\Gamma(\mathbf{u}^+_h), \mathbf{n}) - \{\!\{\widetilde{G}(\mathbf{u}_h)\}\!\}_\Gamma \nabla u^+_h + \tilde{\underline{\delta}}_\Gamma(\mathbf{u}^+_h)\Big) \cdot \tilde{\boldsymbol{\psi}}\,\mathrm{d}s. \end{aligned} \tag{12}$$

with $\tilde{\boldsymbol{\psi}} = (0, \psi_1, \psi_2, \psi_3, 0)^\top$, is a consistent discretization of the force coefficient $J(\mathbf{u})$ in (11). Finally, the discretization of the force coefficients according to (12) is required for a discretization of (1) to be adjoint consistent [14].

3.4 Evaluation of c_p- and c_f-values

Related to the evaluation of integral quantities (the force coefficients) is the evaluation of local quantities at the wall boundary like surface pressure and skin friction as involved in c_{p}- and c_{f}-distributions. The local c_{p}- and c_{f}-values at the wall boundary Γ_{W} are given by

$$c_{\mathrm{p}} = c_{\mathrm{p}}(\mathbf{u}) = \frac{p(\mathbf{u}) - p_\infty}{\frac{1}{2}\rho_\infty v^2_\infty}, \qquad c_{\mathrm{f}} = c_{\mathrm{f}}(\mathbf{u}, \nabla\mathbf{u}) = \frac{\tau_W(\mathbf{u}, \nabla\mathbf{u})}{\frac{1}{2}\rho_\infty v^2_\infty}, \tag{13}$$

where $\tau_W = -(\underline{\tau}\,\mathbf{n}) \cdot \mathbf{t}$ denotes the viscous stress at the wall, and $\underline{\tau} = \mu S$ is the viscous stress tensor defined in Section 3.3. Here, we recall that $\mathbf{n}$ is the unit outward normal vector to the boundary Γ of the (fluid dynamics) computational domain Ω; thus the vector $\mathbf{n}$ points into the

airfoil. Furthermore, $\mathbf{t}$ denotes the unit tangential vector to Γ parallel to the flow direction off the wall and directed such that $\mathbf{t} \cdot \mathbf{v}_\infty > 0$. Similar to evaluating the force coefficients (11) based on (12), the local c_{p}- and c_{f}-values (13) are evaluated based on (cf. [14])

$$c_{\mathrm{p},h} = c_{\mathrm{p},h}(\mathbf{u}_h) = \frac{\hat{\mathbf{h}}_{\Gamma,h} \cdot \tilde{\mathbf{n}} - p_\infty}{\frac{1}{2}\rho_\infty v_\infty^2}, \qquad c_{\mathrm{f},h} = c_{\mathrm{f},h}(\mathbf{u}_h, \nabla \mathbf{u}_h) = -\frac{\left(\hat{\underline{\sigma}}_{\Gamma,h}\mathbf{n}\right) \cdot \tilde{\mathbf{t}}}{\frac{1}{2}\rho_\infty v_\infty^2}, \tag{14}$$

with $\tilde{\mathbf{n}} = (0, n_1, n_2, n_3, 0)^\top$ for $\mathbf{n} = (n_1, n_2, n_3)^\top$, and $\tilde{\mathbf{t}} = (0, t_1, t_2, t_3, 0)^\top$ for $\mathbf{t} = (t_1, t_2, t_3)^\top$.

4 APPLICATION TO THE DLR-F11 HIGH LIFT CONFIGURATION

In the 4th International Workshop on High-Order CFD Methods (HioCFD-4), the DLR-F11 high lift configuration is considered a "computational/meshing challenge". This configuration has already been extensively analyzed for various configuration detail levels in the 2nd AIAA High Lift Prediction Workshop (HiLiftPW-2). In HioCFD-4, the so-called *Config 4* detail level is considered which includes slat tracks and flap track fairings but no slat pressure tube bundles. This configuration is considered at Mach number $M = 0.175$, Reynolds number $Re = 15.1 \cdot 10^6$ based on the mean aerodynamic chord (MAC) of 347.09mm, and angle of attack $\alpha = 7°$. These are the flow conditions of one of the high Reynolds number cases of Case 2b in HiLiftPW-2.

4.1 Linear and quadratic mesh generation process

For the DLR-F11 configuration under consideration, first a linear hybrid mesh has been generated using the CENTAUR grid generator. To reduce the number of surface faces, the curvature clustering was lowered to 6 compared to a typical value of 12-20. The curvature clustering represents the number of points used to model a full circle. To further reduce the number of surface faces, the faces on the wing and high lift surfaces were stretched in the spanwise direction by a factor of 3 near the leading and trailing edges. To properly model the boundary layer region, a 20-layer prismatic mesh was used with the initial layer thickness set to 0.003mm and the stretching set to 1.44. To decrease the number of tetrahedra generated, a tetrahedral stretching value of 2.15 was used. Sources were also used to locally add additional elements where needed to model complex geometric regions like the intersection between the slat and slat tracks.

With 2 365 919 prisms, 42 603 pyramids and 1 116 213 tetrahedra, in total 3 524 735 elements and 1 427 392 nodes, the linear mesh is rather coarse considering the complexity of the configuration. In a second step, also using the CENTAUR grid generator, the wall boundary faces of this mesh have been curved to create a quadratic, 3rd-order boundary approximation, and boundary as well as interior elements have been curved in order to avoid crossover of element faces. The high-order mesh contains the same number of elements and has 11 237 409 nodes. Figure 2 shows the quadratic mesh for the DLR-F11 configuration including curved surface faces on the fuselage. Figure 3 shows the surface mesh for both the linear and quadratic meshes in the region of the outboard flap track fairing. Note the curved surface elements on the trailing edge of the flap track fairing and its lower surface. Figure 4 shows a cut through both the linear and the quadratic hybrid meshes around the slat. The quadratic mesh clearly shows the curved prisms in the boundary layer. Both the first layer elements and elements farther away from the surface have been curved to ensure a good quality mesh in this region. Finally, Figure 5 shows a cut through the prism layers around the slat and leading edge of the main wing and their connection to the tetrahedral farfield mesh. Note the relatively smooth surface of the slat and main wing despite the relatively coarse surface mesh in this region of high curvature.

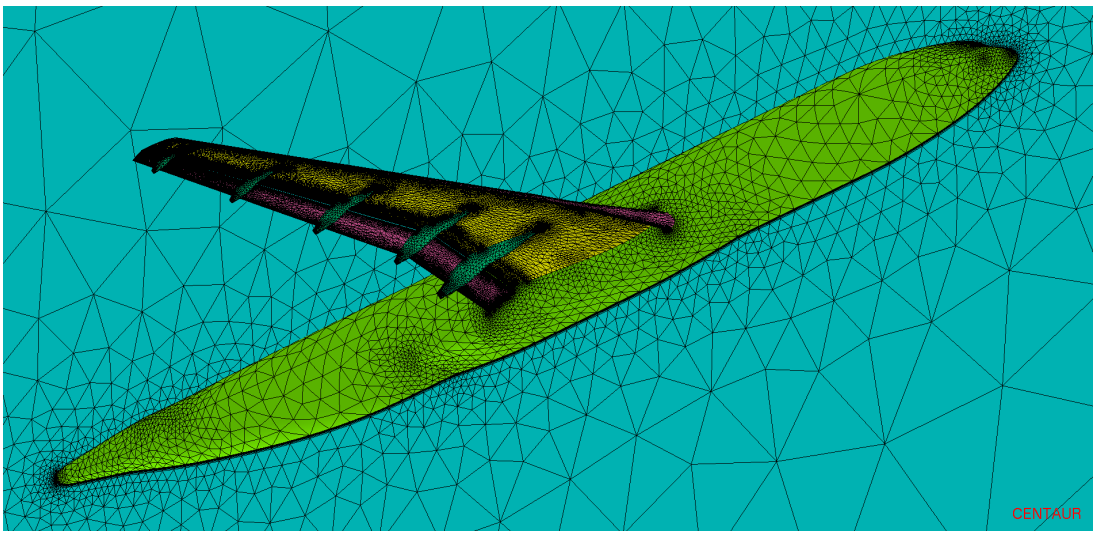
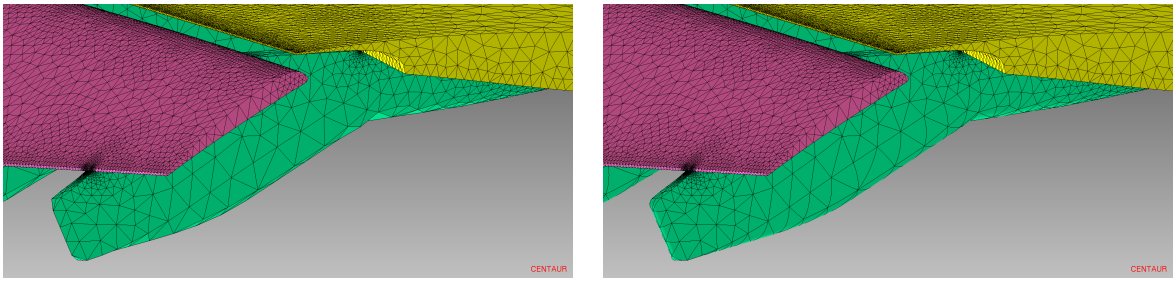
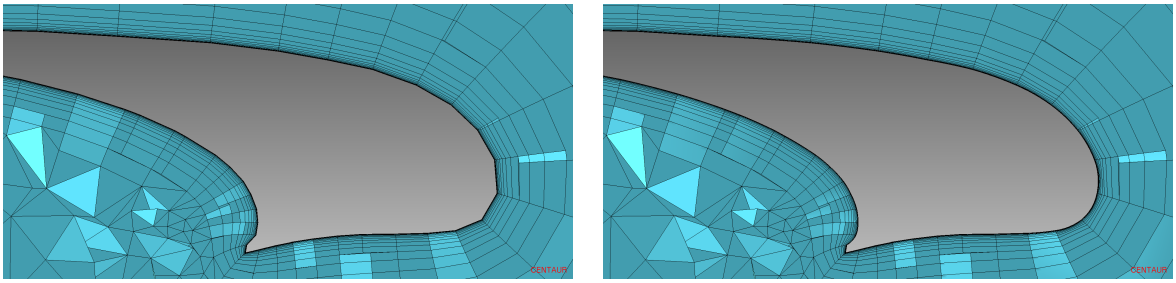


Figure 2: Quadratic mesh for the DLR-F11 configuration.



(a) (b)

Figure 3: Outboard flap track fairing surface mesh: (a) the linear mesh, and (b) the quadratic mesh.



(a) (b)

Figure 4: Hybrid mesh cuts in the slat region: (a) the linear mesh, and (b) the quadratic mesh.

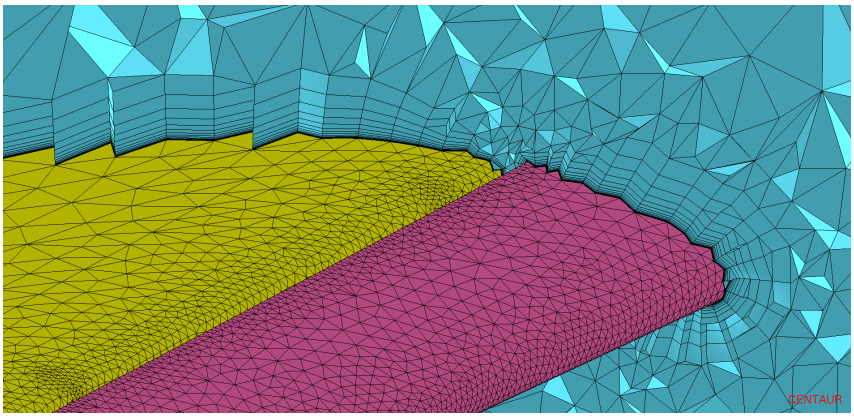


Figure 5: Cut throught the prism layers around the slat and the leading edge of the main wing.

4.2 Flow solution process

The DG solver PADGE [12] employs the discretization of the RANS and Wilcox-$k\omega$ equations as described in Section (3). The discrete equations are solved fully implicitly using a linearized Backward-Euler method. The CFL number of this pseudo-time iteration method starts with a low value and increases as the nonlinear residual decreases. Divergent iteration steps are (automatically) repeated with half the CFL number or even with a lower factor if required. As the solution continues to converge the CFL reduction factor is subsequently increased to unity again. This way, the solver can overcome difficult stages in the solution process like the transient phase but still benefits from (relatively) high CFL numbers in the final stage of the solution process. The linear systems of the implicit iteration scheme are solved using the GMRES method and the ILU(0) preconditioner as offered by the PETSc library [2] which runs within a block-Jacobi iteration in parallel. The Jacobian matrices required in the implicit scheme are hand-differentiated and complete in the sense that the derivatives of all terms are included and all equations are fully coupled.

On the mesh generated as described in Sections 2 and 4.1 we first solve the DG discretization (2) for the polynomial degree $p = 0$, and subsequently for $p = 1$ and $p = 2$. The solver for the $p = 0$ solution starts from freestream values, the solver for each of the $p > 0$ solutions starts from the converged $p-1$ solution. With this p-sequencing the solver for each polynomial degree has a sufficiently good initial guess to start the solution process with. We choose the number of quadrature points per elements depending on the polynomial degree. For a polynomial degree p we use $p+2$ Gaussian quadrature points in each dimension, i. e., $(p+2)^3$ per element. Before the PADGE solver starts solving it checks the regularity of the mesh by evaluating the Jacobian determinants[1] at the quadrature points for the particular polynomial degree. As the degree increases the number of quadrature points per element increases and the coverage of the element by checking the determinants in quadrature points is increased. Thus, a curved mesh which might occur "discretely regular" for one specific degree (and thus quadrature formula) might turn out to be irregular if checked at the quadrature points involved for a higher polynomial degree.

degree	# DoFs/eqn	# quadrature points/elem.	# irregular elements	worst # neg. Jac./element
0	3 524 735	8	0	0/8
1	14 098 940	27	0	0/27
2	35 247 350	64	14	2/64

Table 1: PADGE's "discrete regularity" check of the mesh (with 3 524 735 elements) for $p = 0$, $p = 1$, and $p = 2$.

The related data is collected in Table 1 for $p = 0$, $p = 1$, and $p = 2$. For each of the polynomial degrees it includes the number of degrees of freedoms per equation, the number of quadrature points per element, the number of irregular elements when checking the Jacobian determinants in the quadrature points, and finally the worst/highest number of quadrature points per element at which the Jacobian determinant is negative. Here, we see that the mesh is "discretely regular" when checking in the quadrature points of the discretization of polynomial degrees $p = 0$ and $p = 1$ as in each of the quadrature points involved the Jacobian determinant

[1] Jacobian determinant: The determinant of the derivative of the mapping function from the reference element to the element in physical space.

is positive. Furthermore, we see that for $p = 2$ there are 14 irregular elements. However, in these elements the Jacobian determinants are negative in at most 2 out of 64 quadrature points. Depending on the sensitivity of the discretization to irregular elements this "slightly irregular" mesh might be acceptable to the flow solver or not.

On this mesh we computed $p = 0$, $p = 1$ and $p = 2$ (i. e., 1st-, 2nd-, and 3rd-order) flow solutions. For each of the polynomial degrees the (l2-norm of the) nonlinear residual (vector) is reduced by a factor of 10^{-10} relative to the freestream residual for that polynomial degree. Then the flow solution is transferred to (embedded into) the function space of the next higher polynomial degree, and the solution process is continued. Figure 6 plots the convergence history

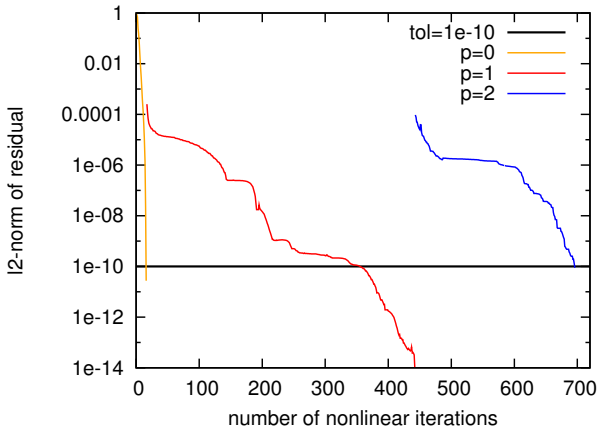


Figure 6: DLR-F11, Config 4: $M = 0.175$, $Re = 15.1 \cdot 10^6$, $\alpha = 7°$: Solver convergence of the nonlinear residual vs. the number of fully implicit iteration steps for the $p = 0$, $p = 1$, and $p = 2$ flow solutions.

of the (l2-norm of the) nonlinear residual (vector) normalized by the freestream residual for each polynomial degree. The $p = 0$ solution for which several of the physical and discretization terms vanish is converged in a very small number (17) of iterations steps. The $p = 1$ solver does not start from freestream (corresponding to a normalized residual of one) but from a lower residual as it uses the $p = 0$ solution as initial solution. The $p = 1$ solution requires a relatively large number (339) of iteration steps. This is a typical effect in p-sequencing on fine meshes where the $p = 1$ solver starts with a rather bad $p = 0$ solution as initial guess. In the convergence history of the $p = 1$ solution there are several stages where the convergence rate is reduced to almost zero. Here, the solver experienced problems and decided to recompute some iteration steps with a decreased CFL number. Given that in this computation 45 out of the 339 iteration steps diverged/failed and needed to be recomputed the initial CFL number and the settings of its subsequent increase seemed to be chosen too high. Even close to the 10^{-10} convergence criterion the solver experienced problems which is why the computation had been restarted to check whether the nonlinear residual would converge below 10^{-14}, and it actually did after 87 further steps 18 of which were recomputed. Overall, the CFL number in the $p = 1$ solution process started with 5 and increased to 500 until residual 10^{-10} and over 40 000 until residual 10^{-14}. Finally, the $p = 2$ solver took the $p = 1$ solution as initial guess and converged the nonlinear residual below the prescribed 10^{-10} tolerance. Here, only 24 out of 254 nonlinear iteration steps failed and needed to be recomputed with a decreased CFL number. Here, the CFL number started at 10 and increased up to 200.

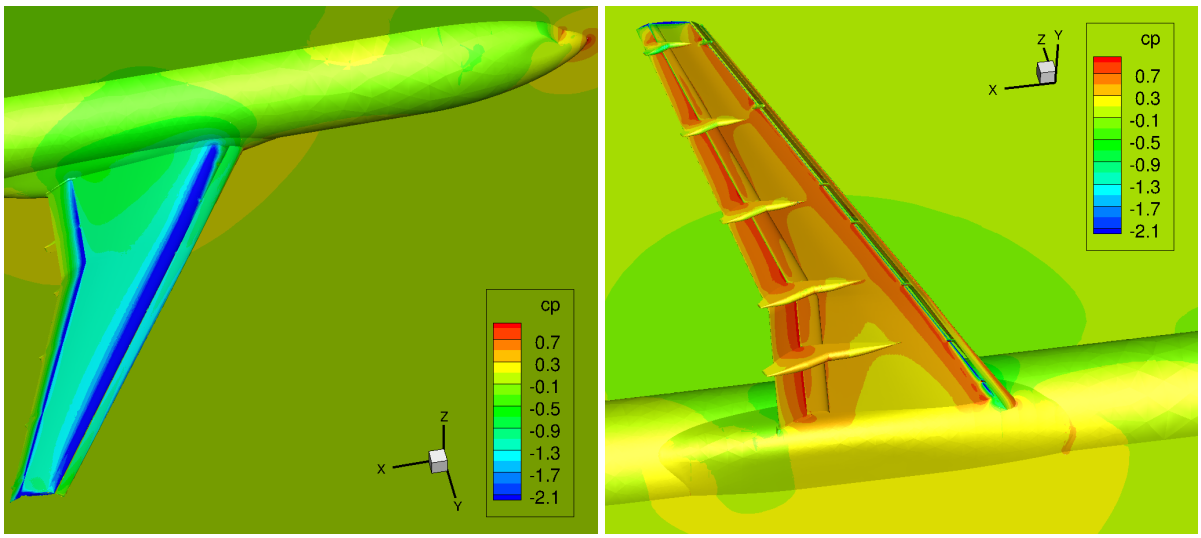
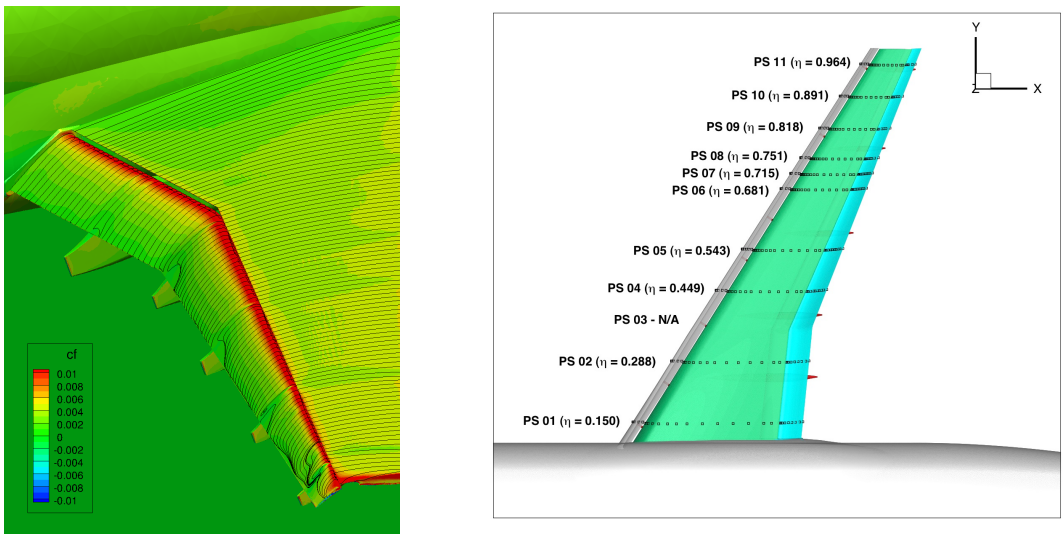


Figure 7: DLR-F11, Config 4: $M = 0.175$, $Re = 15.1 \cdot 10^6$, $\alpha = 7°$: c_p-distributions of the 3rd-order DG solution viewed from above and below the configuration.



(a) (b)

Figure 8: DLR-F11, Config 4: $M = 0.175$, $Re = 15.1 \cdot 10^6$, $\alpha = 7°$: (a) c_f-distribution and streamlines of the 3rd-order DG solution on the main wing and flap. (b) Pressure tap locations in the experiment (figure from [1]).

The c_p-distribution of the resulting $p = 2$ flow solution is shown in Figure 7. Here, we recognize the complexity of the flow which passes the slat, the slat tracks, the upper wing or the lower wing with the flap track fairings, and the flap. Furthermore, Figure 8(a) shows the c_f-distribution and $\mathbf{c}_f$-streamlines zoomed in on the flap and parts of the main wing. Here, we see that the flow stays attached on the main wing as well as on most of the flap. However, close to the flap track fairings (located below the flap) there are small separation regions on the upper side of the flap. Furthermore, there is a larger separation region on the flap near its tip.

Figure 8(b) shows the location of the pressure taps in the experimental setup. Slices of the c_p-distribution at the pressure tap locations PS1, 4, 6, and 10 (at the y-positions $\eta = 0.150$, 0.449, 0.681, and 0.891, respectively) are shown in Figure 9 compared to the experimental data at similar flow conditions. Here, we see that the c_p-distribution matches very well that of the experiments, not only on the main wing but also on the flap and the slat.

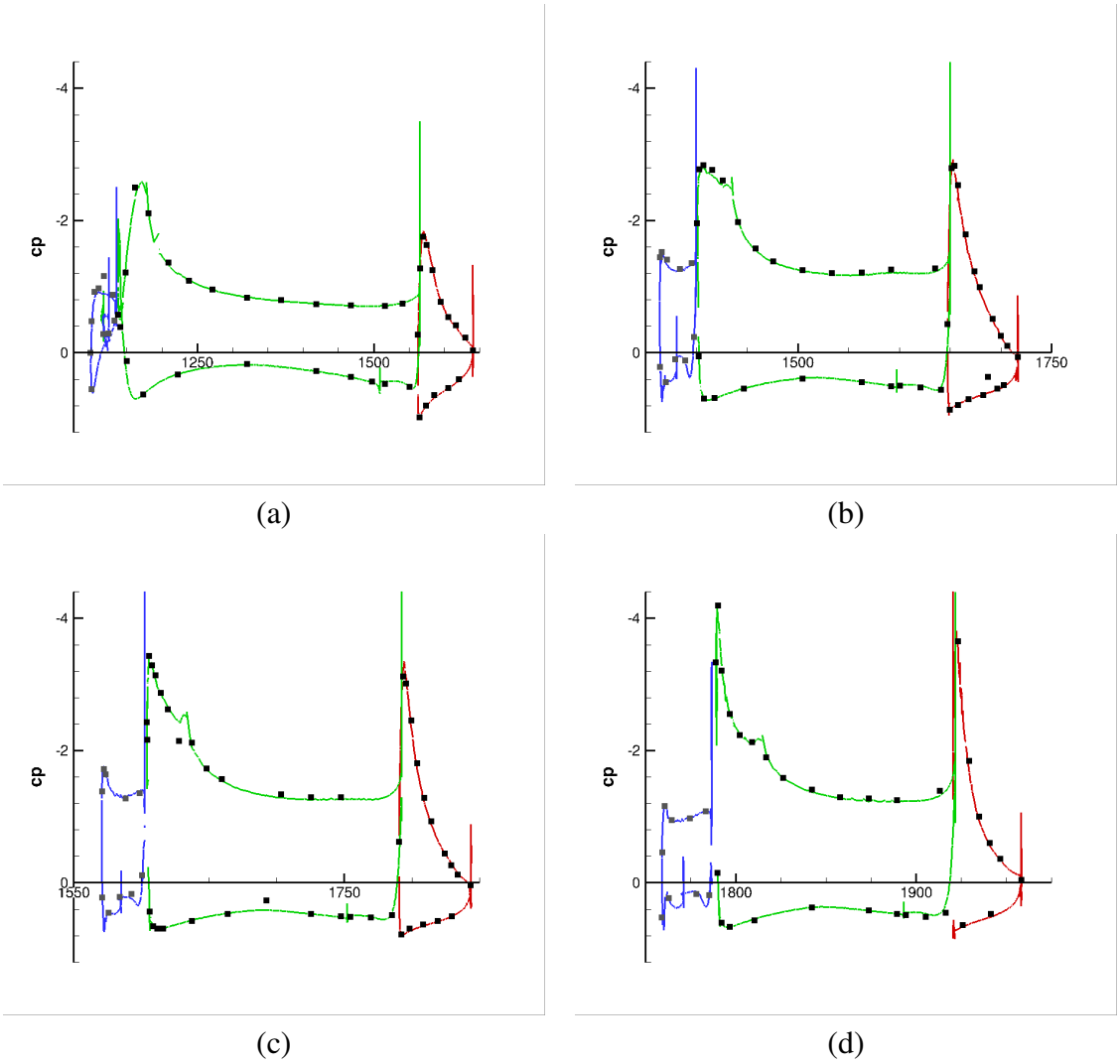


Figure 9: DLR-F11, Config 4: $M = 0.175$, $Re = 15.1 \cdot 10^6$, $\alpha = 7°$: c_p-distributions of the 3rd-order DG solution compared to experimental data (at $M = 0.176$, $Re = 15.1 \cdot 10^6$, $\alpha = 7.03938$) [1]. c_p-slices at pressure tap locations (a) PS01 at $\eta = 0.150$, (b) PS04 at $\eta = 0.449$, (c) PS6 at $\eta = 0.681$, and (d) PS10 at $\eta = 0.891$.

Finally, Table 2 collects the integrated forces. Here we see that the force coefficient values for the lift, drag and moment coefficients, C_L, C_D, and C_M, are quite close to the experimental data. Furthermore, they are in a range similar to that obtained in RANS-SA computations during the HiLiftPW-2, with a slightly larger deviation in the moment but closer to the experiments in the drag coefficient.

5 CONCLUSIONS AND OUTLOOK

The computational/meshing challenge of the DLR-F11 configuration which is "close to industrial needs" was considered to improve and challenge current high-order mesh generation and solver techniques. For this complex geometry, a quadratic curved mesh has been generated using the CENTAUR grid generator [5]. On this 3rd-order mesh a fully turbulent 3rd-order DG solution to the RANS and Wilcox-$k\omega$ turbulence equations has been obtained using the fully implicit DLR-PADGE solver [12]. The numerical results compare very well to experiments, with particularly well matching c_p-distributions, and good approximations were obtained to global

	C_L	C_D	C_M
Exp.	1.9270	0.1615	-0.5390
TAU*	1.8794	0.1681	-0.5647
	-2.5%	+4.1%	-4.8%
PADGE	1.8781	0.1649	-0.5704
	-2.5%	+2.1%	-5.8%

Table 2: DLR-F11, Config 4: $M = 0.175$, $Re = 15.1 \cdot 10^6$, $\alpha = 7°$: Force coefficients of the $p = 2$ flow solution compared to HiLiftPW-2 results (* Rudnik, Melber-Wilkending (DLR), RANS-SAO, 2013) and experiments [1].

force coefficients with deviations in the range of previous HiLiftPW-2 results.

This work can be considered as one of possibly many steps towards offering a complete and automated chain of high-order grid generation and high-order aerodynamic CFD solver technology. At the beginning of this work curved grids were generated for this complex geometry which included a large number (> 100) of heavily irregular elements. In contrast to that the grid on which a 3rd-order solution could finally be computed had only a few (in the range of 10) elements left which were "slightly irregular" only. Given this it is clear that the high-order grid generation took a significant step forward. Nevertheless, further steps are required in this direction. First of all, the final goal would be that all elements in a high-order mesh are regular; in a first step not necessarily "regular" as can be analyzed with a validity check based on Bézier functions [17], but "discretely regular" in the sense that the Jacobian determinants are positive at least in each of the quadrature points involved in a (quadrature-based) discretization of a specific polynomial degree. Secondly, meshes which are generated to be used with high-order methods, and which will be taken as starting point for local mesh (or hp-)refinement [18, 21] should be as coarse as possible but clearly still sufficiently fine to capture the geometry. On this fine line of ostensibly conflicting requirements (coarse but still sufficiently fine) it is particular important that mesh generation capabilities are stable and deliver high quality elements even in case of particularly coarse meshes. Finally, curved grid generation should not stop at quadratic boundary approximation. The difference between using quadratic curved meshes as compared to using linear (straight-sided) meshes is quite large and can be expected to be larger than the difference between e. g., a quadratic and a cubic boundary approximation. Nevertheless, a boundary approximation higher than quadratic – even though introducing additional complexity in the grid generation process – would increase the quality of the boundary approximation in case of a fixed mesh size or would allow coarser meshes with a similar boundary approximation quality.

No doubt, the complexity of the configuration considered stressed the PADGE solver [12] close to its current limits. Using a discretization on the boundary which resembles the discretization employed on interior faces as close as possible, and thus introducing a level of numerical diffusion which is equal/similar throughout the domain until (and including) the boundary, is – to our experience [15] – an essential ingredient to stabilize flow computations on geometries of this complexity. With an appropriate discretization of force coefficients the particular discretization employed on the boundary does not break adjoint consistency [14], in which case corresponding discrete adjoint solutions can be expected to be smooth. Furthermore, evaluating local quantities like surface pressure and skin-friction at the wall accordingly, allows to obtain accurate and smooth c_p- and c_f-distributions [14] even in our case of using numerical flux functions at the wall boundary. A fully implicit solver has been used, with solver recovery techniques enabled which recompute divergent/failed steps with a lower CFL number. Further

development of the high-order discretization and solvers are still required. While PADGE can be considered a research code, not particularly trimmed to efficiency or scalability, high-order DG methods are currently being implemented alongside cell-centered 2nd-order Finite Volume methods into the new DLR flow solver Flucs dedicated (amongst others) to future industrial use. Significant effort will be required to make high-order DG methods readily available to industry for configurations of this complexity, at cruise as well as at high lift conditions.

REFERENCES

[1] 2nd AIAA CFD High Lift Prediction workshop. http://hiliftpw.larc.nasa.gov/index-workshop2.html.

[2] S. Balay, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. PETSc Web page, 2001. http://www.mcs.anl.gov/petsc.

[3] F. Bassi, A. Crivellini, S. Rebay, and M. Savini. Discontinuous Galerkin solution of the Reynolds-averaged Navier-Stokes and $k-\omega$ turbulence model equations. *Computers & Fluids*, 34:507–540, 2005.

[4] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Num. Meth. Part. Diff. Eq.*, 16(4):365–378, 2000.

[5] CentaurSoft. CENTAUR software. Computational grids for Advanced Modelling. www.centaursoft.com.

[6] M. Ceze and K. Fidkowski. Drag prediction using adaptive discontinuous finite elements. 51st AIAA Aerospace Sciences Meeting, Jan. 2013, AIAA 2013-0051, 2013.

[7] C. Geuzaine, A. Johnen, J. Lambrechts, J.-F. Remacle, and T. Toulorge. The generation of curvilinear meshes. In H. Deconinck and R. Abgrall, editors, *VKI LS 2014-03: 37th Advanced CFD Lecture Series: Recent developments in higher order methods and industrial application in aeronautics, Dec. 9-12, 2013*. Von Karman Institute for Fluid Dynamics, Rhode Saint Genèse, Belgium, 2014.

[8] A. Harten and J. Hyman. Self adjusting grid methods for one-dimensional hyperbolic conservation laws. *J. Comput. Phys.*, 50(2):416–433, 1983.

[9] R. Hartmann. Higher-order and adaptive discontinuous Galerkin methods with shock-capturing applied to transonic turbulent delta wing flow. *Int. J. Numer. Meth. Fluids*, 72(8):883–894, 2013.

[10] R. Hartmann. Results for the C3.5 test case with the DLR-PADGE code. 3rd International Worskhop on High-Order CFD methods, Kissimmee, Florida, Jan 3-4, 2015. https://www.grc.nasa.gov/hiocfd/wp-content/uploads/sites/22/C3.5_DLR.pdf, 2015.

[11] R. Hartmann, J. Held, and T. Leicht. Adjoint-based error estimation and adaptive mesh refinement for the RANS and k-ω turbulence model equations. *J. Comput. Phys.*, 230(11):4268–4284, 2011.

[12] R. Hartmann, J. Held, T. Leicht, and F. Prill. Discontinuous Galerkin methods for computational aerodynamics – 3D adaptive flow simulation with the DLR PADGE code. *Aerosp. Sci. Technol.*, 14:512–519, 2010.

[13] R. Hartmann and T. Leicht. Higher order and adaptive DG methods for compressible flows. In H. Deconinck and R. Abgrall, editors, *VKI LS 2014-03: 37th Advanced CFD Lecture Series: Recent developments in higher order methods and industrial application in aeronautics, Dec. 9-12, 2013*. Von Karman Institute for Fluid Dynamics, Rhode Saint Genèse, Belgium, 2014.

[14] R. Hartmann and T. Leicht. Generalized adjoint consistent treatment of wall boundary conditions for compressible flows. *J. Comput. Phys.*, 300:754–778, 2015.

[15] R. Hartmann and T. Leicht. Generation of unstructured curvilinear grids and high-order discontinuous Galerkin discretization applied to a 3D high-lift configuration. *Int. J. Num. Meth. Fluids*, 2016. published online. DOI: 10.1002/fld.4219.

[16] 4th International Worskhop on High-Order CFD methods, Heraklion (Crete), Greece, June 4-5, 2016. http://how4.cenaero.be/.

[17] A. Johnen, J.-F. Remacle, and C. Geuzaine. Geometrical validity of curvilinear finite elements. *J. Comput. Phys.*, 233:359–272, 2013.

[18] Y. Kallinderis, A. Karkoulias, and P. Antonellis. Adaptive high-order hybrid grids for field computations. In *8th Int. Congress on Computational Mechanics, 12-15 July 2015, Volos, Greece*, 2015.

[19] N. Kroll, H. Bieler, H. Deconinck, V. Couallier, H. van der Ven, and K. Sorensen, editors. *ADIGMA - A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*, volume 113 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*. Springer, 2010.

[20] N. Kroll, C. Hirsch, F. Bassi, C. Johnston, and K. Hillewaert, editors. *IDIHOM - Industrialization of High-Order Methods - A Top Down Approach*, volume 128 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*. Springer, 2015.

[21] T. Leicht and R. Hartmann. Error estimation and hp–adaptive mesh refinement for discontinuous Galerkin methods. In Z. J. Wang, editor, *Adaptive High-Order Methods in Computational Fluid Dynamics*, volume 2 of *Advances in Computational Fluid Dynamics*, chapter 3, pages 67–94. World Science Books, 2011.

[22] O. Sahni, X. Luo, K. Jansen, and M. Shephard. Curved boundary layer meshing for adaptive viscous flow simulations. *Finite Elements in Analysis and Design*, 46(1-2):132–139, 2010.

[23] S. Schoenawa and R. Hartmann. Discontinuous Galerkin discretization of the Reynolds-averaged Navier-Stokes equations with the shear-stress transport model. *J. Comput. Phys.*, 262:194–216, 2014.

[24] M. Wallraff, T. Leicht, and M. Lange-Hegermann. Numerical flux functions for Reynolds-averaged Navier-Stokes and $k\omega$ turbulence model computations with a line-preconditioned p-multigrid discontinuous Galerkin solver. *Int. J. Numer. Meth. Fluids*, 71(8):1055–1072, 2013.

[25] Z. J. Wang, K. Fidkowski, R. Abgrall, F. Bassi, D. Caraeni, A. Cary, H. Deconinck, R. Hartmann, K. Hillewaert, H. T. Huynh, N. Kroll, G. May, P.-O. Persson, B. van Leer, and M. Visbal. High-order CFD methods: current status and perspective. *Int. J. Num. Meth. Fluids*, 72(8):811–845, 2013.

[26] D. C. Wilcox. Reassessment of the scale-determining equation for advanced turbulence models. *AIAA J.*, 26(11):1299–1310, 1988. 1988.

[27] D. C. Wilcox. *Turbulence Modeling for CFD*. DCW Industries, Inc., La Canada CA, 1993.

[28] Z. Q. Xie, R. Sevilla, O. Hassan, and K. Morgan. The generation of arbitrary order curved meshes for 3D finite element analysis. *Computational Mechanics*, 51(3):361–374, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

BS1 - DNS OF THE TAYLOR-GREEN VORTEX AT RE $= 1600$

A. Mastellone[1], L. Cutrone[1] and F. Capuano[2]

[1]Centro Italiano per le Ricerche Aerospaziali (CIRA)
Via Maiorise, I-81043 Capua (Italy)
e-mail: a.mastellone@cira.it

[2] Dipartimento di Ingegneria Industriale, Università di Napoli "Federico II"
Via Claudio 21 - I-80125 Napoli (Italy)
e-mail: francesco.capuano@unina.it

Keywords: LES. High-order schemes. Turbulence.

Abstract. *SPARK-LES is a Large-Eddy Simulation (LES) code currently under development at the Italian Aerospace Reseacrh Center (CIRA). It solves the fully compressible Navier-Stokes equations discretized on multi-block, structured grids according to the Finite-Volume (FV) method, using different high-order schemes. In this work a detailed comparison between the schemes is presented in the simulation of a Taylor-Green Vortex in order to assess their accuracy and performances. Temporal behaviour of kinetic energy decay rate and enstrophy is given at different mesh resolutions, showing a good agreeement with DNS results at finer grids. Analysis of power spectra confirms the universal energy cascade behavior.*

1 Code description

SPARK-LES is a Large-Eddy Simulation (LES) code currently under development at CIRA in the framework of the HYPROB Program, funded by the Italian Ministry of Research [1]. The code shall be an advanced tool with high-fidelity numerical methods and state-of-the-art modelling, and able to reproduce the operation of a liquid-rocket thrust chamber at high pressures. This feature will support design of critical components of the chamber, like injectors, cooling channels, etc; and analysis of unsteady phenomena of combustion instabilities.

SPARK-LES solves the fully compressible Navier-Stokes equations discretized on multi-block, structured grids according to the Finite-Volume (FV) method. Explicit second- and fourth-order (2E, 4E), as well as fourth- and sixth-order compact (4C, 6C) cartesian-like operators are available for reconstruction of convective fluxes, whereas an explicit second-order scheme is used for diffusive fluxes [2]. The general interpolation operator can be expressed, for a generic variable ϕ, as

$$\alpha\tilde{\phi}_{i-3/2} + \tilde{\phi}_{i-1/2} + \alpha\tilde{\phi}_{i+1/2} = \sum_{j=1}^{N_C} \gamma_j \left(\bar{\phi}_{i-j} + \bar{\phi}_{i+j-1}\right), \tag{1}$$

where $\tilde{\phi}$ and $\bar{\phi}$ are the face- and cell-averaged values of ϕ, respectively.

The tridiagonal system arising from Eq. (1) when $\alpha \neq 0$ is solved by means of a Thomas algorithm. In the case of a multi-block domain, an *overlap* method is adopted [3]: for each block, an enlarged linear system is solved by overlapping a certain number of neighbour cells from adjacent blocks (in this study 4 cells are used, which are a good tradeoff between parallel communication overhead and scheme accuracy). The boundary points at block-to-block interfaces are evaluated by a high-order centered approximation formula.

Time-advancement is obtained by means of explicit Runge-Kutta schemes of arbitrary number of stages. SPARK-LES features an MPI parallel implementation based on the multi-block partitioning of the computational domain. The MPI calls have been carefully profiled in order to minimize communication overhead. The code makes full use of Fortran 90 capabilities, in terms of dynamic allocation, highly modular architecture and massive use of pointers for CPU efficiency and memory usage optimization.

2 Case summary

The Taylor-Green Vortex problem is defined on a triperiodic cube with sides of 2π, $L = 1$ and the following initial conditions

$$u_x(x,y,z) = U_0 \sin(x/L)\cos(y/L)\cos(z/L), \tag{2}$$

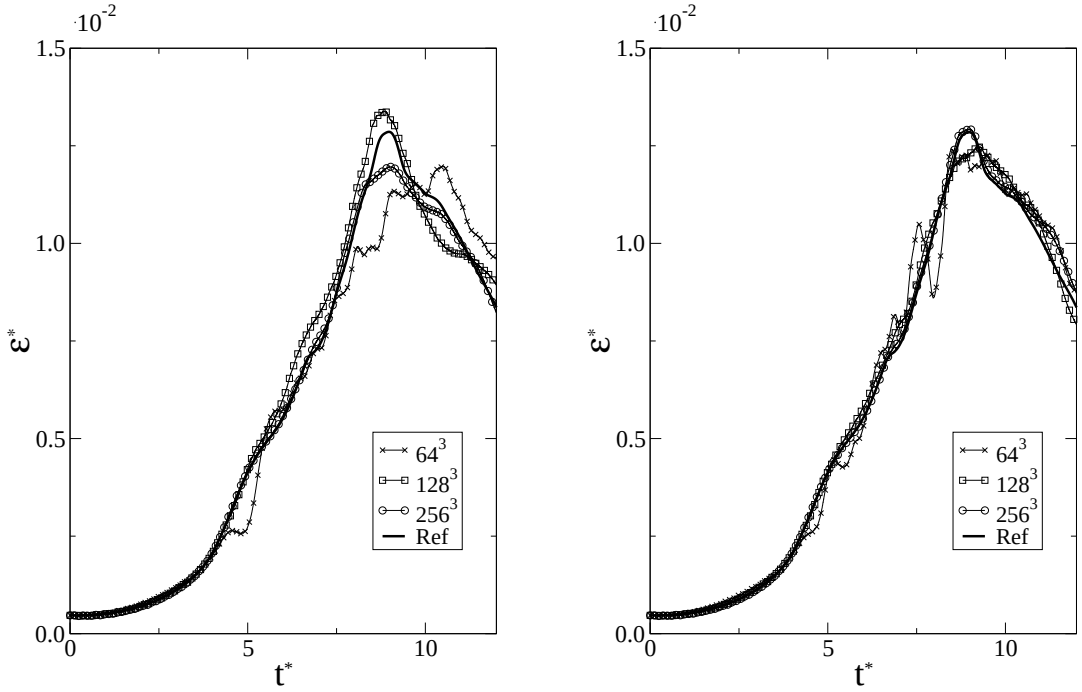
$$u_y(x,y,z) = -U_0 \cos(x/L)\sin(y/L)\cos(z/L), \tag{3}$$

$$u_z(x,y,z) = 0, \tag{4}$$

$$p(x,y,z) = p_0 + \frac{\rho_0 U_0^2}{16}\left[\cos(2x/L) + \cos(2y/L)\right]\left[\cos(2z/L) + 2\right] \tag{5}$$

As time advances, the initial distribution of vorticity is subject to vortex-stretching, thus generating small scales and eventually causing the vortices to break into turbulence. Since there is no forcing to sustain the turbulent motion, a decay is observed after transition. The incompressible problem is entirely governed by the Reynolds number $Re = \frac{\rho U_0 L}{\mu}$, equal to 1600. However,

Figure 1: Kinetic energy dissipation rate as a function of the non-dimensional time t^* for second-order (left) and fourth-order (right) explicit schemes.



since the code solves the compressible flow equations, the other dimensional parameters have been set to yield a nearly incompressible condition, i.e. $M = 0.1$. The ideal gas equation of state is used, along with a constant Prandtl number $Pr = 0.71$. The domain has been discretized by a uniform mesh of increasing resolution, 64^3, 128^3 and 256^3 cells, divided in 64 equally-sized blocks. The meshes are regular cartesian grids and have been generated by an in-house tool. In this study, explicit second- and fourth-order and compact fourth- and sixth-order schemes are used. The coefficients of Eq. (1) are $\gamma_1 = 1/2$ for the explicit second-order scheme ($\alpha = 0$, $N_C = 1$), $\gamma_1 = 7/12$ and $\gamma_2 = -1/12$ for the explicit fourth-order scheme ($\alpha = 0$, $N_C = 2$), $\gamma_1 = 3/4$ for the fourth-order compact scheme ($\alpha = 1/4$, $N_C = 1$), and $\gamma_1 = 29/36$ and $\gamma_2 = 1/36$ for the sixth-order compact scheme ($\alpha = 1/3$, $N_C = 2$). Time-advancement is performed using a third-order Runge-Kutta scheme with CFL $= 0.6$. No artificial dissipation or filters of any type are used. The computations were run on the CIRA cluster FLAME, equipped with Intel Xeon E5-2680 @2.7 GHz processors. All simulations have been carried over 64 MPI nodes, assigning one block to each node. Results from TauBench runs gave an average time of 7.344 s.

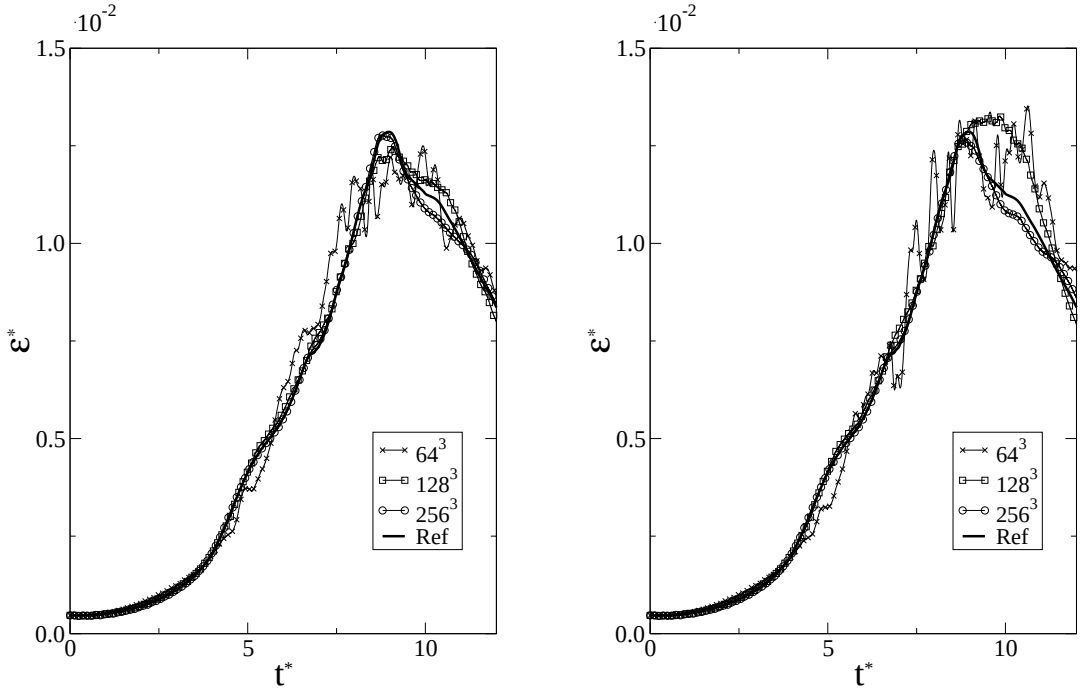
3 Results

Figures 1 and 2 show the time-evolution of the global kinetic energy dissipation rate

$$\varepsilon = \frac{d}{dt}\frac{1}{\rho_0 \Omega}\int_\Omega \rho \frac{u_i u_i}{2} d\Omega \tag{6}$$

for different schemes and for three grid resolutions. The time t^* is adimensionalized by means of L and U_0. A good grid convergence is observed in all cases. The finest grid is in excellent agreement with the reference solution. The coarse grid shows an oscillatory behavior probably

Figure 2: Kinetic energy dissipation rate as a function of the non-dimensional time t^* for fourth-order (left) and sixth-order (right) compact schemes.



due to inadequate resolution as the flow undergoes creation of smaller scales. The integrated enstrophy

$$\zeta = \frac{1}{\rho_0 \Omega} \int_\Omega \rho \frac{\omega_i \omega_i}{2} d\Omega \tag{7}$$

is also considered: this is as an indirect measure of the dissipation rate, through the relation $\varepsilon_\zeta = 2\frac{\mu}{\rho_0}\zeta$. The time-evolution of ε_ζ is shown in Figs. 3 and 4. For the finest resolution, good agreement is observed also for the enstrophy-based dissipation rate, hence no significant spurious dissipation is added to the solution and turbulent scales are well resolved by the numerical scheme. At lower resolutions, the enstrophy-based dissipation rate is lower than the kinetic energy dissipation rate, suggesting that there are numerical dissipation sources [4].

Contours of the vorticity norm on a subset of the plane $x = -\pi L$ at $t^* = 8$ are shown in Figures 5 and 6. While the lowest grid resolution captures only the concentration of vorticity, the main structures of the shear layer are described increasingly well as the mesh is refined. The result on 256^3 cells shows a satisfactory agreement with the reference solution [5]. Although the presence of spurious secondary structures is still observed, the vorticity contours are well defined for high order (> 2) schemes, showing adequate resolution.

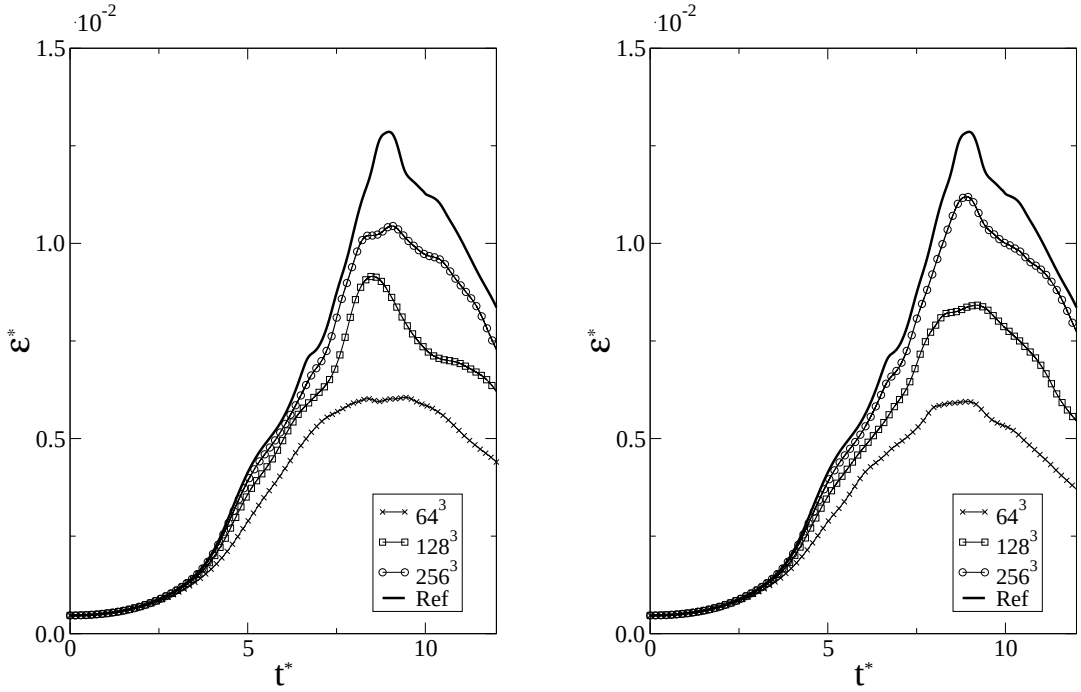
Kinetic energy spectra are defined as

$$P(k) = \frac{1}{2}\left(\tilde{u}_x \tilde{u}_x^* + \tilde{u}_y \tilde{u}_y^* + \tilde{u}_z \tilde{u}_z^*\right), \quad k = \sqrt{k_x^2 + k_y^2 + k_z^2} \tag{8}$$

being $\tilde{u}_i$ ($i \in \{x, y, z\}$) the 3D discrete Fourier transform of the i-th velocity component u_i.

$$\tilde{u}_i(k_x, k_y, k_z) = \frac{1}{N^3} \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \sum_{r=0}^{N-1} u_i(x_p, y_q, z_r) \cdot \exp\left(-ik_x x_p\right) \exp\left(-ik_y y_q\right) \exp\left(-ik_z z_r\right) \tag{9}$$

Figure 3: Enstrophy-based dissipation rate as a function of the non-dimensional time t^* for second-order (left) and fourth-order (right) explicit schemes.



and the wavenumbers are limited by the condition $|k| < \pi N/L$, being N^3 the number of cells in the domain. They give a precise idea of the energy scale behavior. Figures 7 and 8 show that the spectra are substantially independent of the scheme giving the expected universal behavior at finer grids. The lowest resolution shows significant pile-up of energy at higher wavenumbers due to the accumulation of aliasing errors. The pile-up disappears as the number of mesh points is increased.

Simulation times are reported in Table 1. A decrease of performance is present in parallel runs due to the communication overhead, increasing with the number of layers used for different schemes (2E : one layer; 4E : two layers; 4C and 6C : four layers).

REFERENCES

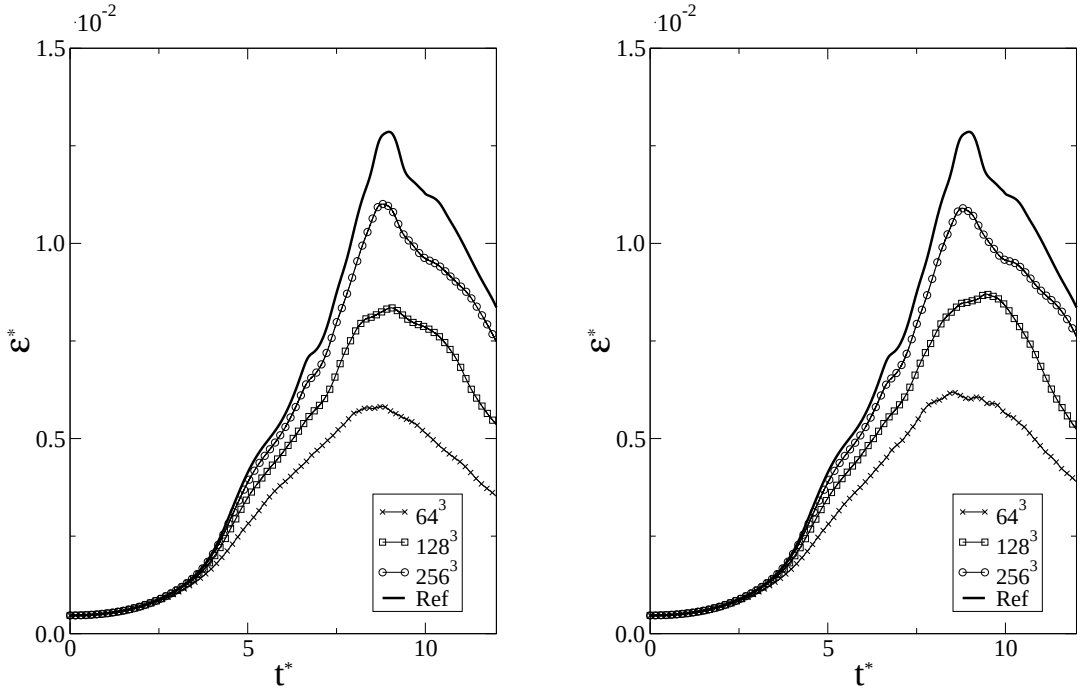
[1] Borrelli, S., de Matteis, P., Schettino, A. and Ferrigno, F., "The HYPROB Program: mastering key technologies, design and testing capabilities for space transportation rocket propulsion evolution," *Proceedings of the 63th International Astronautical Congress*, Naples (2012), pp. 7189-7190.

[2] Fosso, P., A., Deniau, H., Sicot, F. and Sagaut, P., "Curvilinear finite-volume schemes using high-order compact interpolation," *J. Comput. Phys.* (2010), **229**:5090–5122.

[3] Capuano, F., and Mastellone, A., "Parallelization of compact finite-volume schemes for turbulent compressible flow", *Proceedings of 27th International Conference on Parallel Computational Fluid Dynamics (PARCFD 2015)*, Montreal (2015), in press.

[4] DeBonis, J. R., "Solutions of the Taylor-Green Vortex Problem Using High-Resolution Explicit Finite Difference Methods", NASA/TM2013-217850.

Figure 4: Enstrophy-based dissipation rate as a function of the non-dimensional time t^* for fourth-order (left) and sixth-order (right) compact schemes.



[5] van Rees, W.M., Leonard, A., Pullin, D.I. and Koumoutsakos, P., "A comparison of vortex and pseudo-spectral methods for the simulation of periodic vortical flows at high Reynolds numbers," *J. Comput. Phys.* (2011), **230**:2794–2805.

Table 1: TauBench-normalized costs to integrate until $t^* = 10$. MPI and serial refer to runs over 64 and single node. Serial runs have been verified to show about same execution time for all the schemes.

N	Scheme	Total Work Units	
		MPI	Serial
64	2E	1180	942
64	4E	1420	
64	4C	1626	
64	6C	1662	
128	2E	20945	15125
128	4E	23571	
128	4C	25633	
128	6C	25978	
256	2E	311457	243567
256	4E	352837	
256	4C	380245	
256	6C	382699	

Figure 5: Contour of dimensionless vorticity norm $\frac{L}{U_0}\|\omega\| = 1, 5, 10, 20, 30$ in a subset of the periodic face $\frac{x}{L} = -\pi$ at time $t^* = 8$ on the three grids. Comparison between reference results in [5] (red) and SPARK-LES results (black) for second-order (left) and fourth-order (right) explicit schemes.

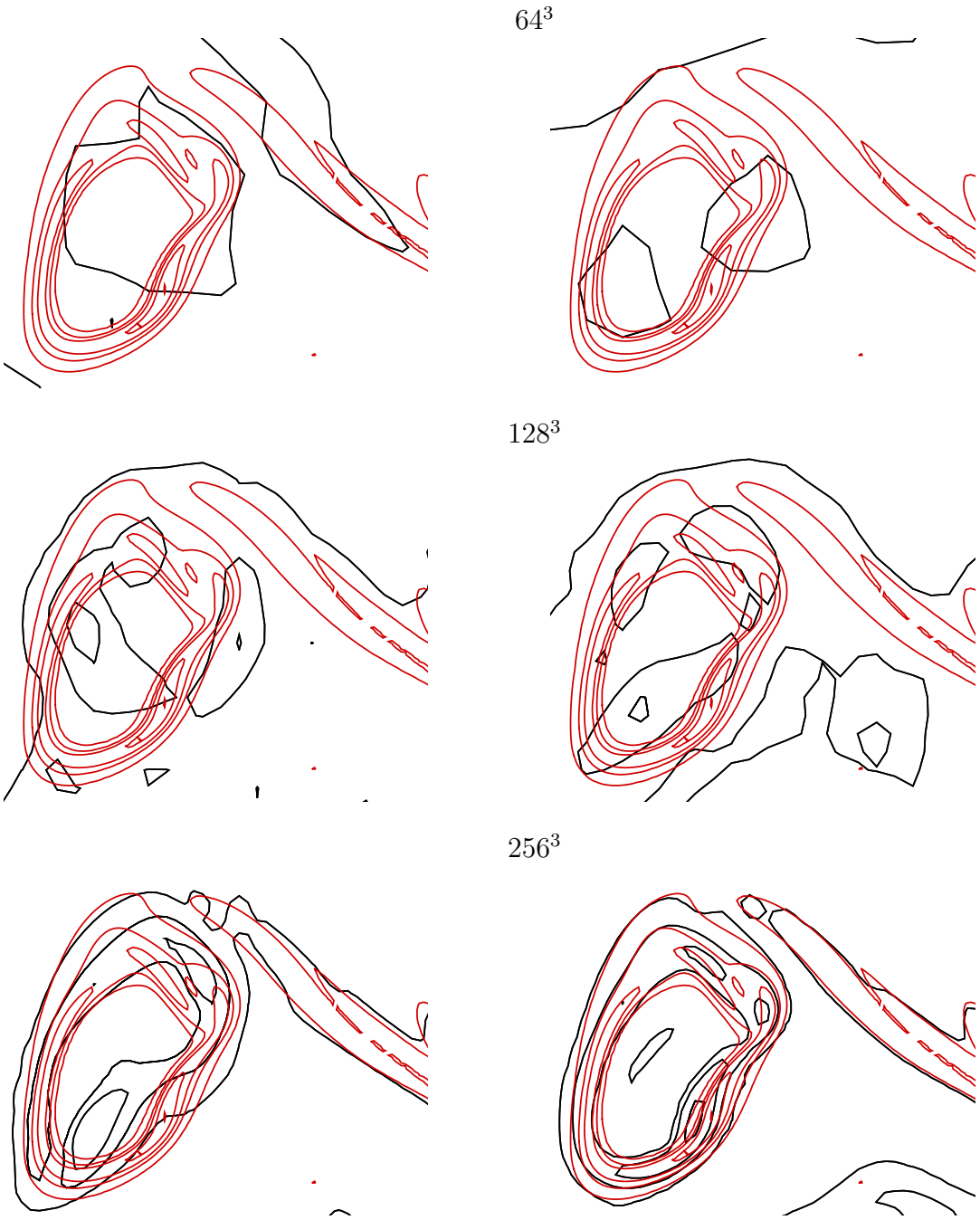


Figure 6: Contour of dimensionless vorticity norm $\frac{L}{U_0}\|\omega\| = 1, 5, 10, 20, 30$ in a subset of the periodic face $\frac{x}{L} = -\pi$ at time $t^* = 8$ on the three grids. Comparison between reference results in [5] (red) and SPARK-LES results (black) for fourth-order (left) and sixth-order (right) compact schemes.

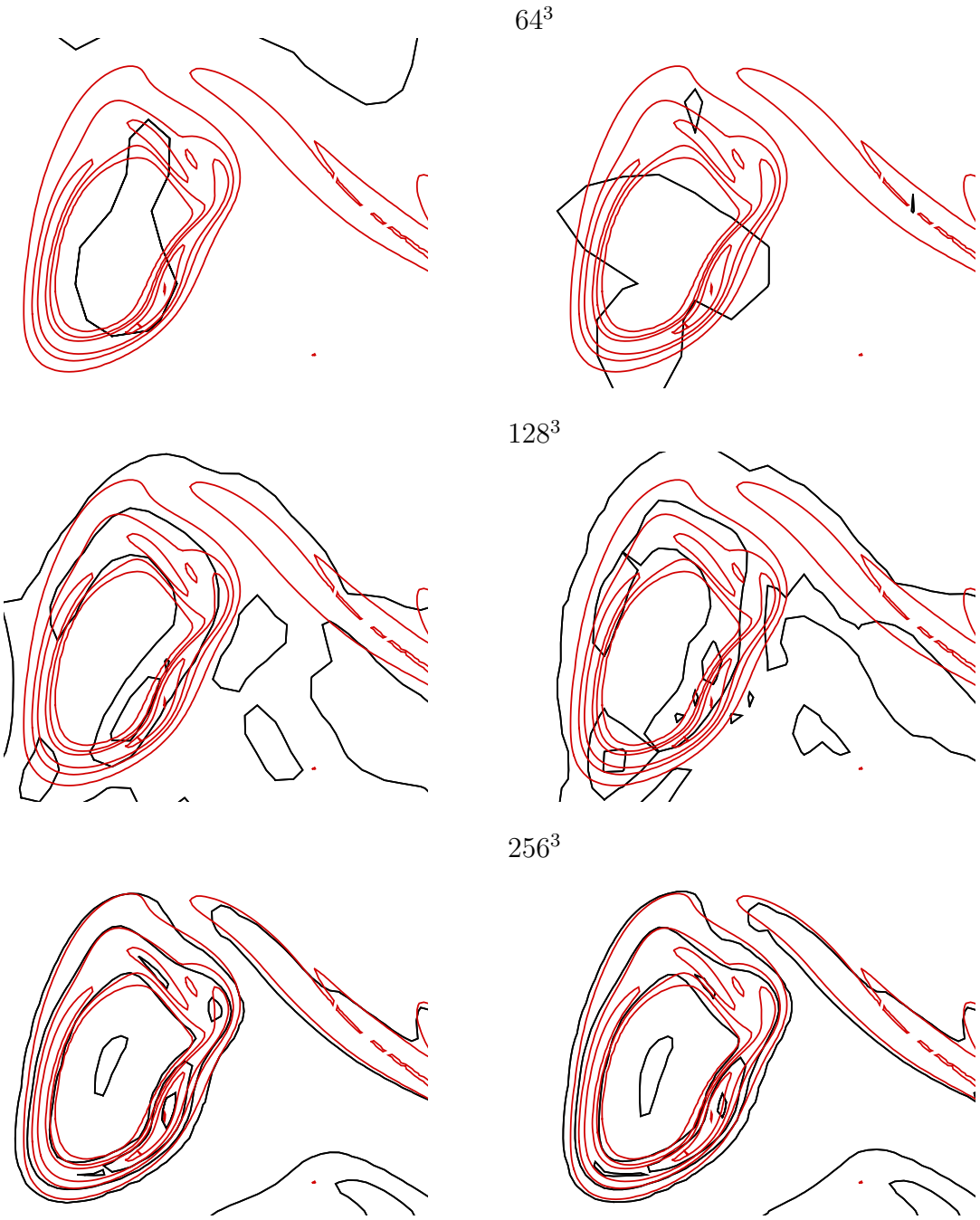


Figure 7: Power spectrum $E^*(k)$ for second-order (left) and fourth-order (right) explicit schemes.

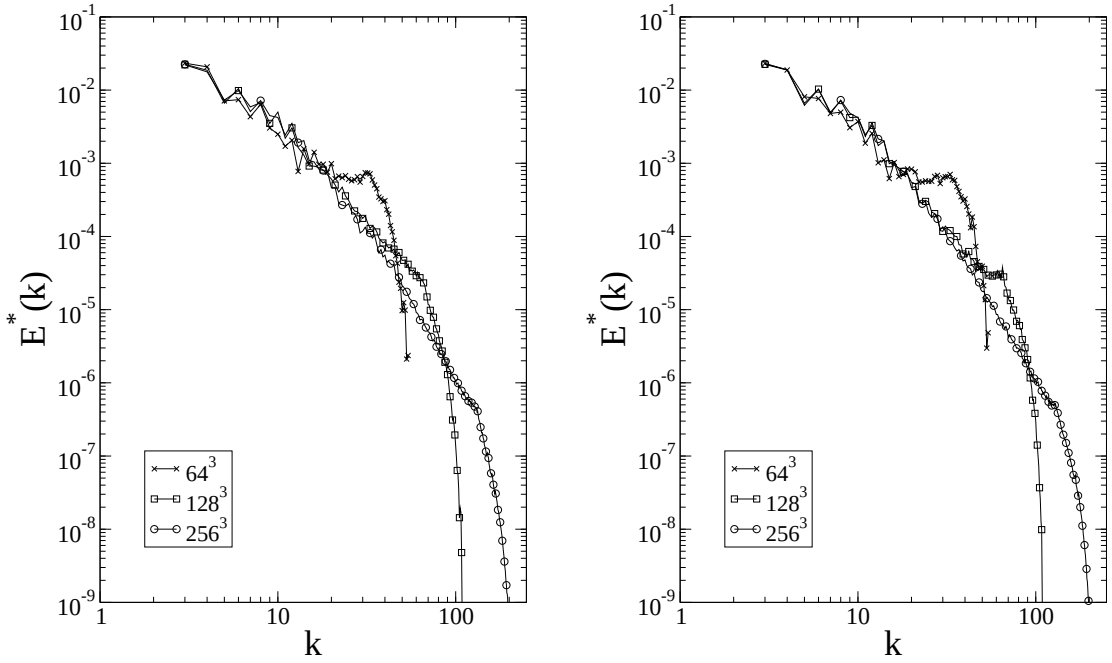
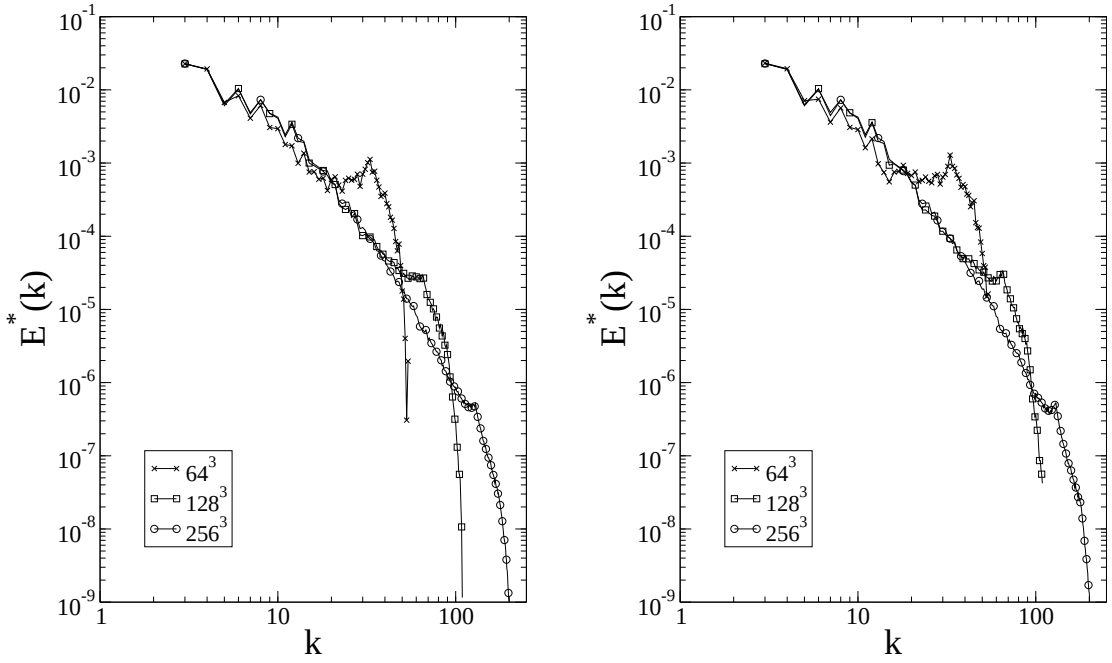


Figure 8: Power spectrum $E^*(k)$ for fourth-order (left) and sixth-order (right) compact schemes.



ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THE STRUCTURAL SYMMETRY WITHIN THE CONTEXT OF NONLOCAL ELASTICITY

A.A. Pisano[1], **P. Fuschi**[1]

[1] Dept. PAU, University Mediterranea of Reggio Calabria,
via Melissari, I-89124 Reggio Calabria, Italy
e-mail: aurora.pisano@unirc.it; paolo.fuschi@unirc.it

Keywords: Structural Symmetry, Nonlocal Elasticity of Integral Type, Nonlocal Finite Element Method.

Abstract. *The paper deals with symmetric structures made of nonlocal elastic materials. In particular, it refers to a nonlocal elastic strain integral constitutive model. In this context, in contrast to what happens in local elasticity, the solution obtained by analysing a standard symmetric portion of the structure leads to incorrect results. This drawback is due to the loss of nonlocal effects induced on the selected symmetric portion by the removed portion. To recover the mechanical equivalence with the original (entire) structure an enlarged symmetric model has to be considered together with the application of appropriate boundary conditions, hereafter discussed with the aid of a numerical example.*

1 INTRODUCTION

Many structural/mechanical problems are characterized by being symmetric in terms of geometry, material properties and boundary conditions. As well known, the presence of symmetries can drastically simplify the solution of such problems, halving or more the computational effort related to their analysis.

In the context of simple (local) materials the symmetric portions of a structure to be analyzed and the appropriate boundary conditions that have to be applied along the lines of symmetry are clearly identified. The solution computed on a reduced symmetric portion, once mirrored, gives the correct solution for the entire structure. The same cannot be asserted for symmetric structures made of nonlocal materials for which the constitutive law at a given point involves weighted averages of a state variable over a certain neighbor of the point [1]. The wideness of this zone is related to an internal length material scale parameter. For such structures the removal of one or more symmetric portions implies the loss of the nonlocal effects exerted by the removed portions on the one selected for the analysis.

The present paper suggests two possible remedies in order to obtain the expected (correct) solution of a nonlocal symmetric problem by recovering the missing nonlocal effects; precisely:

i) an enlarged symmetric model, whose dimension is related to the internal length material scale, has to be considered;

ii) appropriate (smeared) boundary conditions have to be applied to the above mentioned enlarged model.

Although the arguments developed seem to be of general validity in the context of nonlocal elasticity, in the present paper they are referred to a nonlocal approach of integral type known as strain-difference-based model [2] and the presented numerical example is solved by means of a nonlocal finite element method promoted by the authors [3].

2 THE THEORETICAL FRAMEWORK

2.1 The constitutive model

This paper considers the nonlocal integral elasticity model theorized by Polizzotto et *al.* in [2]. The quoted model, known as strain-difference-based model, proposes the following stress-strain constitutive relation:

$$\boldsymbol{\sigma}(\boldsymbol{x}) = \boldsymbol{D}(\boldsymbol{x}) : \boldsymbol{\varepsilon}(\boldsymbol{x}) - \alpha \int_V \boldsymbol{\mathcal{J}}(\boldsymbol{x}, \boldsymbol{x}') : [\boldsymbol{\varepsilon}(\boldsymbol{x}') - \boldsymbol{\varepsilon}(\boldsymbol{x})] \, \mathrm{d}\, \boldsymbol{x}' \quad \forall\, (\boldsymbol{x}, \boldsymbol{x}') \in V, \tag{1}$$

in which the stress is conceived as the sum of two contributions, the first is the one of the classical elasticity, while the second (integral) contribution is of nonlocal nature involving the nonlocal tensor $\boldsymbol{\mathcal{J}}(\boldsymbol{x}, \boldsymbol{x}')$ and the strain difference field $\boldsymbol{\varepsilon}(\boldsymbol{x}') - \boldsymbol{\varepsilon}(\boldsymbol{x})$. The nonlocal tensor $\boldsymbol{\mathcal{J}}(\boldsymbol{x}, \boldsymbol{x}')$ is defined as:

$$\boldsymbol{\mathcal{J}}(\boldsymbol{x}, \boldsymbol{x}') := [\gamma(\boldsymbol{x})\boldsymbol{D}(\boldsymbol{x}) + \gamma(\boldsymbol{x}')\boldsymbol{D}(\boldsymbol{x}')] \, g(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{q}(\boldsymbol{x}, \boldsymbol{x}') \quad \forall\, (\boldsymbol{x}, \boldsymbol{x}') \in V, \tag{2}$$

with:

$$\gamma(\boldsymbol{x}) := \int_V g(\boldsymbol{x}, \boldsymbol{x}') \, \mathrm{d}\, V'; \tag{3}$$

$$\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{x}') := \int_V g(\boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{x}', \boldsymbol{z}) \boldsymbol{D}(\boldsymbol{z}) \, \mathrm{d}\, V^z. \tag{4}$$

In the above equations $\boldsymbol{D}(\boldsymbol{x})$ is the symmetric and positive definite elastic moduli tensor, while $g(\boldsymbol{x},\boldsymbol{x}')$ is a positive scalar attenuation function depending on the internal length material scale, say ℓ, as well as on the Euclidean distance between points $\boldsymbol{x}$ and $\boldsymbol{x}'$ in V. The attenuation function assigns a *weight* to the nonlocal effects induced at the field point $\boldsymbol{x}$ by a phenomena acting at the source point $\boldsymbol{x}'$. This function has a peak at $\boldsymbol{x}\equiv\boldsymbol{x}'$ and rapidly decreases with increasing distance, becoming practically null beyond the so-called *influence distance*, say L_R which is a multiple of ℓ. It is worth noting that the model of Eq.(1) possesses, beside ℓ, a second material model parameter, α, which controls the proportion of the nonlocal contribution. Both material parameters ℓ and α should be experimentally determined.

2.2 The NL-FEM formulation

Let us consider a nonlocal elastic body occupying a volume V whose boundary surface is S. The body is subjected to body forces $\boldsymbol{b}(\boldsymbol{x})$ in V and surface tractions $\boldsymbol{t}(\boldsymbol{x})$ on S_t. Kinematic boundary conditions $\boldsymbol{u}(\boldsymbol{x})=\bar{\boldsymbol{u}}(\boldsymbol{x})$ are also specified on $S_u=S-S_t$. Moreover, the governing constitutive relation is the one given by Eq.(1). The pertinent boundary value problem is governed by the standard equilibrium and compatibility equations together with the stress strain relation (1). Following Polizzotto et *al.* [2] and Fuschi et *al.* [3], it can be shown that the nonlocal total potential energy functional associated to such nonlocal boundary value problem can be written as:

$$\begin{aligned}\Pi\left[\boldsymbol{u}(\boldsymbol{x})\right] := &\frac{1}{2}\int_V \nabla\boldsymbol{u}(\boldsymbol{x}):\boldsymbol{D}(\boldsymbol{x}):\nabla\boldsymbol{u}(\boldsymbol{x})\,\mathrm{d}\,V+\\ &+\frac{\alpha}{2}\int_V \nabla\boldsymbol{u}(\boldsymbol{x}):\gamma^2(\boldsymbol{x})\boldsymbol{D}(\boldsymbol{x}):\nabla\boldsymbol{u}(\boldsymbol{x})\,\mathrm{d}\,V+\\ &-\frac{\alpha}{2}\int_V\int_V \nabla\boldsymbol{u}(\boldsymbol{x}):\boldsymbol{\mathcal{J}}(\boldsymbol{x},\boldsymbol{x}'):\nabla\boldsymbol{u}(\boldsymbol{x}')\,\mathrm{d}\,V'\,\mathrm{d}\,V+\\ &-\int_V \boldsymbol{b}(\boldsymbol{x})\cdot\boldsymbol{u}(\boldsymbol{x})\,\mathrm{d}\,V-\int_{S_t}\boldsymbol{t}(\boldsymbol{x})\cdot\boldsymbol{u}(\boldsymbol{x})\,\mathrm{d}\,S.\end{aligned}\tag{5}$$

Grounding on Eq.(5) a nonlocal finite element formulation can be obtained from the following discretized form of functional Π:

$$\begin{aligned}\Pi\left[\boldsymbol{d_n}\right] = &\frac{1}{2}\sum_{n=1}^{N_e}\boldsymbol{d}_n^T\,\boldsymbol{k}_n^{loc}\,\boldsymbol{d}_n+\frac{\alpha}{2}\sum_{n=1}^{N_e}\boldsymbol{d}_n^T\,\boldsymbol{k}_n^{nonloc}\,\boldsymbol{d}_n+\\ &-\frac{\alpha}{2}\sum_{n=1}^{N_e}\sum_{m=1}^{N_e}\boldsymbol{d}_n^T\,\boldsymbol{k}_{nm}^{nonloc}\,\boldsymbol{d}_m-\sum_{n=1}^{N_e}\boldsymbol{d}_n^T\,\boldsymbol{f}_n,\end{aligned}\tag{6}$$

where $\boldsymbol{d}_n$ and $\boldsymbol{d}_m$ are the nodal displacements vectors of finite elements $\#n$ and $\#m$ respectively, $\boldsymbol{f}_n$ is the equivalent nodal forces vector, $\boldsymbol{k}_n^{loc}$ denotes the local element stiffness matrix, while $\boldsymbol{k}_n^{nonloc}$ and $\boldsymbol{k}_{nm}^{nonloc}$ are the element nonlocal stiffness matrices. Finally, N_e is the number of finite elements in which V has been subdivided.
With reference to Eq.(6) the following positions hold true:

$$\boldsymbol{k}_n^{loc} := \int_{V_n}\boldsymbol{B}_n^T(\boldsymbol{x})\,\boldsymbol{D}\,(\boldsymbol{x})\,\boldsymbol{B}_n(\boldsymbol{x})\,\mathrm{d}\,V_n;\tag{7}$$

$$\boldsymbol{k}_n^{nonloc} := \int_{V_n}\boldsymbol{B}_n^T(\boldsymbol{x})\,\gamma^2(\boldsymbol{x})\,\boldsymbol{D}\,(\boldsymbol{x})\,\boldsymbol{B}_n(\boldsymbol{x})\,\mathrm{d}\,V_n;\tag{8}$$

$$\boldsymbol{k}_{nm}^{nonloc} := \int_{V_n}\int_{V_m} \boldsymbol{B}_n^T(\boldsymbol{x})\, \boldsymbol{\mathcal{J}}\,(\boldsymbol{x},\boldsymbol{x}')\, \boldsymbol{B}_m(\boldsymbol{x}')\, \mathrm{d}\, V_m\, \mathrm{d}\, V_n; \tag{9}$$

$$\boldsymbol{f}_n := \int_{V_n} \boldsymbol{N}_n^T(\boldsymbol{x})\, \boldsymbol{b}(\boldsymbol{x})\, \mathrm{d}\, V_n + \int_{S_{t(n)}} \boldsymbol{N}_n^T(\boldsymbol{x})\, \boldsymbol{t}(\boldsymbol{x})\, \mathrm{d}\, S_n. \tag{10}$$

In the above expressions $\boldsymbol{N}_n(\boldsymbol{x})$ and $\boldsymbol{B}_n(\boldsymbol{x})$ denote the matrices of the *n-th* element shape functions and their Cartesian derivatives, respectively.
It is worth noting that matrix $\boldsymbol{k}_n^{nonloc}$ accounts for the influence exerted on the *n-th* element by the nonlocal diffusive processes over the whole domain and this by the presence of $\gamma^2(\boldsymbol{x})$; matrix $\boldsymbol{k}_{nm}^{nonloc}$ accounts explicitly for the nonlocal effects exerted by the *m-th* element on the *n-th* one and this by the presence of $\boldsymbol{B}_n(\boldsymbol{x})$ and $\boldsymbol{B}_m(\boldsymbol{x}')$ related to the elements $\#n$ and $\#m$, respectively. $\boldsymbol{k}_{nm}^{nonloc}$ is a *set of nonlocal matrices* pertaining to element $\#n$, precisely: a self-stiffness matrix, obtained for $m = n$, plus all the cross-stiffness matrices given by $m = 1, 2, ..., N_e$ with $m \neq n$.
Following the standard rationale of the FEM, Eq.(6) can be rewritten in terms of global DOFs and by minimization would furnish the solving global linear equation system, (see [3] for more details).

3 A SYMMETRIC NONLOCAL SQUARE PLATE

As already asserted in the introductory section, the analysis of a symmetric structure made of a nonlocal elastic material does not lead to a correct solution when a standard symmetric portion of the structure is considered. This assertion appears now more evident if the implications given by the constitutive assumption of Eq.(1) are considered. The boundary zone adjacent to the symmetry line and belonging to the removed symmetric portion affects the mechanical behavior of its mirror reflection on the portion to be analyzed. The dimension of the quoted boundary zone is related to the influence distance L_R within which all the nonlocal effects are active. With the aid of the following numerical example the effects on the solution due to nonlocality will be highlighted, together with the remedies proposed to recover the correct solution.
Let's consider the nonhomogeneous square plate shown in Fig.1a having side length equal to 5 cm and thickness t equal to 0.5 cm. The plate is made up of a nonlocal elastic material for which the parameter ℓ is set equal to 0.1 cm, while the parameter α is set equal to 50. A central square part of the plate, of sides $a = 1$ cm, has Young's modulus $E_1 = 84$ GPa while the remaining part has Young's modulus $E_2 = 260$ GPa; a Poisson's coefficient $\nu = 0.2$ is assumed for the whole structure. The plate is fixed at the left edge (i.e. at $x = 0$ cm) and it is subjected to a uniform prescribed displacement $\bar{u}_x = 0.001$ cm at the opposite edge ($x = 5$ cm). In the performed analysis a normalized bi-exponential attenuation function is considered in the form:

$$g(\boldsymbol{x},\boldsymbol{x}') := \frac{1}{2\pi\ell^2 t} \exp\left(-|\boldsymbol{x}-\boldsymbol{x}'|/\ell\right). \tag{11}$$

The structure of Fig.1a is symmetric with respect to geometry, materials and boundary conditions and Fig.1c reports the classical symmetric scheme to be utilized in the analysis of the plate. However, if the model of Fig.1c is assumed when the considered structure is made of a nonlocal elastic material, the results obtained from models a) and c) turn out to be more different as much more are the nonlocal effects.
This is shown by the results given in terms of strain component profiles ε_x, ε_y and ε_{xy} plotted at the mid-plate horizontal section, $\bar{y} \backsimeq 2.5$ cm drawn in Figs.2a-c for models b), complete,

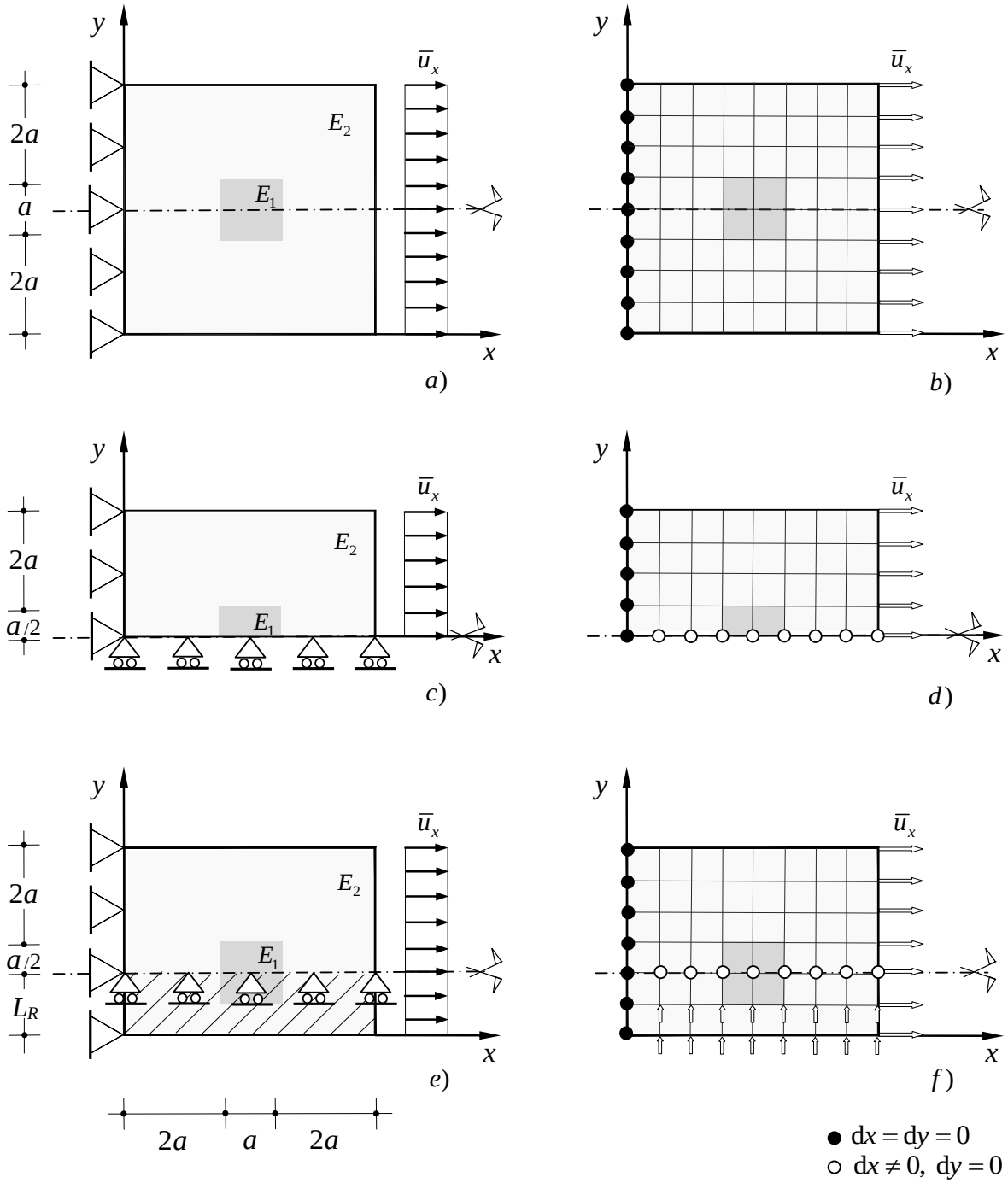


Figure 1: Nonlocal elastic symmetric square plate under tension with piecewise constant Young modulus. Mechanical model of: *a*) whole structure; *c*) symmetric half structure; *e*) enlarged symmetric half structure. *b*), *d*) and *f*) FE models of structures a), c) and e), respectively.

Figure 1: Nonlocal elastic symmetric square plate under tension with piecewise constant Young modulus. Mechanical model of: *a*) whole structure; *c*) symmetric half structure; *e*) enlarged symmetric half structure. *b*), *d*) and *f*) FE models of structures a), c) and e), respectively.

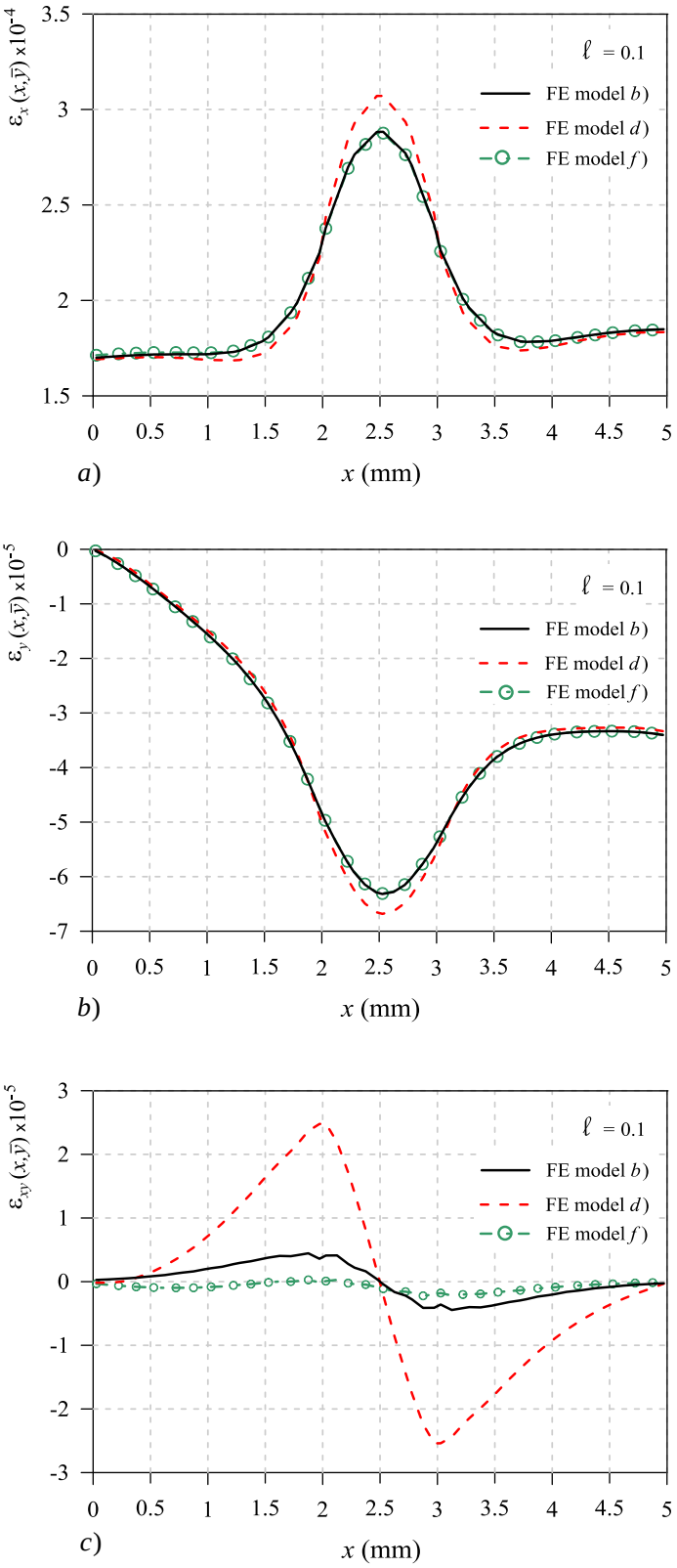


Figure 2: Nonlocal elastic symmetric square plate. Strain components profiles ε_x, ε_y, and ε_{xy} versus x at $y \backsimeq 2.5$. Solution referred to the entire plate (solid lines), solution referred to the half symmetric portion (dashed lines), solution referred to the enlarged half symmetric portion (dashed lines with circles).

and d), reduced. In particular the structure of Fig.1a has been discretized into 400 FEs of equal dimensions and consequently the model of the half classical symmetric structure is made of 200 FEs.

By inspection of Figure 2 appears evident how the half symmetric model furnishes a solution that deviates from the expected one. In order to recover the correct nonlocal solution the symmetric scheme of Fig.1c has to be enlarged, with respect to the symmetry axis, by adding a *symmetrical boundary zone* whose wideness is equal to the influence distance L_R as sketched in Fig.1e by the filled area below the symmetry axis. The related FE model of Fig.1f utilizes 320 FEs. To the enlarged model have also to be applied appropriate kinematic boundary conditions able to mimic the missing nonlocal effects. However, such kinematic conditions are unknown and the question of their specification is a crucial point. The need of them, to attain the exact solution, is here proved by making use of the solution in terms of displacements of the original (entire) structure solved on the scheme of Fig.1a. As in fact, if as nonlocal boundary conditions are assumed the vertical displacements coming from the solution (in the corresponding zone) of the original (entire) structure, the enlarged symmetric model gives exactly the expected solution for the strain profiles ε_x and ε_y and a very good approximation of the strain profile ε_{xy}, as shown in Figures 2.

4 CONCLUSIONS

The paper has pointed out some incoherences arising in the study of a symmetric structure made up of a nonlocal elastic material. In particular, it is highlighted that the physical reason for such incoherences is due to the fact that the removal of a symmetric portion produces the loss of the nonlocal effects exerted by that portion. It is shown that, to re-introduce in the solution the missed nonlocal effects, an enlarged symmetric model, with respect to the standard symmetric one, has to be employed. Appropriate nonlocal boundary conditions have also to be considered.

REFERENCES

[1] Z.P. Bažant, M.Jirásek. Nonlocal integral formulations of plasticity and damage: survey of progress. *Journal of Engineering Mechanics* **11**, 1119–1149, 2002.

[2] Polizzotto C., Fuschi P. and Pisano A.A. "A nonhomogeneous nonlocal elasticity model," *European Journal of Mechanics A/Solids* **25**, 308–333, 2006.

[3] Fuschi P., Pisano A.A. and De Domenico D. "Plane stress problems in nonlocal elasticity: finite element solutions with a strain-difference-based formulation," *Journal of Mathematical Analysis and Applications* **431**, 714–736, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ISOGEOMETRIC GALERKIN METHODS FOR GRADIENT-ELASTIC BARS, BEAMS, MEMBRANES AND PLATES

Jarkko Niiranen[1], **Sergei Khakalo**[1], **Viacheslav Balobanov**[1], **Josef Kiendl**[2], **Antti H. Niemi**[1], **Bahram Hosseini**[1], **Alessandro Reali**[3]

[1]Department of Civil Engineering, Aalto University
P.O. Box 12100, 00076 AALTO, Finland
e-mail: jarkko.niiranen@aalto.fi

[2]Institute of Applied Mechanics, Technische Universität Braunschweig
Bienroder Weg 87, 38106 Braunschweig, Germany

[3]Department of Civil Engineering and Architecture, University of Pavia
via Ferrata 3, 27100 Pavia (PV), Italy

Keywords: Isogeometric Analysis, Gradient Elasticity, Bars, Beams, Membranes, Plates.

Abstract. *Isogeometric Galerkin methods are used to analyse plate and beam bending problems as well as membrane and bar models based on Mindlin's strain gradient elasticity theory for generalized continua. The current strain gradient models include higher-order displacement gradients combined with length scale parameters enriching the strain and kinetic energies of the classical elasticity and hence resulting in higher-order partial differential equations with corresponding non-standard boundary conditions. The problems are first formulated within appropriate higher-order Sobolev space settings and then discretized by utilizing Galerkin methods with isogeometric NURBS basis functions providing appropriate higher-order continuity properties. Example benchmark problems illustrate the convergence properties of the methods.*

1 INTRODUCTION

Isogeometric Analysis (IGA) was introduced by Hughes et al. [9] roughly ten years ago having its primary focus on the integration of industrial design-analysis processes by performing finite element analysis by B-splines and NURBS utilized in computer aided design. Compared to the classical finite element analysis based on polynomial basis functions, isogeometric analysis provides some significant benefits originating from higher-order continuities of basis functions provided by basic isogeometric methods in a natural and straightforward manner, as C^{p-1} continuity for standard NURBS patches of order p. As a related particular implication, isogeometric Galerkin methods have turned out to be applicable for solving problems governed by higher-order partial differential equations which require higher-order requlatities for function spaces of Galerkin methods. In the present work, this capability is utilized for a group of problems following a strain gradient elasticity theory.

Typically, the aim of the generalized continuum theories as gradient elasticity is to take into account the effect of the microstructure of the material on its mechanical behaviour. The current strain gradient models, in particular, include higher-order displacement gradients combined with one length scale parameter [1, 2] enriching the strain energy of the classical elasticity theory [3]. Accordingly, these models result in higher-order partial differential equations with related non-standard boundary conditions [5, 6, 7, 8].

The paper is organized as follows: In Section 2, we introduce our notation by recalling the variational energies for the three-dimensional theory of gradient elasticity. Section 3 is devoted to the governing equations and weak forms of the corresponding dimensionally reduced bar, beam, membrane and plate problems, whereas in Section 4 isogeometric Galerkin discretizations are briefly formulated. Finally, in Section 5, we give some examples of numerical benchmarks illustrating the convergence properties of the methods.

2 GRADIENT ELASTICITY THEORY

Let us denote the classical linear (second order) strain tensor by $\boldsymbol{\varepsilon}$, defined as the symmetric part of the gradient of the displacement vector $\boldsymbol{u}$ in the form

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\big(\boldsymbol{\nabla}\boldsymbol{u} + (\boldsymbol{\nabla}\boldsymbol{u})^T\big). \tag{1}$$

As in the classical elasticity theory, the strain tensor is assumed to be related to its work conjugate, the classical (second order) Cauchy stress tensor, through the generalized Hooke's law

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\varepsilon} + \lambda \mathrm{tr}\,\boldsymbol{\varepsilon}\boldsymbol{I}, \tag{2}$$

with the Lame material parameters μ and λ, and $\boldsymbol{I}$ denoting the identity tensor.

In Mindlin's gradient elasticity theory of Form II [3], the (third-order) micro-deformation gradient tensor is defined as $\boldsymbol{\gamma} = \boldsymbol{\nabla}\boldsymbol{\varepsilon}$. Its work conjugate, the (third-order) double stress tensor $\boldsymbol{\tau}$, is defined in the simplest one parameter variant [1, 2] of Mindlin's theory as $\boldsymbol{\tau} = g^2\boldsymbol{\nabla}\boldsymbol{\sigma}$, with g denoting the gradient-elastic modulus describing the length scale of the micro-structure of the material. The virtual internal work expression over a body $\mathcal{B} \subset \mathbb{R}^3$ then takes the form

$$\delta W^g_{\text{int}} = \int_{\mathcal{B}} \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\delta\boldsymbol{u})\,\mathrm{d}\mathcal{B} + \int_{\mathcal{B}} g^2\boldsymbol{\nabla}\boldsymbol{\sigma} \vdots \boldsymbol{\nabla}\boldsymbol{\varepsilon}(\delta\boldsymbol{u})\,\mathrm{d}\mathcal{B}. \tag{3}$$

Applying integration by parts for the second term above gives the so called total stress tensor [1] in the form $\boldsymbol{\sigma}^g = (1 - g^2\Delta)\boldsymbol{\sigma}$, where g is assumed to be constant.

For analysing vibrations or time-dependent problems within the current gradient elasticity theory, an additional gradient parameter introducing a micro-inertia term has been proposed [3] in order to achieve a physically satisfactory dispersion relation for a large range of wave numbers [4]. The variation of the kinetic energy is then written in the form

$$\delta \int_T W_{\text{kin}}^{\gamma} \,\mathrm{d}\tau = -\int_T \Big(\int_{\mathcal{B}} \rho \ddot{\boldsymbol{u}} \cdot \delta \boldsymbol{u} \,\mathrm{d}\mathcal{B} + \int_{\mathcal{B}} \gamma^2 \rho \boldsymbol{\nabla} \ddot{\boldsymbol{u}} : \boldsymbol{\nabla} \delta \boldsymbol{u} \,\mathrm{d}\mathcal{B} \Big) \,\mathrm{d}\tau, \tag{4}$$

with ρ and T denoting the mass density and a time interval of the time variable τ, respectively, and finally γ standing for the gradient parameter related to the micro-inertia.

3 GOVERNING EQUATIONS AND WEAK FORMS

In this section, we briefly recall the governing equations and variational formulations for the bar and beam problems and then for the membrane and plate problems. For simplicity, we focus on statics, whereas for extensions to dynamics we refer to [5, 6, 8].

3.1 Bars and beams

Let us consider a thin bar or beam structure which occupies a three-dimensional body $\mathcal{B} = \Omega \times A$, where $\Omega = (0, L) \subset \mathbb{R}$ denotes the central axis of the structure with L standing for the length of the structure, and $A \subset \mathbb{R}^2$, with $\mathrm{diam}(A) \ll L$, denoting a cross-section of the structure. For simplicity, the cross-section is assumed to be constant.

With a constant Young's modulus E, the governing equation of the bar problem reads as [5]

$$EAu''(x) - g^2 EAu''''(x) + Af(x) = 0 \quad \forall x \in \Omega, \tag{5}$$

with *doubly* clamped, *singly* clamped and free boundaries, respectively:

$$u(x) = \overline{u}, \quad u'(x) = \overline{w} \quad \forall x \in \Gamma_{\mathrm{C_d}}, \tag{6}$$

$$u(x) = \overline{u}, \quad g^2 EAu''(x) = \overline{G} \quad \forall x \in \Gamma_{\mathrm{C_s}}, \tag{7}$$

$$EAu'(x) - g^2 EAu'''(x) = \overline{P}, \quad g^2 EAu''(x) = \overline{G} \quad \forall x \in \Gamma_{\mathrm{F}}. \tag{8}$$

The weak formulation of the problem reads as follows [5]: For $f \in L^2(\Omega)$, find $u \in U \subset H^2(\Omega)$ such that

$$a(u, v) = l(v) \quad \forall v \in V \subset H^2(\Omega), \tag{9}$$

where the bilinear form $a : U \times V \to \mathbb{R}$, $a(u, v) = a^{\mathrm{c}}(u, v) + a^{\nabla}(u, v)$, and the load functional $l : V \to \mathbb{R}$, respectively, are defined by

$$a^{\mathrm{c}}(u, v) = \int_\Omega EAu'v' \,\mathrm{d}x, \quad a^{\nabla}(w, v) = \int_\Omega g^2 EAu''v'' \,\mathrm{d}x, \quad l(v) = \int_\Omega Af\, v \,\mathrm{d}x. \tag{10}$$

The trial and test function sets are denoted by U and V, respectively. The problem is continuous and coercive with respect to the H^2 norm [5].

With the moment of inertia I, the governing equation of the Euler–Bernoulli beam bending problem reads as [6]

$$EIw''''(x) - g^2 EIw''''''(x) = f(x) \quad \forall x \in \Omega, \tag{11}$$

while the boundary conditions can be found in [6]. The weak formulation of the problem reads as follows [6]: For $f \in L^2(\Omega)$, find $w \in W \subset H^3(\Omega)$ such that

$$a(w,v) = l(v) \quad \forall v \in V \subset H^3(\Omega), \tag{12}$$

where the bilinear form $a : W \times V \to \mathbb{R}$, $a(w,v) = a^{\mathrm{c}}(w,v) + a^{\nabla}(w,v)$, and the load functional $l : V \to \mathbb{R}$, respectively, are defined as

$$a^{\mathrm{c}}(w,v) = \int_\Omega EIw''v'' \,\mathrm{d}x, \quad a^{\nabla}(w,v) = \int_\Omega g^2 EIw'''v''' \,\mathrm{d}x, \quad l(v) = \int_\Omega f\, v \,\mathrm{d}x. \tag{13}$$

The trial and test function sets are denoted by W and V, respectively. The problem is continuous and coercive with respect to the H^3 norm [6].

3.2 Membranes and plates

Let us consider a thin planar membrane or plate structure which occupies a three-dimensional body $\mathcal{B} = \Omega \times (-t/2, t/2)$, where the domain $\Omega \subset \mathbb{R}^2$ denotes the midsurface of the structure and $t \ll \mathrm{diam}(\Omega)$ stands for the thickness of the structure. For simplicity, the thickness is assumed to be constant.

The governing equation of the (plane stress) membrane problem is written as [5]

$$(1 - g^2\Delta)\mathbf{div}\,\boldsymbol{\sigma} + \boldsymbol{f} = \boldsymbol{0} \quad \text{in}\,\Omega, \tag{14}$$

while the boundary conditions can be found in [5]. The weak formulation of the problem reads as follows [5]: For $\boldsymbol{f} \in [L^2(\Omega)]^2$, find $\boldsymbol{u} \in \boldsymbol{U} \subset [H^2(\Omega)]^2$ such that

$$a(\boldsymbol{u},\boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \boldsymbol{V} \subset [H^2(\Omega)]^2, \tag{15}$$

where the bilinear form $a : \boldsymbol{U} \times \boldsymbol{V} \to \mathbb{R}$, $a(\boldsymbol{u},\boldsymbol{v}) = a^{\mathrm{c}}(\boldsymbol{u},\boldsymbol{v}) + a^{\nabla}(\boldsymbol{u},\boldsymbol{v})$, and the load functional $l : \boldsymbol{V} \to \mathbb{R}$, respectively, are defined as

$$a^{\mathrm{c}}(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega (2\mu\,\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \mathrm{tr}\,\boldsymbol{\varepsilon}(\boldsymbol{u})\boldsymbol{I}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) \,\mathrm{d}\Omega, \tag{16}$$

$$a^{\nabla}(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega \boldsymbol{\nabla} g^2 (2\mu\,\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \mathrm{tr}\,\boldsymbol{\varepsilon}(\boldsymbol{u})\boldsymbol{I}) \vdots \boldsymbol{\nabla}\boldsymbol{\varepsilon}(\boldsymbol{v}) \,\mathrm{d}\Omega, \quad l(\boldsymbol{v}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \,\mathrm{d}\Omega. \tag{17}$$

The trial and test function sets are denoted by $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. The problem is continuous and coercive with respect to the H^2 norm [5].

The governing equation of the Kirchhoff plate bending problem is written as

$$D\Delta^2 w - g^2 D\Delta^3 w = f \quad \text{in}\,\Omega, \tag{18}$$

with the bending rigidity defined as

$$D = \frac{Et^3}{12(1-\nu^2)}. \tag{19}$$

The boundary conditions of the problem can be found in [7].

The weak formulation of the problem reads as follows [7]: For $f \in L^2(\Omega)$, find $w \in W \subset H^3(\Omega)$ such that

$$a(w,v) = l(v) \quad \forall v \in V \subset H^3(\Omega), \tag{20}$$

where the bilinear form $a : W \times V \to \mathbb{R}$, $a(w,v) = a^{\mathrm{c}}(w,v) + a^{\nabla}(w,v)$, and the load functional $l : V \to \mathbb{R}$, respectively, are defined as

$$a^{\mathrm{c}}(w,v) = \int_\Omega \boldsymbol{E}\boldsymbol{\varepsilon}(\nabla w) : \boldsymbol{\varepsilon}(\nabla v)\,\mathrm{d}\Omega, \tag{21}$$

$$a^{\nabla}(w,v) = \int_\Omega g^2 \boldsymbol{\nabla}(\boldsymbol{E}\boldsymbol{\varepsilon}(\nabla w)) \vdots \boldsymbol{\nabla}\boldsymbol{\varepsilon}(\nabla v)\,\mathrm{d}\Omega, \quad l(v) = \int_\Omega f\, v\,\mathrm{d}\Omega. \tag{22}$$

The symmetric positive definite bending moduli tensor is defined, in the case of constant bending rigidity, by the relation

$$\boldsymbol{E}\boldsymbol{\varepsilon} = D\big((1-\nu)\boldsymbol{\varepsilon} + \nu\,\mathrm{tr}\,\boldsymbol{\varepsilon}\boldsymbol{I}\big). \tag{23}$$

The trial and test function sets are denoted by W and V, respectively. The problem is continuous and coercive with respect to the H^3 norm [7].

4 CONFORMING ISOGEOMETRIC GALERKIN APPROACH

Within the gradient elasticity theory, the bar and mebrane problems deal with H^2 spaces, whereas for the beam and plate problems H^3 is the appropriate Sobolev space. Therefore, C^1 and C^2 continuity, respectively, are required for the corresponding conforming Galerkin approximations. Within the classical elasticity theory, instead, these problems deal with less regular function spaces, H^1 and H^2, respectively, requiring C^0 and C^1 continuity, respectively, for conformity. Isogeometric methods provide higher-order regularities for the approximation spaces in a straightforward manner. Furthermore, with continuity and coercivity, conformity implies Cea's lemma type convergence results for the methods [5, 6, 8].

Let us next briefly recall the isogeometric tensor product discretizations [9] which can be applied for solving the problems formulated in the previous section. As an example, we construct a discrete space for the plane stress problem.

First, we introduce an isoparametric discrete space S_h for the approximation of the displacement field such that $\boldsymbol{u}_h \in [S_h]^2$ with

$$S_h = \{R^{p,q}_{i,j} \circ \boldsymbol{F}^{-1}\}. \tag{24}$$

The geometrical mapping between the two-dimensional parametric space $[0,1] \times [0,1]$ and the midsurface $\bar{\Omega}$ is defined by $\boldsymbol{F} : [0,1] \times [0,1] \to \bar{\Omega}$ as

$$\boldsymbol{F}(\xi,\eta) = \sum_{i=1}^{n}\sum_{j=1}^{m} R^{p,q}_{i,j}(\xi,\eta)\boldsymbol{B}_{i,j} \tag{25}$$

providing an isogeometric NURBS discretization. Above, $\boldsymbol{B}_{i,j}$, $i = 1 \ldots n$, $j = 1 \ldots m$, denote the control points, while the NURBS basis functions are defined as

$$R^{p,q}_{i,j}(\xi,\eta) = \frac{N_{i,p}(\xi)M_{j,q}(\eta)w_{i,j}}{\sum_{\hat{i}=1}^{n}\sum_{\hat{j}=1}^{m} N_{\hat{i},p}(\xi)M_{\hat{j},q}(\eta)w_{\hat{i},\hat{j}}}. \tag{26}$$

The B-spline basis functions $N_{i,p}$ and $M_{j,q}$ of order p and q, respectively, associated to the open knot vectors $\{0 = \xi_1, \ldots, \xi_{n+p+1} = 1\}$, $\{0 = \eta_1, \ldots, \eta_{m+q+1} = 1\}$, respectively, are defined as follows:

$$N_{i,0}(\xi) = \begin{cases} 1, & \xi_i \le \xi < \xi_{i+1}, \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

$$N_{i,p}(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(\xi). \tag{28}$$

The tensor product mesh of the midsurface, with the mesh size h, is denoted by

$$\mathcal{T}_h = \{\boldsymbol{F}([\xi_i, \xi_{i+1}] \times [\eta_j, \eta_{j+1}]) | i \in [n+p], j \in [m+q]\}. \tag{29}$$

By assuming $p = q$ and global regularity C^{p-1} over $\mathcal{T}_h$, with $p \geq 2$, it holds that $S_h \subset H^2(\Omega)$, which provides a conforming and consistent Galerkin method following the formulation (15) with $\boldsymbol{U}_h = [S_h]^2 \cap \boldsymbol{U}$, $\boldsymbol{V}_h = [S_h]^2 \cap \boldsymbol{V}$.

5 NUMERICAL RESULTS

As simple examples of convergence results for the proposed methods, we study the beam and plate bending problems with constant material values and smooth distributed loadings. For numerical results concerning free vibrations, we refer again to [5, 6, 8].

In Figures 1 and 2, for the beam and plate problems, respectively, the convergence of the relative error in the H^3 norm is plotted against the number of degrees of freedom (in logarithmic scales). It can be seen that for the NURBS order $p = 3, 4, 5$ the convergence rates follow the theoretical rates [6, 8] depicted by the dashed slopes.

REFERENCES

[1] E. C. Aifantis. On the role of gradients in the localization of deformation and fracture. *International Journal of Engineering Science*, **30**, 1279–1299, (1992).

[2] C. Q. Ru and E. C. Aifantis. A simple approach to solve boundary-value problems in gradient elasticity. *Acta Mechanica*, **101**, 59–68, (1993).

[3] R. D. Mindlin. Micro-structure in linear elasticity. *Archive for Rational Mechanics and Analysis*, **16**, 51–78, (1964).

[4] S. Papargyri-Beskou, D. Polyzos, D. E. Beskos. Wave dispersion in gradient elastic solids and structures: A unified treatment. *International Journal of Solids and Structures*, **40**, 385–400, (2009).

[5] J. Niiranen, S. Khakalo, V. Balobanov, A. H. Niemi. Variational formulation and isogeometric analysis for fourth-order boundary value problems of gradient-elastic bar and plane strain/stress problems, *in review*.

[6] J. Niiranen, V. Balobanov, J. Kiendl, B. Hosseini. Isogeometric analysis of gradient-elastic Euler–Bernoulli beams, *submitted*.

[7] J. Niiranen, A. Niemi. Variational formulation and general boundary conditions for sixth order boundary value problems of gradient-elastic Kirchhoff plates, *in review*.

[8] J. Niiranen, A. H. Niemi, J. Kiendl and A. Reali. Isogeometric analysis for sixth-order boundary value problems of gradient-elastic Kirchhoff plates, *submitted*.

[9] T.J.R. Hughes, J.A. Cottrell and Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, **194**, 4135–4195, (2005).

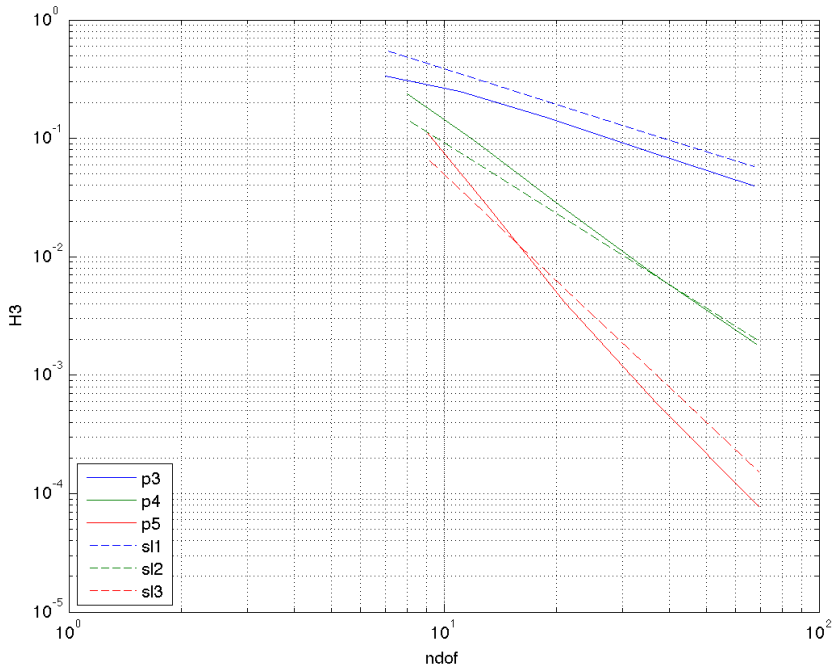


Figure 1: Doubly simply supported beam: Convergence of the relative error in the H^3 norm for $p = 3, 4, 5$ with $g = t = L/20$.

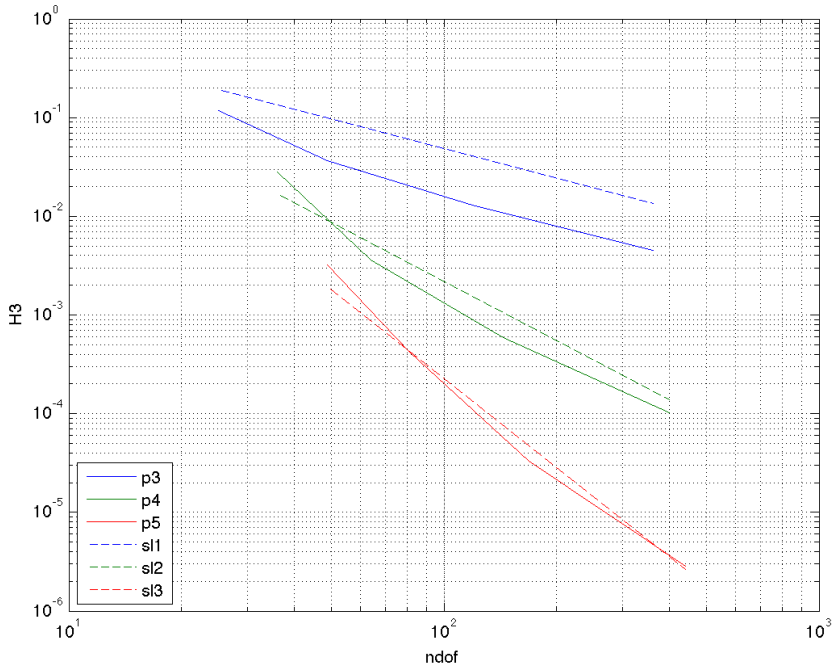


Figure 2: Doubly simply supported square plate: Convergence of the relative error in the H^3 norm for $p = 3, 4, 5$ with $g = t = 0.1$.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ADAPTIVE MESH REFINEMENT TECHNIQUES FOR HIGH-ORDER FINITE-VOLUME WENO SCHEMES

Harshavardhana Srinivasan[1], **Panagiotis Tsoutsanis**[1],

[1] Fluid Mechanics and Computational Sciences Centre, Cranfield University
College Road,Cranfield, Bedfordshire, MK 43 0AL
e-mail: h.srinivasan@cranfield.ac.uk, Panagiotis.tsoutsanis@cranfield.ac.uk

Keywords: Adaptive Mesh Refinement, Unstructured mesh, Finite-volume, Weno

Abstract. *This paper demonstrates the capabilities of Adaptive Mesh Refinement Techniques (AMR) on 2D hybrid unstructured meshes, for high order finite volume WENO methods. The AMR technique developed is a conformal adapting unstructured hybrid quadrilaterals and triangles (quads & tris) technique for resolving sharp flow features in accurate manner for steady-state and time dependent flow problems. In this method, the mesh can be refined or coarsened which depends on an error estimator, making decision at the parent level whilst maintaining a conformal mesh, the unstructured hybrid mesh refinement is done hierarchically. When a numerical method can work on a fixed conformal mesh this can be applied to do dynamic mesh adaptation. Two Refinement strategies have been devised both following a H-P refinement technique, which can be applied for providing better resolution to strong gradient dominated problems. The AMR algorithm has been tested on cylindrical explosion test and forward facing step problems.*

1 INTRODUCTION

Adaptive mesh refinement (AMR) techniques are well known and vastly used technique for the accurate capturing of the solution features in a steady or an unsteady simulation. The adaptive refinement enables to capture complex solution features by performing refinement in critical areas without having to refine the whole mesh. AMR has become a standard practice in triangular and tetrahedral meshes for various applications, the unique topological properties of these elements allow for local refinement and maintaining good element quality and retaining the conformity of the mesh [6]. For a quadrilateral mesh the general approach of refinement generates non-conformal elements. This non-conformity may allow local refinement but introduces hanging nodes, which requires special augmentation of the PDE solution to deal with these special nodes. Hanging nodes are generally dealt by constraining the solution at these nodes to be dependent on the solution at the nodes of the edge it lies on using constraint equations [4].

The adaptive procedures automatically try to refine, coarsen or relocate the mesh or tries to adjust the solution basis to achieve a specific accuracy in an optimal way. The computations generally begin with a trial solution generated on a coarse mesh which has a lower order basis where the error of this solution is assessed. If this fails to satisfy the required accuracy, adjustments are made to obtain the required solution with minimum effort, where we try to reduce the discretisation error to its required. Adaptive methods have been studied for nearly twenty years now and there are still only a few known optimal strategies and few of the common procedures studied till date include, the local refining or coarsening of a mesh (h-refinement), relocating or moving the nodes in a mesh (r-refinement) and locally varying the polynomial degree (p-refinement).

This paper demonstrates the capabilities of AMR on 2D hybrid unstructured meshes, for high order finite volume WENO methods. The AMR technique developed is a conformal adapting unstructured hybrid (quads & tris) technique for resolving sharp flow features in accurate manner for steady-state and time dependent flow problems. In this method, the mesh can be refined or coarsened which depends on an error estimator, making decision at the parent level whilst maintaining a conformal mesh, the unstructured hybrid mesh refinement is done hierarchically. When a numerical method can work on a fixed conformal mesh this can be applied to do dynamic mesh adaptation [4]. The adaptation strategy devised follows a H-P refinement technique, which can be applied for providing better resolution to strong gradient dominated problems.

2 NEED FOR REFINEMENT

Most of the physical problems that are considered for numerical simulation have features with multiple scales in both space and time, which has been a problem for numerical analysis in attempts to resolve more of the scales with evenly spaced grids which require more computational resources, storage and time to execute, because of this the meshes were pre adapted to known features in the solution. This method works well for steady solutions with well-known locations of the local features within the solution, but for solution features which are not well known, multiple solution or remeshing were needed to get the considerable resolution level which required a complex remeshing code. These problems become absurd when unsteady solutions are considered, a small movement in the solution feature may render extensive remeshing of no use. To overcome this, zonal refinement techniques have also been used which also fails when unrestrained feature movement exists. For the adaptive mesh techniques to be useful for unsteady or steady flows where there is no proper knowledge of the solution, it has to adapt dynamically and automatically as the solution evolves. Using a standard mesh generation code

in this situation will require stopping the solution periodically, remeshing, interpolating and then restarting, to resolve multiple scale unsteady problems a dynamic solution adaptive grid technique has to be developed.
With references to the study made by Joe.F.Thompson[11] on grid generation and adaptive techniques, an adaptive mesh refinement technique has to fulfil a few criteria which can be termed as the goals of adaptive mesh refinement. According to this study amr has to,

- Reduce spatial discretisation error.
- Remove grid dependency of the solution to maximum extent.
- Preserve mesh quality as far as possible.
- The results with adaptation have to be quantifiable.
- If solution is time dependent, the adaptation should be dynamic and has to preserve temporal accuracy.
- Once initial criteria is selected, the adaptive process should continue without any user intervention.
- The adaptive technique should be effective.
- There should be minimal error added to the solution.

3 PREVIOUS WORK

There has been extensive research on the conformal refinement of triangular meshes for adaptive simulations as treatment of these element types are easy [2]. On the contrary for quadrilateral meshes it is common to use non-conformal quad tree type refinement with special treatment for the hanging nodes [1]. There have been only a few researchers describing the coarsening and refinement of quadrilateral mesh and relatively fewer which can also handle hybrid mesh elements and also deal with the dynamic settings.

The best Paper on conformal quadrilateral refinement written by Schneider's [8] discusses the method of refinement based on bisection and trisection of the edges. He implies that the trisection of edge strategy simplifies the algorithm, where the information of refinement is communicated from elements to nodes and the templates for refinement are defined based on the number of marked nodes. The refinement templates are selected to keep the scheme stable, where the quality of the elements do not depreciate with increasing refinement levels. In his paper, uniformly refined quadrilaterals which are trisected have 9 child cells and the templates for the adjacent cells which terminate the refinement have bisected edge. Schneider's schemes are more complicated to implement than the scheme to be presented here, but still is a valid and tested scheme for conformal quadrilateral refinement and has also been used by others like Zhang and Bajaj [12]. Schneider's extended his work to hexahedral but says that certain refinement patterns for the faces of hexahedra may not admit a valid decomposition of the parent hexahedron. Ito et al. [5] has also used the Schneider's approach for octree hexahedral refinement templates.

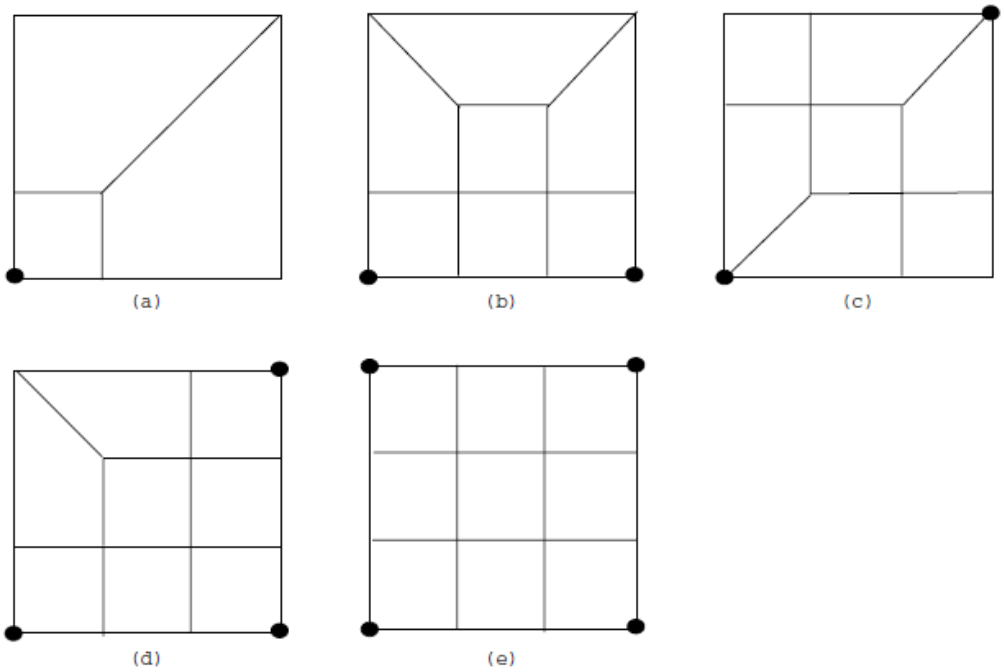


Figure 1. Schniders' subdivision templates showing the trisection and bisection of edges. [8]

Tchon et al. [10 (Tchon, 2004)] has also worked on quadrilateral refinement strategy where they find layers of elements, shrink the layers of elements and reconnect the shrunk layer with the surrounding mesh. This strategy assumes a particular structure to the mesh and a few specific patterns may ignore the issues of multiple levels of refinement, mesh quality and dynamic adaptation which makes this method limited in utility.

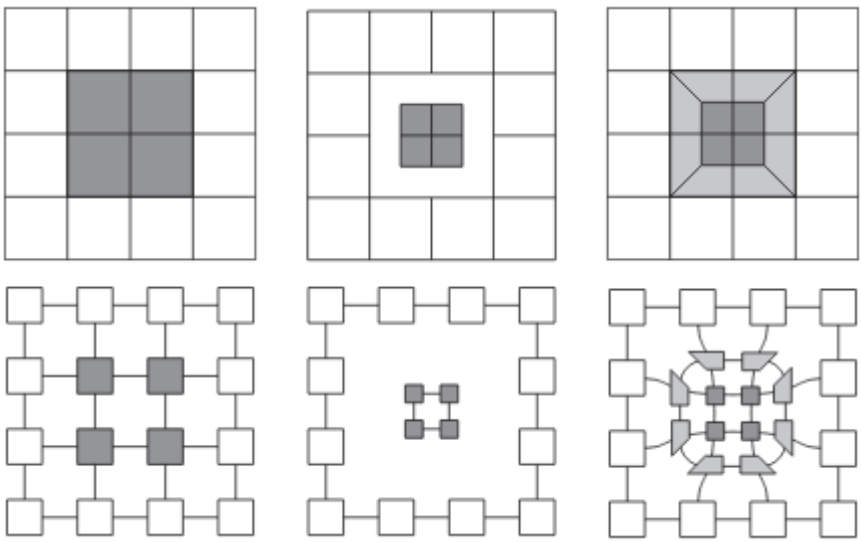


Figure 2. Tchon et al.'s conformal shrink and reconnect method [10]

Benzley et al. [9] proposed quadrilateral mesh coarsening strategies which are general and have an advantage over nested refinement strategies where they can be coarsened beyond the original resolution of the mesh.

In the work presented by Sandhu et al. [7] they use node marking and trisection of edges to define templates for refinement, which is quite similar to the work done by Garimella Rao [4] where he proposes a technique for multilevel adaptive refinement of quadrilateral meshes where the elements are kept conformal all the time. Garimella uses one less defined template than Sandhu et al.. With all these methods into consideration, they show only static refinement and aspects of dynamic adaptation and solution mapping have not been explored.
The research that is closest to the presented work is the paper by Michael Dumbser et al. [3] which talks about high order ader-weno finite volume scheme with AMR. The higher order spatial accuracy is got by using a weno reconstruction and the high order one step time discretisation is obtained through discontinuous galerkin predictor method. The AMR strategy is implemented cell-by-cell with standard tree type algorithm and has also been parallelised. The strategies have been tested on nonlinear systems including the Euler equations and with varying orders of accuracy to show the results of using AMR.

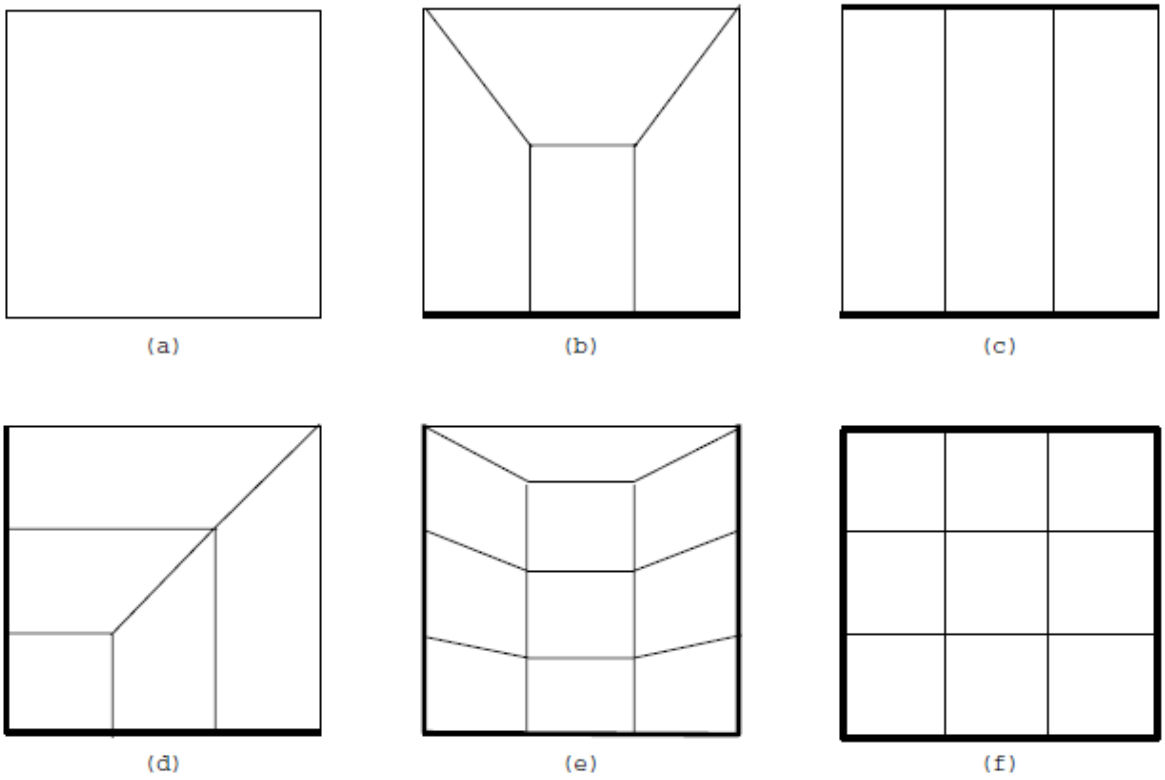


Figure 3. Graimella Rao's subdivision templates demonstrating edge based refinement [4]

4 AMR ALGORITHM

The mesh adaptation technique that has been developed is used in conjunction with the high order unstructured finite volume solver (ucns) which is capable of obtaining 7th order spatial accuracy with weno schemes. The AMR technique implemented is a fully conformal method that can work on hybrid meshes, which follows a hierarchical tree based data structure. The refinement or coarsening is node based and the decision is always taken at the parent level i.e. the data of the initial parent cell is preserved and is always carried over to the number of adaptations done during the process which limits the coarsening to be done only up to the initial mesh level. The main point of this technique is to preserve the order of accuracy with varying

gradients and also at points where the solution is smooth using solution transfer/remapping techniques. The other advantage of this method developed here is that the refinement or the coarsening process can be done as local as possible and also can be done per cell basis which gives the user more control over the process and is very essential and would aid in better parallelisation of the algorithm.

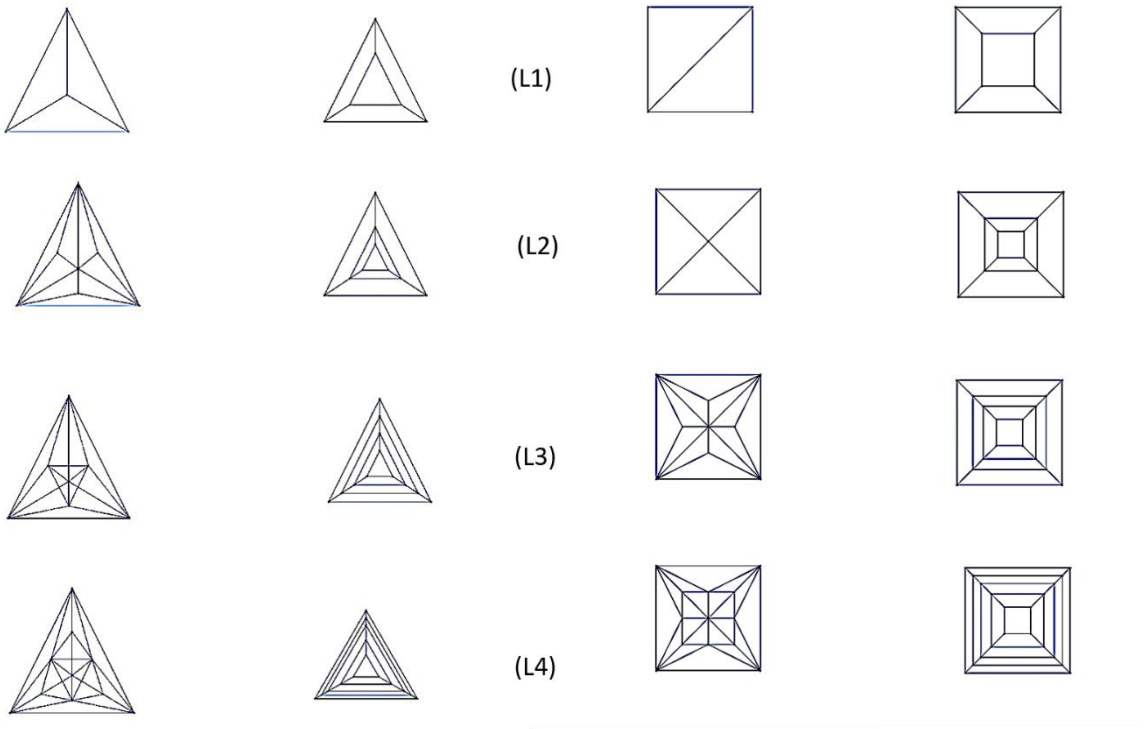


Figure 4. The two strategies of refinement on quadrilateral and triangular cells, showing different levels of refinement.

Two strategies of refinements have been developed for quadrilaterals and triangles which can refine the cells up to 4 levels where the maximum number of children per cell can go up to 21 cells. The above figure shows the subdivision of cells for both the refinement strategies. The subdivision is done based on the nodes and the centre of the cells, where for each level the midpoint is calculated on the diagonals between the end node and the cell centre. This is done repetitively until four levels for the first strategy. The second strategy of subdivision was developed for enabling better stencil marking for the weno schemes to communicate better with the neighbour cells. The subdivision of the cells were kept uniform by following a step by step subdivision where the new cells generated can take up an identical or an averaged area or volume for better solution remapping and transfer.

For the AMR to be activated it is necessary for the solver to start, to get the initial gradients and the solution. The refinement or coarsening is done based on the non dimensionalised gradient levels. With the initial mesh fed to the solver we get the gradients and the solution, the criteria and the levels of refinement are pre-defined in a parameter file based on which gradient to choose. The AMR can be activated in two ways, for every n iterations of the simulation or for every n-th time step of the simulation based on the type of problem being solved. From the

initial mesh with the parameters given for refinement the cells are marked, it has to be noted that there can be no coarsening happening at the initial level as the algorithm does not support coarsening of the mesh beyond the initial mesh. As soon as the cells are marked the neighbours of the marked cells as well as the neighbours of the neighbours are marked and only the exclusive neighbour list is populated and is stored for smooth transition of refinement levels which makes sure that the adjacent cell is always only one level different than the target cell. The non dimensionalised gradients are now used to specify the level of refinement to the marked target cells along with their neighbours for a smoother transition. The refinement is done for the marked cells and the data of the parent cells are stored with the children flag and the child cell numbers, this data is used in the next step of adaptation where the decision is taken at a parent level for refinement or coarsening. For the parent cells there is an option of choosing the maximum gradients of the child cell or the average gradients of the cells for obtaining more resolution with adaptation. The solution is remapped on to the cells and it is made sure that the order of accuracy is conserved.

4.1 Solution Remapping & Transfer

Adaptation methods to reduce the solution errors of solving a PDE is highly dependent on the remapping or transfer of quantities from the base mesh to the adapted mesh and while coarsening of the mesh. It is important to remap the solution quantities like the integral quantities such as mass or energy and pointwise quantities such as diffusivity [13]. The remapping of both these quantities have to be done very accurately and more importantly for the integral quantities it has to be conservative. Consider the density of a child, it must be transferred such that the total mass of the parent is conserved. If a group of elements are to be coarsened the integral quantities can be summed up over the children and can be assigned to the parent. With the pointwise quantities the children cells can be averaged weighted by their volume. For conditions of refinement, the mass can be distributed equally over the children or a linear reconstruction of the density can be made over the parent and integrate over the child [13,14]. In the proposed algorithm here, the solution transfer between the refinement levels is treated differently where a volume average is take and remapped to the cells and for refinement the quantities of the parent are mapped as it is to the children cells. Using a summation of masses of the children and passing it to the parent may be a poor choice and might lead to lower order of accuracy.

4.2 Methodology

The AMR algorithm is employed with MUSCL and WENO type of spatial discretisation schemes on hybrid unstructured meshes based on the implementation of [18,19], where they have been applied to a range of inviscid, laminar, transitional and turbulent flows simulations [14-17,20] demonstrating the advantages of high-order schemes in conjunction with hybrid unstructured meshes. The HLLC Riemann solver is used along with an explicit 3rd-order three stage strong stability preserving Runge-Kutta time stepping for advancing the solution in time.

5 RESULTS & DISSCUSSION

The developed AMR is tested on two test cases namely the cylinder explosion problem and the forward facing step. Both the cases demonstrates the capability of dynamic mesh adaptation through time and changing gradients over single element and hybrid mesh domains. The tests are done with three numerical schemes namely muscl second order, weno 3rd order and weno 5th order to study the AMR behavior over increasing order of numerical accuracy.

5.1 Cylinder Explosion Test Case

The two dimensional Euler equation for gas dynamics to solve the considered problem is given as,

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ \rho u_x\\ \rho u_y\\ E\end{bmatrix}+\frac{\partial}{\partial x}\begin{bmatrix}\rho u_x\\ \rho u_x^2+p\\ \rho u_x u_y\\ (E+p)u_x\end{bmatrix}+\frac{\partial}{\partial y}\begin{bmatrix}\rho u_y\\ \rho u_y u_x\\ \rho u_y^2+p\\ (E+p)u_y\end{bmatrix}=0 \tag{1}$$

Where the pressure p is related to the conserved quantities through the equation of state,

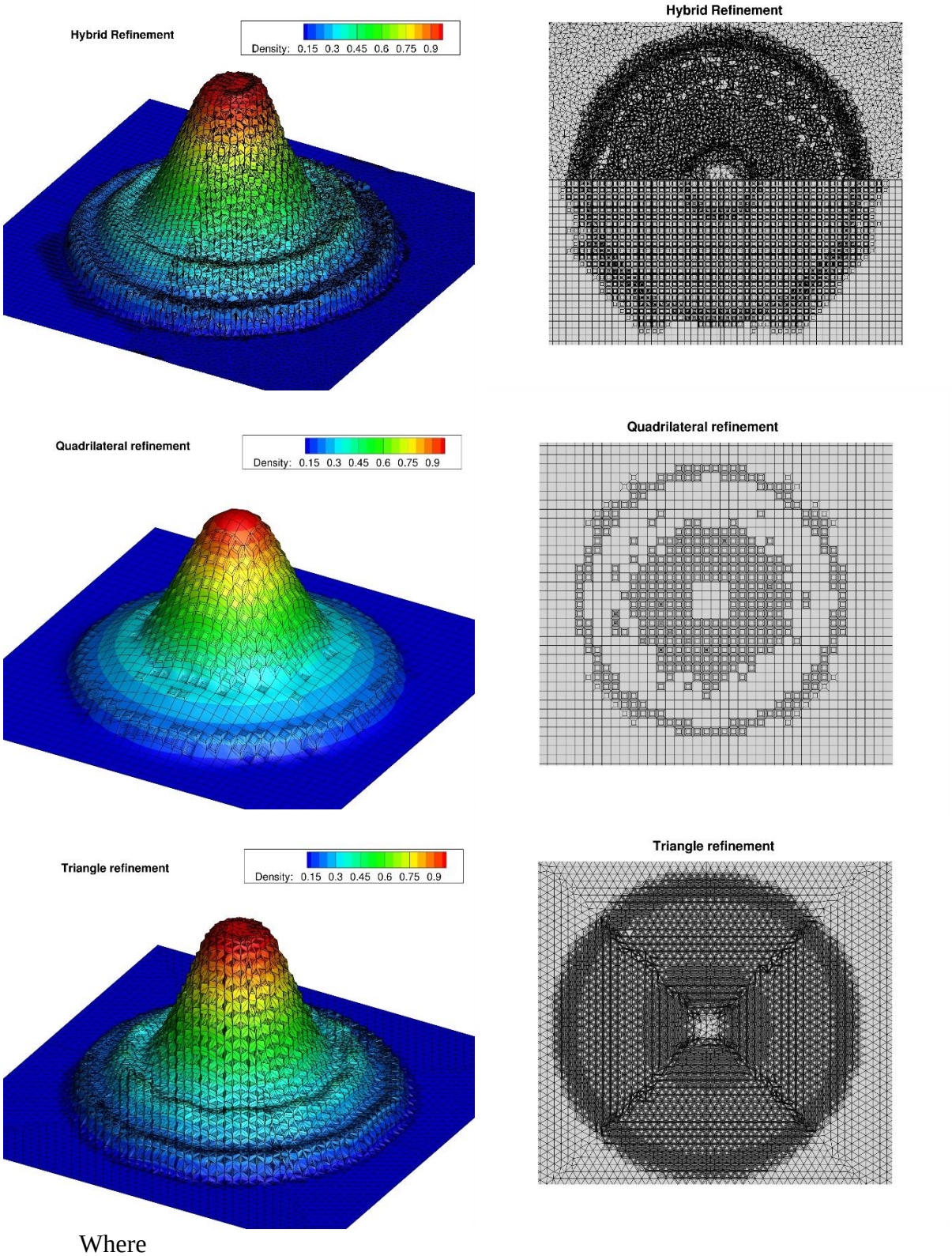
$$p=(\gamma-1)(E-\frac{1}{2}\rho(u_x^2+u_y^2)) \tag{2}$$

With γ=1.4, which is solved on a square domain in the xy-palne. The initial condition has a circular discontinuity of radius 0.4 centered at (1, 1). The initial data for the problem is defined on a non-dimensional domain of [-1:1] x[-1:1] and has two regions of constant but different values of gas parameters, the initial conditions are,

$$(\rho, p)=\begin{cases}(1.0,1.0), r\le 0.4\\ (0.125,0.1), r>0.4\end{cases}, u=v, r^2=x^2+y^2 \tag{3}$$

For the cylinder explosion problem in 2D, a set of three meshes namely a quad mesh with 1600 quadrilaterals, a triangular mesh with 5458 triangle elements and a hybrid mesh with 1711 and 5250 quads and tris respectively are used as the base mesh. These meshes are run on three numerical schemes muscl2, weno3 and weno5 to capture the solution with increasing numerical accuracy. The AMR for all the cases have been set to be moderately aggressive, the simulation is run for a time of two seconds with adaptation set to happen every 0.09 seconds which accounts for 3 adaptations during the course of the simulation. We expect the results to be sharper and where there are sharp gradients and also to conserve the parts where the solution is smooth with varying AMR settings. Refinement levels of l2, l3 and l4 were used for all conditions and the varying levels of non-dimensionalised density was used as the gradient markers for marking the cells for adaptation.

The solution exhibits a circular shock wave propagating away from the center, a circular contact surface travelling in the same direction and a circular rarefaction moving towards the origin. As time evolves it can be observed that a complex wave pattern emerges. The circular shock wave moves outwards and becomes weaker, the contact surface also follows the shock and becomes weaker and at some point in time the contact comes to rest and starts to move inwards. The rarefaction moving towards the center reflects and over expands the flow to create an inward moving shock wave which implodes to the origin, reflects and moves out colliding with the contact surface.



Where

Figure 5. Cylinder explosion with strategy_1 hybrid, quad and tri grid on weno3 scheme, meshes at final time step

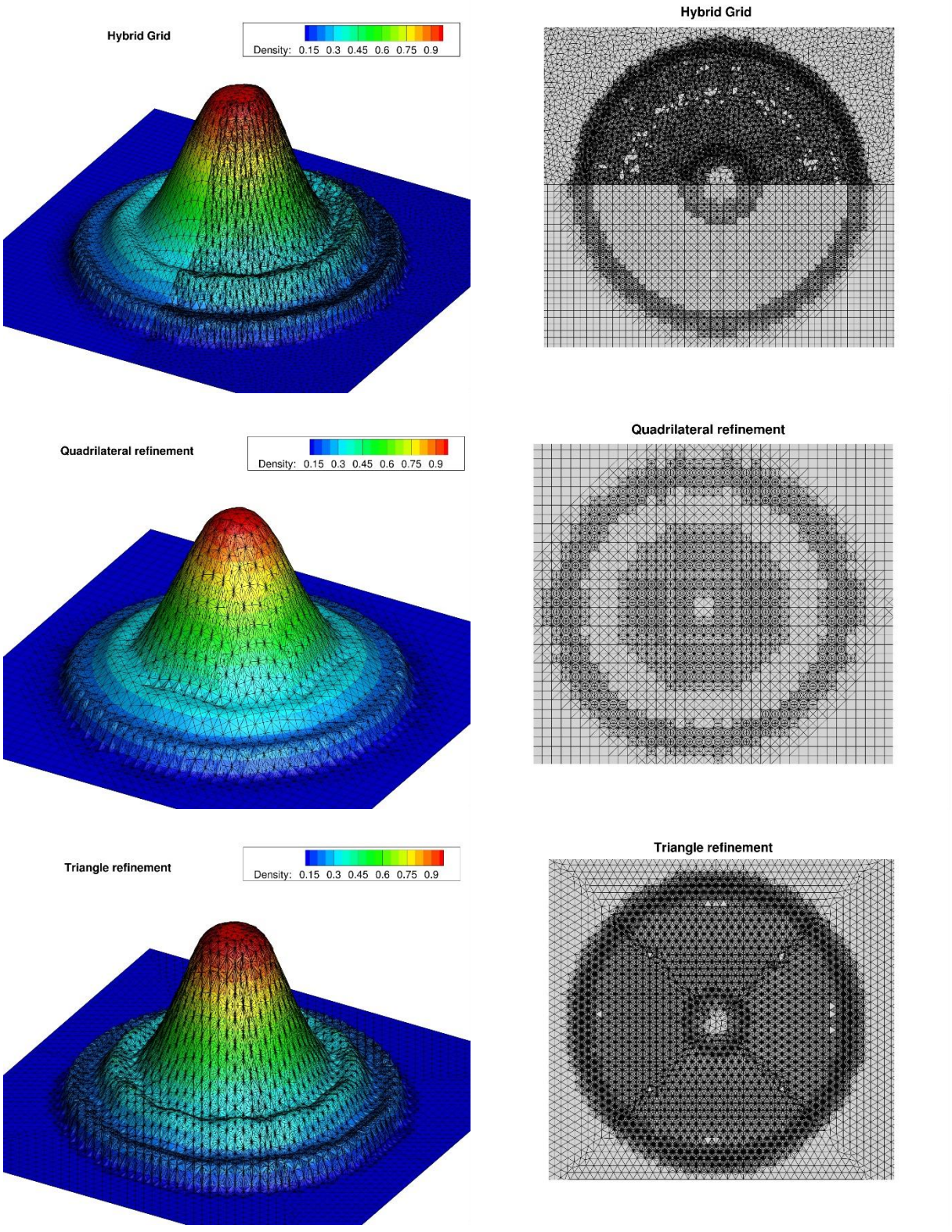


Figure 6.Cylinder explosion with strategy_2 hybrid, quad and tri grid on weno3 scheme, meshes at final time step

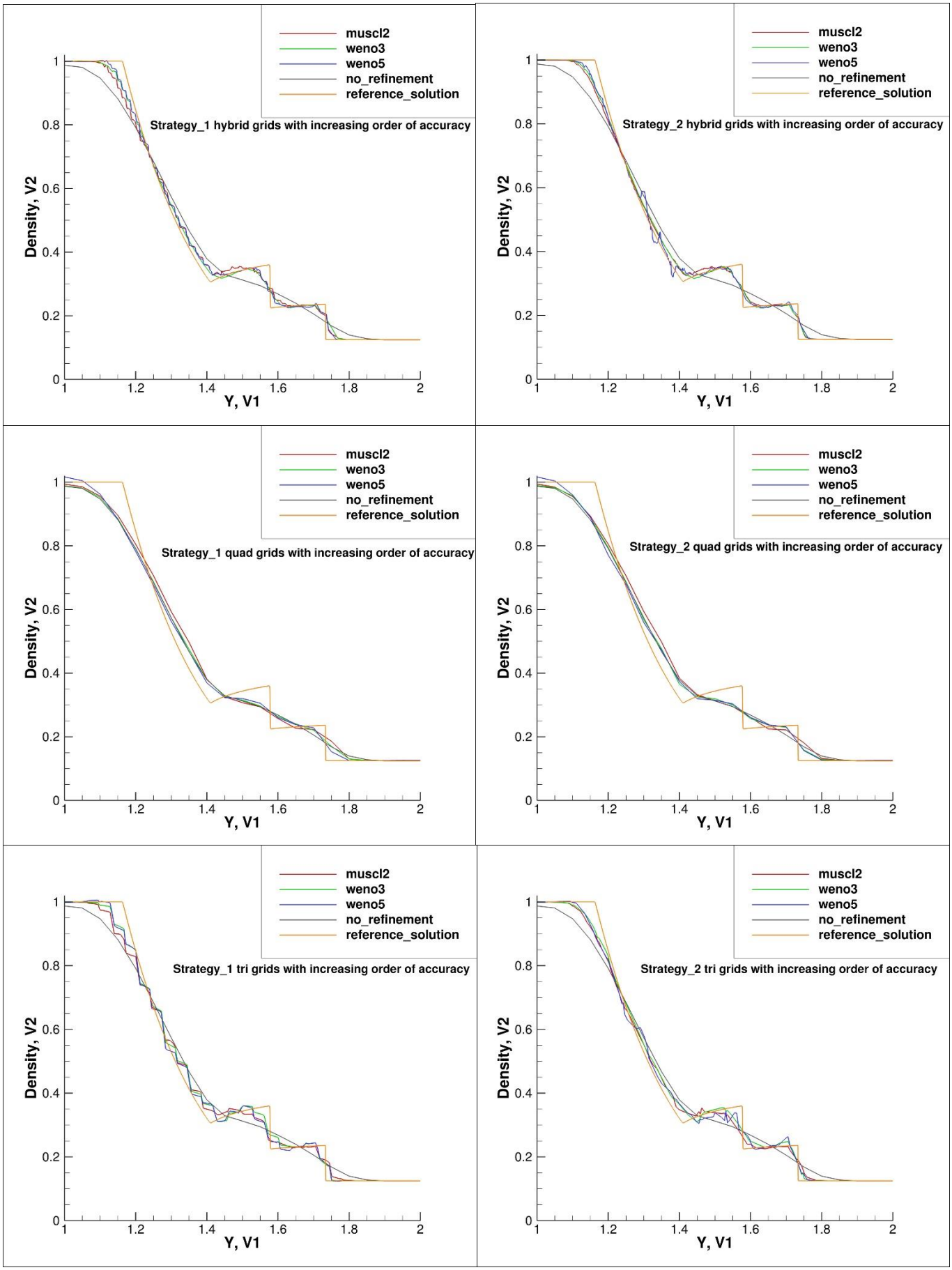


Figure 7.Density plots on different grids with higher order schemes

Figures 5 & 6 show the visualization of density and the final mesh at the end of the simulation, for both the adaptation strategies which demonstrates the capability of the algorithm to perform dynamic adaptation with any element type separately and also on hybrid grids. It can be observed that the use of strategy 2 refinement methods the solutions are comparatively smoother than the refinement of grids with strategy1.

In figure 7 we have the density plots compared with increasing order of numerical schemes which helps us to study the performance of AMR with higher order schemes. These results are compared to t an analytical solution to check for the accuracy of the schemes with AMR. It can be observed that all the numerical schemes capture the peak with all the grids except for quads which tend to over predict with a weno5 numerical scheme on both the strategies. The quad grid with both the strategies and with higher order numerical schemes fail to capture the total diffusivity happening in the simulation, but they tend to stay smooth throughout. On the other hand with the hybrid and triangular grids the diffusivity capturing is better and closer to the reference solution but comes with a noise generated due to the weno weights of the triangular elements present in the grid. It is important to notice that the solution is less susceptible to noise with hybrid and tri grids with the use of the strategy 2 for refinement.

5.2 Forward Facing Step Test Case

The forward facing step is a classic test case to study the flow separation and recirculation caused by a sudden contraction in a channel. For this case depending on the ratio of the boundary layer thickness at the step to the step of the height there might be one to three recirculation regions, one upstream, the other in the downstream and the other just immediately above the step, the flow might also separate from the upper sharp corner of the step generating a recirculation region behind the step.

For the case under consideration a domain where length (L) of the domain is equal to 3 units and the height (H) of the domain is equal to 1 unit, the step is located at (l) 0.2 units from the inlet and with a height (hl) of 0.2 units is considered,

$$\begin{aligned} 0 \le x \le L, &\begin{cases} y = 0 \\ y = H \end{cases} \{\phi \\ 0 \le y \le H, &\begin{cases} x = 0 \\ x = L \end{cases} \end{aligned} \tag{4}$$

$where, \phi = u, v, p$

The above equation (4) describes the boundary conditions, where the boundaries of the geometry are considered to be a wall, friction is created at this wall due to the viscosity and momentum of the fluid (air). The inlet is taken as the velocity inlet where the velocity is given in the u direction equal to mach 3, the outlet of the channel is considered as the pressure outlet. The pressure is taken as 1atm and the density as 1.4 kg/m^3 for the problem which is solved using the Euler equations as mentioned in equations (1) and (2).

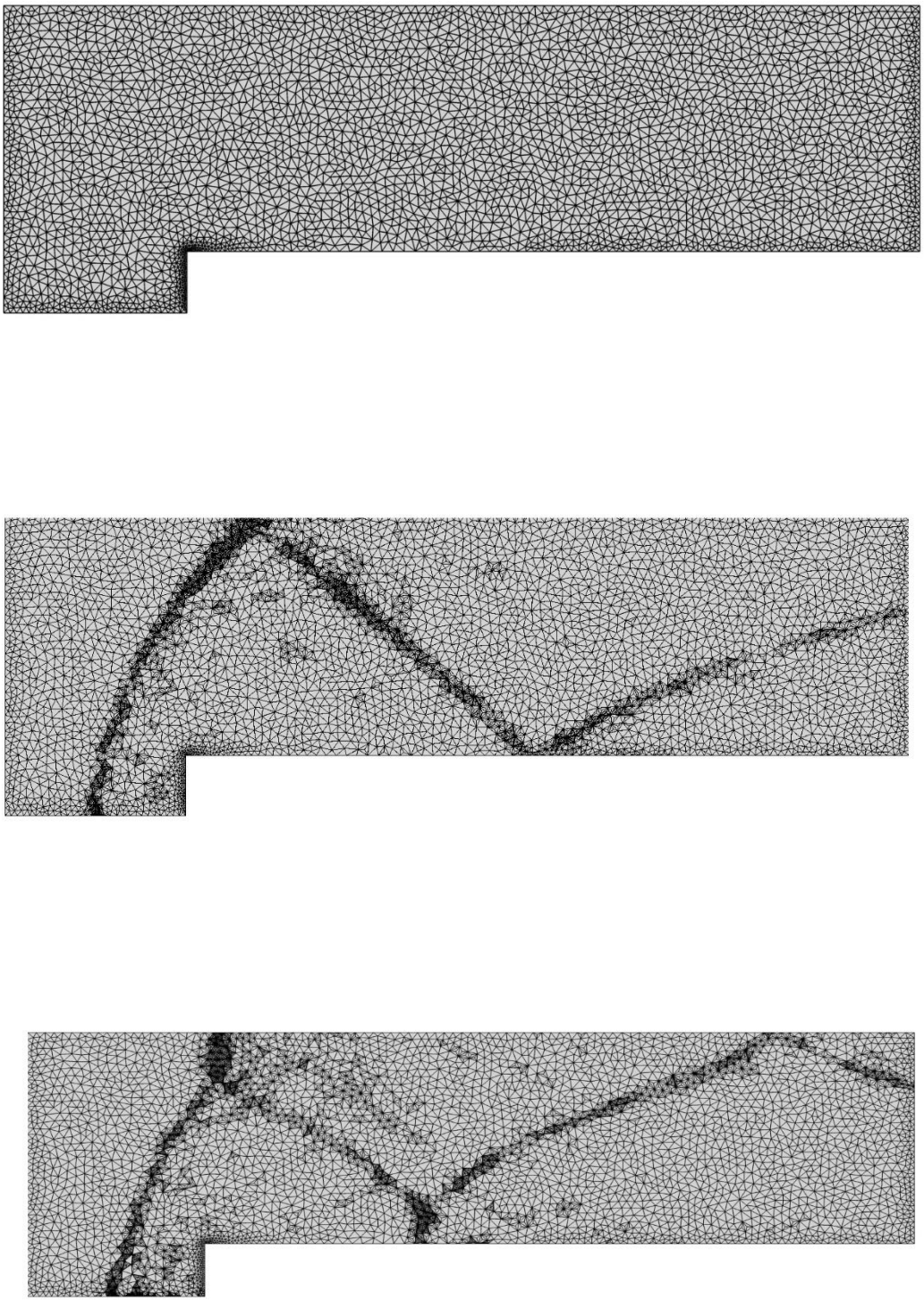


Figure 8. Forward facing step, initial mesh, mest at t=2.0secs, final mesh

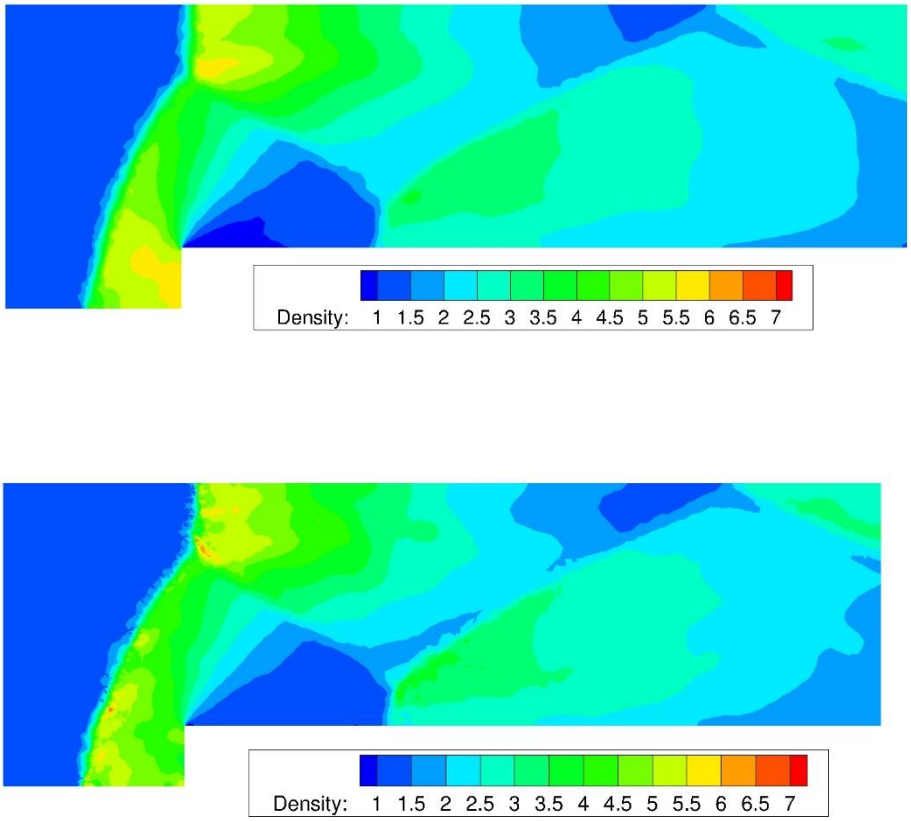


Figure 9. Contours of density, with no-refinement and with refinement

The domain is made up of triangular mesh which are moderately coarse for this setup. The simulation is run with a weno3 numerical scheme and a moderately aggressive AMR setup with second, third and fourth levels of refinement for a total simulation time of 4 seconds and AMR being activated every 0.9 seconds based on the gradients of density form the simulation, hence we would have five levels of adaptation happening in the course of the simulation. Figure 8 shows the triangular mesh of the case at the initial instance, mesh with adaptation at t=2.0secs and the mesh at t=4.0 secs.

When the flow enters through the left boundary at the velocity of Mach 3 and exits through the right pressure outlet boundary. During the first few time steps a strong shock is generated due to the step and an expansion wave appears originating from the top corner of the step. The shock then moves progressively towards the top wall and gets reflected at the wall, after a few tome steps the flow exhibits several shock wave reflections at the top and bottom walls before the shock exits the domain. Figure 9 shows the contours of density with a case run with no refinement and the case run with the AMR. It can be observed that the AMR helps in capturing the propagating shock with a better definition and also helps in preserving solution accuracy at places where the gradients are not that steep. This case clearly demonstrates the capability of tracking steep gradients by the AMR by accurately capturing the shock being propagated through the domain.

6 CONCLUSION

This paper proposed a mesh adaptation procedure for hybrid unstructured grids that gives a conformal mesh with node based refinement which follows a tree based hierarchical data structure where the decision on adaptation is always taken at the parent level. The quality of the refined mesh is kept comparatively high even with a few irregular elements present in the grid. From the test cases considered it can be said that the AMR efficiently performs the refinement where required and coarsens where it is not. The advantage of this method developed here is that the refinement or the coarsening process can be done as local as possible and also can be done per cell basis which gives the user more control over the process and is very essential and would aid in better parallelisation of the algorithm.

Solution transfer or remapping of solution of from child to parent and vice versa has always been a concern for conserving the integral and pointwise quantities of the solution, an efficient method has been devised with this algorithm where solution averaging based on volume weights have been done for remapping solution between the different levels of refinement and solution from the parent is remapped as a whole to the child cells, hence conserving the quantities. It was observed that the quadrilateral refinement was quite smooth than the triangle and hybrid refinement but failed to capture the diffusivity in the solution. Strategy 2 is a better option for refinement as there was less noise observed with triangular and hybrid grids.

This algorithm in conjunction with the high order solver can be used for solving a variety of problems pertaining to capturing of steep gradients and also can perform well with RANS cases. The present work will be extended to 3D domain for hexahedral, tetrahedral and prism elements and also the AMR algorithm would be parallelized.

References

[1] Bass, J. M. (1987). Adaptive finite element methods for a class of evolution problems in viscoplasticity. *International Journal of Engineering Science*, pp 623-653.

[2] Botkin, M. E.-P. (2002, April). Adaptive refinement of quadrilateral finite element meshes based upon a posteriori error estimation of quantities of interest. . *In Proceedings of the AIAA SDM Conference*, (pp. 22-25). Boulder, CO.

[3] Dumbser, M. Z. (2013). ADER-WENO finite volume schemes with space–time adaptive mesh refinement. . *Journal of Computational Physics*, *248*, pp 257-286.

[4] Garimella, R. (2009). Conformal refinement of unstructured quadrilateral meshes. *In Proceedings of the 18th International Meshing Roundtable Springer Berlin Heidelberg*, pp. 31-44.

[5] Ito, Y. S. (2009). Octree-based reasonable-quality hexahedral mesh generation using a new set of refinement templates. . *International Journal for Numerical Methods in Engineering*, pp.1809-1833.

[6] Liu, A. a. (1994). On the shape of tetrahedra from bisection. *Mathematics of computation 63.207*, pp.141-154.

[7] Sandhu, J. S. (1995). Hierarchical mesh adaptation of 2D quadrilateral elements. *Engineering fracture mechanics 50(5/6)*, pp.727-735.

[8] Schneiders, R. (1996). Refining quadrilateral and hexahedral element meshes. *In:proceedings of the 5th International Conference on Numerical Grid Generation in Computational Fluid Simulations,*, (pp. 679-688). Mississipi State University.

[9] Staten, M. L. (2008). A methodology for quadrilateral finite element mesh coarsening. *Engineering with Computers*, pp.241-251.

[10] Tchon, K. F. (2004). Automated refinement of conformal quadrilateral and hexahedral meshes. . *International Journal for Numerical Methods in Engineering*, pp.1539-1562.

[11] Thompson, J. F. (1985). *Numerical grid generation: foundations and applications (Vol. 45).* Amsterdam: North-holland.

[12] Zhang, Y. &. (2006). Adaptive and quality quadrilateral/hexahedral meshing from volumetric data. *Computer methods in applied mechanics and engineering*, pp 942-960.

[13] Garimella, R. K. ((2007)). An efficient linearity and bound preserving conservative interpolation (remapping) on polyhedral meshes. *Computers & fluids, 36(2)*, 224-237.

[14] Bathe, K. J. (1997). On the constraint function method for contact problems. *Computers & Structures, 64(5)*, 1069-1085.

[15] A. F. Antoniadis, D. Drikakis, I. Kokkinakis, P. Tsoutsanis, and Z. Rana (2015). Highorder methods for hypersonic shock wave turbulent boundary layer interaction flow. In AIAA, *20th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, number AIAA 2015-3524, Glasgow, Scotland, July 2015.

[16] A. F. Antoniadis, P. Tsoutsanis, and D. Drikakis(2015). Numerical accuracy in rans computations of high-lift multi-element airfoil and aircraft configurations. In AIAA, editor, *53rd AIAA Aerospace Sciences Meeting*, number AIAA 2015-0317, Kissimmee, Florida, January 2015.

[17] D. Drikakis, A. F. Antoniadis, P. Tsoutsanis, I. Kokkinakis, and Z. Rana(2015). Azure: An advanced cfd software suite based on high-resolution and high-order methods. In AIAA, editor, *53rd AIAA Aerospace Sciences Meeting*, number AIAA 2015-0813, Kissimmee, Florida, January 2015.

[18] V.A. Titarev, P. Tsoutsanis, and D. Drikakis(2010). Weno schemes for mixed-element unstructured meshes. *Communications in Computational Physics*, 8(3):585–609.

[19] P. Tsoutsanis, V.A. Titarev, and D. Drikakis (2011). Weno schemes on arbitrary mixedelement unstructured meshes in three space dimensions. *Journal of Computational Physics*, 230 (4):1585–1601.

[20] P. Tsoutsanis, A.F. Antoniadis, and D. Drikakis(2014). Weno schemes on arbitrary unstructured meshes for laminar, transitional and turbulent flows. *Journal of Computational Physics*, 256:254–276.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

IMPLEMENTATION OF A LOW-MACH NUMBER MODIFICATION FOR HIGH-ORDER FINITE-VOLUME SCHEMES FOR ARBITRARY HYBRID UNSTRUCTURED MESHES

N.Simmonds[1], P.Tsoutsanis[2], and A.Gaylard[2]

[1,2] Fluid Mechanics and Computational Sciences, Cranfield University,
Cranfield MK43 0AL, United Kingdom
e-mail: n.simmonds@cranfield.ac.uk, panagiotis.tsoutsanis@cranfield.ac.uk

[3] Jaguar Land Rover
Banbury Road, Gaydon, Warwick, CV35 0RR
e-mail: agaylar1@jaguarlandrover.com

Keywords: Instructions, ECCOMAS Congress, Computational Methods, Engineering Sciences, Proceedings.

Abstract. *An implementation of a novel low-mach number treatment for high-order finite-volume schemes using arbitrary hybrid unstructured meshes is presented in this paper. Low-Mach order modifications for Godunov type finite-volume schemes have been implemented successfully for structured and unstructured meshes, however the methods break down for hybrid mesh topologies containing multiple element types. The modification is applied to the UCNS3D finite-volume framework for compressible flow configurations, which have been shown as very capable of handling any type of grid topology. The numerical methods under consideration are the Monotonic Upstream-Centered Scheme for Conservation Laws (MUSCL) and the Weighted Essentially Non-Oscillatory (WENO) schemes for two-dimensional mixed-element type unstructured meshes. In the present study the HLLC Approximate Riemann Solver is used with an explicit TVD Runge-Kutta 3rd-order method due to its excellent scalability. These schemes (up to 5th-order) are applied to well established two-dimensional and three-dimensional test cases. The challenges that occur when applying these methods to low-mach flow configurations is thoroughly analysed and possible improvements and further test cases are suggested.*

1 Introduction

Obtaining accurate solutions in an efficient manner for a wide-range of flow problems where the compressible Navier-Stokes are used is the main goal of the majority of the numerical methods and schemes developed in this context. The most challenging part of the development of all the high-resolution numerical schemes used is that they should be adaptive. Adaptive in the sense of identifying regions of sharp-gradients often encountered in compressible flows, and preventing or eliminating any spurious oscillations that can occur and contaminate the solution; but at the same time they should be adaptive and achieve high-order of accuracy in smooth regions of the flow. However there is a delicate balance between the two requirements and is dependent upon the spatial discretisation method, the shock-capturing algorithms, the grid types, the Riemann solvers, the time-stepping algorithms and the quadrature rules used for integration to name a few.

The first generation numerical methods for unstructured grids exhibited lower accuracy and were computationally more demanding than structured grids. However, the numerical methods for unstructured grids have matured and numerous elegant approaches [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13] and algorithms have been developed in the finite volume framework for a wide range of applications for Computational Fluid Dynamics(CFD). Other state-of-the-art approaches are also available, such as the Discontinuous Galerkin (DG) [14, 15, 2, 16, 11, 17], and Spectral Finite-Volume (SFV) methods [18, 19, 20, 12, 21, 22] that have been successfully applied for CFD applications. For the finite volume framework the first class of high-resolution methods developed for unstructured grids included the ENO type schemes [23, 24], followed by the WENO type schemes [25, 26, 27, 28]. In the WENO case, the high-order accuracy was achieved by non-linearly combining a series of high-order reconstruction polynomials arising from a series of reconstruction stencils. Recently, a class of WENO type methods [9, 8] has been successfully extended to hybrid unstructured meshes with various geometrical shapes such as tetrahedrals, hexahedrals, prisms, and pyramids. The schemes can achieve the very high order of spatial accuracy across interfaces between cells of different types, and at the same time essentially non-oscillatory profiles are produced for discontinuous solutions. This gives greater flexibility to handle complex geometrical shapes in an efficient and accurate manner.

For the majority of the finite-volume numerical methods for compressible flows their dissipation characteristics are proportional to the speed of sound, therefore the low Mach number features are damped by the numerical scheme as noted by [29]. This is particular important at regions of the flow where the Mach number is low such as close to the wall at the boundary layer, and at vortices arising from shear layers.

In the present paper we present a modification to the output of the reconstruction, or the input of the Riemann solver in order to remove the Mach number dependence, and improve the resolution at low-Mach regions of the flow. It has to be noted that the

subject modification is different in nature from a preconditioning step used for changing the flow variables, since our main motivation is not to relax the restriction in terms of the time-step size, but rather to increase the resolution at the low-Mach regions. The original modification proposed by [29] is not directly transferable to any grid-type since different mesh elements have different dissipation characteristics, therefore a unified modification is implemented that is suitable for all element types and through the computational results presented we demonstrate the difference between the original modification, and the revised one. The nice feature of this modification is that is quite simple to implement in any compressible code, for any numerical scheme that uses a Riemann solver and the additional computational expense is negligible.

2 Governing Equations

The compressible inviscid Navier-Stokes equations are solved; writen in compact form as:

$$\frac{\partial \mathbf{U}(\mathbf{x},t)}{\partial t} + \nabla \mathbf{F_c}(\mathbf{U}) = 0 \tag{1}$$

where $\mathbf{U}$ is the vector of the conserved mean flow variables; $\mathbf{F_c}$, is the inviscid flux vector given by:

$$\begin{aligned}
\mathbf{U} &= \left[\rho, \rho u, \rho v, \rho w, \rho E\right]^T, \\
F_c^{\mathbf{x}} &= \left[\rho u, \rho u^2 + p, \rho uv, \rho uw, u(E+p)\right]^T \\
F_c^{\mathbf{y}} &= \left[\rho v, \rho uv, \rho v^2 + p, \rho vw, v(E+p)\right]^T \\
F_c^{\mathbf{z}} &= \left[\rho w, \rho uw, \rho vw, \rho w^2 + p, w(E+p)\right]^T
\end{aligned} \tag{2}$$

In the above equations, ρ is the density; u, v, w are the velocity components in x, y and z Cartesian directions, respectively. Calorically perfect gas is assumed where the total energy per unit mass is computed according to the equation of state as $E = p/\left(\gamma - 1\right) + (1/2)\rho(u^2+v^2+w^2)$, where p is the pressure, the ratio of specific heats is set as $\gamma = 1.4$ for air at normal atmoshperic conditions.

3 Numerical Framework

The discretization in a domain Ω is achieved by combining conforming arbitrary shaped elements of volume $|V_i|$. Integrating Eq. (1) over a mesh element using the finite-volume formulation the following ordinary differential equation is obtained

$$\frac{d\mathbf{U}_i}{dt} = -\frac{1}{|V_i|}\sum_{l=1}^{N_f}\sum_{\alpha=1}^{N_{qp}} \mathbf{F_c^{n,l}}\left(\mathbf{U}(\mathbf{x}_\alpha,t)\right)\omega_\alpha |A_l| \tag{3}$$

where $\mathbf{U}_i$ is the volume averaged conserved variables in a volume element, N_f is the number of faces per element, N_{qp} is the number of quadrature points used for approximating the surface integrals. $|A_l|$ is the surface area of the corresponding face, and α corresponds to different Gaussian integration points $\mathbf{x}_\alpha$ and weights ω_α over the face. The weight and distribution of the quadrature points depend upon the Gaussian rule order, higher integration rule will result in enhanced intercell flux approximation. The convective flux tensors are defined as:

$$\mathbf{F}_c^{\mathbf{n,l}} = F_c^{\mathbf{x}} n_x + F_c^{\mathbf{y}} n_y + F_c^{\mathbf{z}} n_z \tag{4}$$

where n_x, n_y and n_z are the Cartesian components of the normal vector on the intercell surface. The intercell fluxes are computed based on volume averaged quantities which are obtained by an interpolation technique. The solution is obtained by a polynomial reconstruction from cell-averaged data. The following sections describe the methodology adopted for space and time discretization.

3.1 Spatial Discretization

The spatial discretisation is based on the approach of [8, 9], which is suitable for unstructured meshes with various types of element shapes in 2D and 3D, and it has been previously used successfully for laminar, transitional and turbulent flows [13, 30, 31, 32, 33]. Therefore, only the key characteristics of this approach are going to be described in this paper. The main objective of the reconstruction process is to build a high-order polynomial $p_i(x, y, z)$ of arbitrary order r, for each considered element V_i that has the same average as a general quantity $\mathbf{U}_i$. This can be formulated as

$$\mathbf{U}_i = \frac{1}{|V_i|} \int_{V_i} \mathbf{U}(x, y, z)\, dV = \frac{1}{|V_i|} \int_{V_i} p_i(x, y, z)\, dV. \tag{5}$$

The reconstruction is carried out in a transformed system of coordinates in order to minimize scaling effects that appear in stencils consisting of elements of different size as well as to improve the condition number of the system of equations [13, 8]. The transformation is achieved by decomposing each element into tetrahedrals.

The reconstruction polynomial at the transformed cell V_i' is expanded over local polynomial basis functions labeled as $\phi_k(\xi, \eta, \zeta)$, which are given by

$$p(\xi, \eta, \zeta) = \sum_{k=0}^{K} a_k \phi_k(\xi, \eta, \zeta) = \mathbf{U}_0 + \sum_{k=1}^{K} a_k \phi_k(\xi, \eta, \zeta) \tag{6}$$

where ξ, η, ζ are the coordinates in the reference system, a_k are the degrees of freedom and the upper index in the summation of expansion K is related to the order of the polynomial r by $K = \frac{1}{6}(r+1)(r+2)(r+3) - 1$ for 3D and $K = \frac{1}{3}(r+1)(r+2) - 1$ for

2D . For computing the degrees of freedom a_k, a minimum of K cells is required in the stencil in addition to the target cell. Using the minimum possible number of cells in the stencil, $M \equiv K$, may produce ill-conditioned systems, hence using $M = 2 \cdot K$ improves the robustness of the method as described in [8, 34]. A linear least-square method is adopted to enable the system of equations of 6 for the unknown degrees of freedom a_k. The final form of the linear system is solve with the QR decomposition algorithm. Discontinuous solutions are often encounter in external aerodynamics, thus limiting functions are essential for maintaning numerical stability and suppress any spurious oscillations. Two approaches are assessed in this work the MUSCL and the WENO schemes.

3.1.1 MUSCL

The MUSCL scheme is employed in this work is partially based on the Barth and Jespersen slope limiter. The design of the scheme requires the minimum and maximum values in the stencil's neighborhood, i.e., $U_i^{min} = \min(U_i, U_l)$ and $U_i^{max} = \max(U_i, U_l)$, where $l = 1, ..L$; L is the total number neighbors of element i. The gradient of a general quantity is defined as ∇U_i and is an approximation of the solution gradient inside the element i. The gradient is computed during the reconstruction process by incorporating information from the entire central stencil. The scheme can be written as

$$U_{i\alpha} = U_i^c + \phi_i \nabla U_i \cdot \mathrm{x}_\alpha \tag{7}$$

where U_i^c is the value for the general quantity at the element centroid and x_α are the coordinates of the quadrature point. The limiter seeks the minimum value of the slope limiter for all the quadrature points that satisfy the following conditions

$$\phi_i = \min(\overline{\phi}_{i,m_1}, \overline{\phi}_{i,m_2}, ...\overline{\phi}_{i,M}) \tag{8}$$

Then, the limiting function is applied, composed by three different states according to the difference of the reconstructed value at the quadrature points of the considered element $U_{(i,\alpha)}$ and each of its neighbors U_l, yielding

$$\overline{\phi}_{i,\alpha} = \begin{cases} \min\left(1, \dfrac{U_i^{max} - U_{i,\alpha}}{U_l - U_{i,\alpha}}\right), & \text{if } U_l - U_{i,\alpha} > 0 \\ \min\left(1, \dfrac{U_i^{min} - U_{i,\alpha}}{U_l - U_{i,\alpha}}\right), & \text{if } U_l - U_{i,\alpha} < 0 \\ 1, & \text{if } U_l - U_{i,\alpha} = 0 \end{cases} \tag{9}$$

3.1.2 WENO

WENO schemes use a non-linear combination of various reconstruction polynomials, where each polynomial is weighted according to the smoothness of its solution. The

polynomials are given by

$$p_i^{\text{weno}} = \sum_{m=1}^{m_s} \omega_m p_m(\xi, \eta, \zeta) \tag{10}$$

where m_s is the total number of WENO stencils. Substituting back to Eq. (6) for $p_m(\xi, \eta, \zeta)$, we obtain the following expression

$$p_m(\xi, \eta, \zeta) = \sum_{m=0}^{K} a_k^{(m)} \phi_k(\xi, \eta, \zeta) \tag{11}$$

Using the condition that the sum of all weights is unity, yields

$$\begin{aligned} p_i^{\text{weno}} &= \mathbf{U}_0 + \sum_{k=1}^{K} \left(\sum_{m=0}^{m_s} \omega_m a_k^m \right) \phi_k(\xi, \eta, \zeta) \\ &\equiv \mathbf{U}_0 + \sum_{k=1}^{K} \tilde{a}_k \phi_k(\xi, \eta, \zeta) \end{aligned} \tag{12}$$

where $\tilde{a}_k$ are the reconstructed degrees of freedom; and the non-linear weight ω_m is defined by [13, 8, 34]:

$$\omega_m = \frac{\tilde{\omega}_m}{\sum_{m=0}^{m_s} \tilde{\omega}_m} \quad \text{where} \quad \tilde{\omega}_m = \frac{\lambda_m}{(\epsilon + \mathcal{I}_m)^b} \tag{13}$$

The smoothness indicator is given by

$$\mathcal{I}_m = \sum_{1 \leq |\beta| \leq r} \int_{\tilde{E}_i} \left(\mathcal{D}^\beta p_m(\xi, \eta, \zeta) \right)^2 (d\xi, d\eta, d\zeta) \tag{14}$$

where β is a multi-index, r is the polynomial's order, λ_m is the linear weight, and D is the derivative operator. The smoothness indicator is a quadratic function of the degrees of freedom (a_k^m) and can be expressed as a universal mesh-independent oscillation indicator matrix as defined in [34]. The WENO reconstruction can be carried out in terms of conserved or characteristic variables. In this work, the conserved variables have been employed to assess the low-Mach number reconstruction modification. The various reconstruction polynomials arise from different sets of stencils that satisfy some geometrical conditions. The reader is referred to [9, 8] for the definition of geometrical sectors., and references therein, for a detailed explanation of the different set of geometrical conditions and WENO characteristic reconstruction.

3.2 Low-Mach Number Modification

Our low-Mach number modification follows the previous work of [29], where the left and right state of the reconstructed velocities is modified in a linear way with respect to Mach number, so when the Mach number approaced zero the arithmetic mean of them is used in the following manner:

$$\begin{aligned} u_L^* &= \frac{(1+z)u_L + (1-z)u_R}{2} \\ u_R^* &= \frac{(1+z)u_R + (1-z)u_L}{2} \\ z &= min(1, max(M_L, M_R)) \end{aligned} \tag{15}$$

The Mach numbers are calculated based on the velocity magnitude of all the velocity components independent of the normal direction in which the flux is computed. However we investigate two approaches one where all the component of the reconstructed velocities are modified as the original implementation [29], and a second one where only the velocity components normal to the face are modified, based on the findings of [29] where it was highlighted that the tangential velocities require more numerical dissipation than the normal components.

3.3 Fluxes Approximation

Having reconstructed the cell-averaged solutions, the intercell fluxes can be evaluated. The Riemann problem is solved with the approximate Harten-Lax-van Leer-Contact (HLLC) solver of Toro [35]. The HLLC solver is also employed for the convective part of the turbulence transport equation. The HLLC flux function is given by

$$\hat{\mathrm{F}}^{\mathbf{HLLC}} \begin{cases} \hat{F}^-, & \text{if } 0 \leq S^-, \\ \hat{F}^{*-} = \hat{F}^- + S^- \left(\hat{W}^{*-} - \hat{W}^-\right), & \text{if } S^- \leq 0 \leq S^*, \\ \hat{F}^{*+} = \hat{F}^+ + S^+ \left(\hat{W}^{*+} - \hat{W}^+\right), & \text{if } S^* \leq 0 \leq S^+, \\ \hat{F}^+, & \text{if } 0 \geq S^+, \end{cases} \tag{16}$$

where

$$\hat{W}^{*\pm} = \rho^\pm \left(\frac{S^\pm - u^\pm}{S^\pm - S^*}\right) \begin{bmatrix} 1 \\ S^* \\ v^\pm \\ w^\pm \\ \frac{E^\pm}{\rho^\pm}\left(S^* - u^\pm\right) + \left(S^* + \frac{p^\pm}{\rho^\pm\left(S^\pm - u^\pm\right)}\right) \\ \frac{\tilde{\nu}}{\rho^\pm} \end{bmatrix} \tag{17}$$

$\hat{W}^{*\pm}$ is computed either for the considered element "$-$", or for its neighbour "$+$" and the wave speeds are estimated according to the sign.

3.4 Temporal discretisation

Having constructed the numerical fluxes $\mathbf{F}_{n,j}$ as expressed in the semi-discrete conservative formulation, the next step involves the advancement of the solution in time. The explicit SSP Runge-Kutta 3rd-order method [36] has been employed for the time integration

$$\left.\begin{aligned} \mathbf{U_i}^1 &= \mathbf{U_i}^n + \Delta t \cdot \mathbf{R}_i\left(\mathbf{U_i}^n\right) \\ \mathbf{U_i}^2 &= \tfrac{3}{4}\mathbf{U_i}^n + \tfrac{1}{4}\mathbf{U_i}^1 + \tfrac{\Delta t}{4} \cdot \mathbf{R}_i\left(\mathbf{U}^1\right) \\ \mathbf{U_i}^{n+1} &= \tfrac{1}{3}\mathbf{U_i}^n + \tfrac{2}{3}\mathbf{U_i}^2 + \tfrac{2\Delta t}{3} \cdot \mathbf{R}_i\left(\mathbf{U_i}^2\right) \end{aligned}\right\} \tag{18}$$

with $\mathbf{R}_i$ being the residual.

The time step Δt is computed as follows

$$\Delta t = K \min_i \frac{h_i}{S_i \cdot V_i}, \tag{19}$$

where h_i is the radius of the inscribed sphere of each cell i and V_i its corresponding volume, $K \leq 1/3$ is the CFL number for unsplit finite-volume schemes [35], and S_i is the maximum propagation speed in each cell i given by

$$S_i = sp_x \cdot n_x + sp_y \cdot n_y + sp_z \cdot n_z, \tag{20}$$

where

$$sp_x = |u + a|\,, \quad , sp_y = |v + a|\,, \quad sp_x = |w + a|\,,$$

with $\mathbf{n} = (n_x, n_y, n_z)$ being the outward unit normal vector and a is the speed of sound.

4 Results

This section presents the results obtained for a variety of cases to illustrate the influence of the low-mach modification compared to the previously implemented version.

4.1 2D Vortex Evolution

The two-dimensional vortex evolution problem is an important case to consider for convergence and accuracy. This is a well used problem for low-Mach testing and illustrating the power of high order methods, [37].

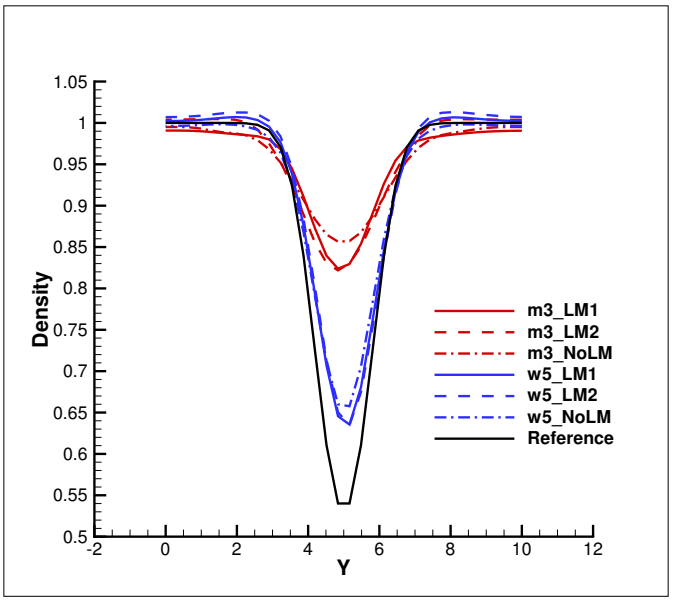


Figure 1: Density plots using a MUSCL 3rd and WENO 5th order schemes for the the vortex evolution flow problem on a quadrilatel mesh at $t = 2s$

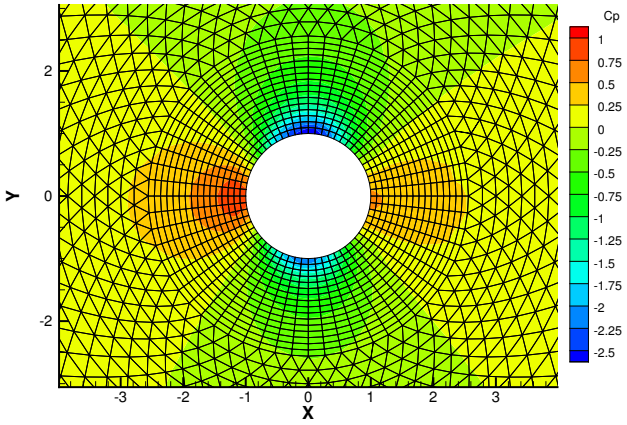
We examine a square domain of area $\Omega = 100$, [0,10]x[0,10], with periodic boundary conditions. The mean flow of the problem is $\rho = 1, P = 1, u = 1, v = 1$, therefore there is a diagonal flow through the domain moving to the top right corner. The flow is pertubated by an isotropic vortex in the u and v fields with a temperature $T = \frac{P}{\rho}$ and no pertubation in the entropy $S = \frac{P}{\rho^\gamma}$, i.e. $\delta S = 0$. The pertubations are given by

$$(\delta u, \delta v) = \frac{\epsilon}{2\pi} e^{0.5(1-r^2)}(5-y, x-5)$$

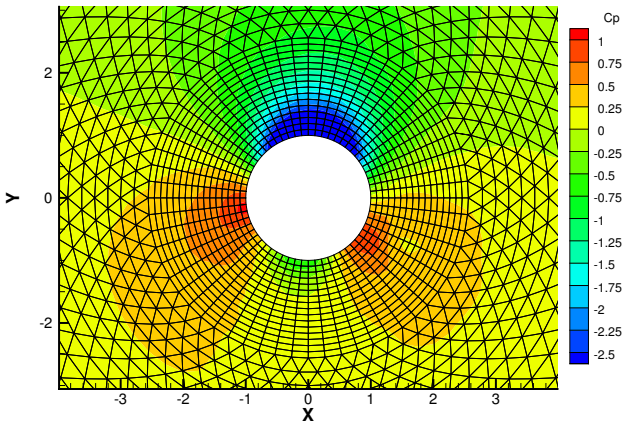
$$\delta T = -\frac{(\gamma-1)\epsilon^2}{8\gamma\pi^2} e^{1-r^2}$$

$$r^2 = (x-5)^2 + (y-5)^2$$

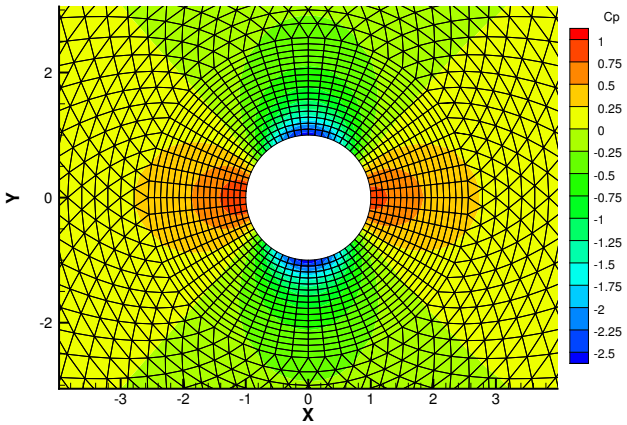
The solution is computed at $t = 2s$ for both low-mach corrections using a 3rd order MUSCL scheme and a 5th order WENO scheme. These are compared to the initial solution $t = 0s$ on the same grid of quadrilateral elements with 32x32 elements. From Fig. 1, it can immediately been seen that the higher order WENO 5 scheme reduces the noticeable dissipation of the MUSCL scheme, with and without the low-mach corrections. The influence of the low-mach corrections on the 3rd order MUSCL simulations is higher and a clear difference can be seen in its absence resulting in less dissipation. It is noted that for this particular test case there is not a large difference between the two modifications.



(a) No Modification



(b) Low Mach-number Modification of normal and tangential velocity components



(c) Low Mach-number Modification of normal velocity component

Figure 2: Pressure Coefficient plots using a WENO 3rd order scheme for the flow past a 2D cylinder at $t = 200s$

4.2 2D Low Mach number past a Cylinder

Examined is the steady-inviscid problem of flow past a circular cylinder with radius R=1. This is cut out of a cylindrical computational domain of radius R=20, with a free-stream inflow of Mach number $M_\infty = 0.1$, density $\rho_\infty = 1$ and the pressure $p_\infty = \frac{\rho_\infty}{\gamma}$, where $\gamma = 1.4$. The flow is initialised with a flow direction parallel to the x-axis.

The simulations were undertaken on two different mesh topologies, a completely triangular mesh with a linear piecewise construction for the cylinder and a hybrid mesh with prism cells near the wall of the cylinder before moving to triangular cells. The computations were ran until $t = 200s$ when the initial transients would have vanished and the residuals would have converged.

The numerical simulations on the fully triangular mesh did not suffer the same dissipation that can be found in the hybrid mesh. Presented in Fig. 2 are the hybrid mesh and pressure coefficient C_p distribution on a medium grid of containing 960 quadrilateral and 3316 triangular elements, computed using a WENO 3 reconstruction scheme. It can be seen that both the simulation without any treatment and the simulation using the second treatment suffered dissipation and did not exhibit the symmetry expected from this case. The presence of the first low-mach modification gives a much more accurate solution, with a clear symettry about $x = 0$.

4.3 2D Cylindrical Explosion

The presented case of an explosion in two dimension is outlined by Toro [35] as a classical Riemann solver problem. Contained within a square domain $[0, 2] \times [0, 2]$, exists a circle of radius $r = 0.4$ centered at (1,1), whose internal initial conditions vary from it's external initial conditions.

$$(\rho, P) = \begin{cases} (0.125, 0.1) & r > 0.4 \\ (1, 1) & r \leq 0.4 \end{cases}, u = v = 0, r^2 = x^2 + y^2$$

The solution is the result of a circular shock occurring from the Riemann problem which propagates outwards from the centre, this is followed by a circular contact discontinuity that propagates in the same direction. This is turns leads to a circular rarefaction travelling in the opposite direction to the aforementioned flow features. Over time the shock wave and contact discontinuity will become weaker, resulting in the contact discontinuity to stop and travel inwards while the rarefaction wave with reflect when it reaches the centre.

The solution presented in Fig. 3 was obtained after running the solution until $t = 0.2$. A reference solution for density was computer by solving the 1D problem on a very fine grid. The results indicate that the solver is capable of resolving the shock within 2-3 cells on the unstructured grid and 3-4 on the structured grid, without the need for a low-mach correction. The results using the WENO 5 reconstruction demonstrate that both low-mach corrections can be implemented without inducing any other numerical errors.

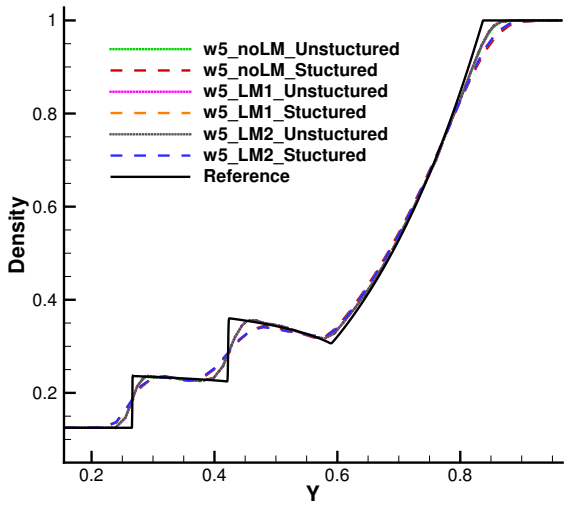


Figure 3: Density plots using Weno 5 numerical scheme with and without the low-mach corrections on both structured and unstructured grids.

5 Conclusions

The conclusion drawn from the subject study are listed below:

- A low-Mach number correction to the reconstruction of finite-volume numerical schemes improves the dissipation characteristics of the schemes
- A direct implementation of the original low-Mach number modification of [29] can not be used for all types of elements since they have different dissipation characteristics
- A revised modification is presented since the reconstructed tangential velocity components require more dissipation than the normal velocity components with respect to the interface between two cells
- The subject modification is quite robust to be utilised for all Mach number flows since it is not producing any spurious oscillations at regions of high-Mach number that can contaminate the solution.

Acknowledgments

The authors would like to thank Jaguar Land Rover for granting permission to publish this work. The work has been undertaken due to funding from EPSRC Award. Ref. Number: 13794.

REFERENCES

[1] M. Dumbser, M. Kaser, V. Titarev, E. Toro, Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems, Journal of Computational Physics 226 (1) (2007) 204–243.

[2] M. Dumbser, D. Balsara, E. Toro, C.-D. Munz, A unified framework for the construction of one-step finite volume and discontinuous galerkin schemes on unstructured meshes, Journal of Computational Physics 227 (18) (2008) 8209–8253.

[3] A. Haselbacher, A weno reconstruction algorithm for unstructured grids based on explicit stencil construction, 2005, pp. 3369–3378.

[4] W. Li, Y. Ren, Quadrature-free non-oscillation finite volume scheme for solving navier-stokes equations on unstructured grids, Vol. 1376, 2011, pp. 639–641.

[5] X. Nogueira, L. Cueto-Felgueroso, I. Colominas, F. Navarrina, M. Casteleiro, A new shock-capturing technique based on moving least squares for higher-order numerical schemes on unstructured grids, Computer Methods in Applied Mechanics and Engineering 199 (37-40) (2010) 2544–2558.

[6] C. Ollivier-Gooch, Quasi-eno schemes for unstructured meshes based on unlimited data-dependent least-squares reconstruction, Journal of Computational Physics 133 (1) (1997) 6–17.

[7] C. O. Gooch, M. V. Altena, A high-order-accurate unstructured mesh finite-volume scheme for the advection-diffusion equation, Journal of Computational Physics 181 (2) (2002) 729–752.

[8] P. Tsoutsanis, V. Titarev, D. Drikakis, Weno schemes on arbitrary mixed-element unstructured meshes in three space dimensions, Journal of Computational Physics 230 (4) (2011) 1585–1601.

[9] V. Titarev, P. Tsoutsanis, D. Drikakis, Weno schemes for mixed-element unstructured meshes, Communications in Computational Physics 8 (3) (2010) 585–609.

[10] W. Wolf, J. Azevedo, High-order eno and weno schemes for unstructured grids, International Journal for Numerical Methods in Fluids 55 (10) (2007) 917–943.

[11] Z. Xu, Y. Liu, C.-W. Shu, Hierarchical reconstruction for discontinuous galerkin methods on unstructured grids with a weno-type linear reconstruction and partial neighboring cells, Journal of Computational Physics 228 (6) (2009) 2194–2212.

[12] Z. Xu, Y. Liu, C.-W. Shu, Hierarchical reconstruction for spectral volume method on unstructured grids, Journal of Computational Physics 228 (16) (2009) 5787–5802.

[13] P. Tsoutsanis, A. Antoniadis, D. Drikakis, Weno schemes on arbitrary unstructured meshes for laminar, transitional and turbulent flows, Journal of Computational Physics 256 (2014) 254–276.

[14] B. Cockburn, C.-W. Shu, Runge-kutta discontinuous galerkin methods for convection-dominated problems, Journal of Scientific Computing 16 (3) (2001) 173–261.

[15] J. Dennis, R. Nair, H. Tufo, M. Levy, T. Voran, Development of a scalable global discontinuous galerkin atmospheric model, Int. J. Comp. Sci. Eng.

[16] A. Uranga, P.-O. Persson, M. Drela, J. Peraire, Implicit large eddy simulation of transition to turbulence at low reynolds numbers using a discontinuous galerkin method, International Journal for Numerical Methods in Engineering 87 (1-5) (2011) 232–261.

[17] J. Zhu, J. Qiu, C.-W. Shu, M. Dumbser, Runge-kutta discontinuous galerkin method using weno limiters ii: Unstructured meshes, Journal of Computational Physics 227 (9) (2008) 4330–4353.

[18] Z. Wang, Spectral (finite) volume method for conservation laws on unstructured grids. basic formulation, Journal of Computational Physics 178 (1) (2002) 210–251.

[19] Z. Wang, Y. Liu, Spectral (finite) volume method for conservation laws on unstructured grids. ii. extension to two-dimensional scalar equation, Journal of Computational Physics 179 (2) (2002) 665–697.

[20] Z. Wang, L. Zhang, Y. Liu, Spectral (finite) volume method for conservation laws on unstructured grids iv: Extension to two-dimensional systems, Journal of Computational Physics 194 (2) (2004) 716–741.

[21] Y. Zhou, Z. Wang, Implicit large eddy simulation of transitional flow over a sd7003 wing using high-order spectral difference method, 2010.

[22] C. Breviglieri, A. Maximiliano, E. Basso, J. Azevedo, Improved high-order spectral finite volume method implementation for aerodynamic flows, 2009.

[23] R. Abgrall, On essentially non-oscillatory schemes on unstructured meshes: analysis and implementation, Journal of Computational Physics 114 (1) (1994) 45–58.

[24] D. Stanescu, W. Habashi, Essentially nonoscillatory euler solutions on unstructured meshes using extrapolation, AIAA Journal 36 (8) (1998) 1413–1416.

[25] G. S. Jiang, C. W. Shu, Efficient implementation of weighted eno schemes, Journal of Computational Physics 126 (1) (1996) 202–228.

[26] J. Shi, C. Hu, C.-W. Shu, A technique of treating negative weights in weno schemes, Journal of Computational Physics 175 (1) (2002) 108–127.

[27] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, Journal of Computational Physics 77 (2) (1988) 439–471.

[28] O. Friedrich, Weighted essentially non-oscillatory schemes for the interpolation of mean values on unstructured grids, Journal of Computational Physics 144 (1) (1998) 194–212.

[29] B. Thornber, A. Mosedale, D. Drikakis, D. Youngs, R. Williams, An improved reconstruction method for compressible flows with low mach number features, Journal of Computational Physics 227 (10) (2008) 4873–4894.

[30] P. Tsoutsanis, I. Kokkinakis, L. Konozsy, D. Drikakis, R. Williams, D. Youngs, An investigation of the accuracy and efficiency of structured and unstructured, compressible and incompressible methods for the vortex pairing problem, Computer Methods in Applied Mechanics and Engineering 293 (2015) 207–231.

[31] A. F. Antoniadis, P. Tsoutsanis, D. Drikakis, Numerical accuracy in rans computations of high-lift multi-element airfoil and aircraft configurations, in: AIAA (Ed.), 53rd AIAA Aerospace Sciences Meeting, no. AIAA 2015-0317, Kissimmee, Florida, 2015.

[32] D. Drikakis, A. F. Antoniadis, P. Tsoutsanis, I. Kokkinakis, Z. Rana, Azure: An advanced cfd software suite based on high-resolution and high-order methods, in: AIAA (Ed.), 53rd AIAA Aerospace Sciences Meeting, no. AIAA 2015-0813, Kissimmee, Florida, 2015.

[33] A. F. Antoniadis, D. Drikakis, I. Kokkinakis, P. Tsoutsanis, Z. Rana, High-order methods for hypersonic shock wave turbulent boundary layer interaction flow, in: AIAA (Ed.), 20th AIAA International Space Planes and Hypersonic Systems and Technologies Conference, no. AIAA 2015-3524, Glasgow, Scotland, 2015.

[34] M. Dumbser, M. Käser, V. A. Titarev, E. F. Toro, Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems, Journal of Computational Physics 226 (1) (2007) 204–243.

[35] E. F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics - A Practical Introduction, Berlin: Springer-Verlag, 1997.

[36] S. Gottlieb, On high order strong stability preserving runge-kutta and multi step discretizations, Journal of Scientific Computing 25 (112) (2005) 105–128.

[37] C. Hu, C. W. Shu, Weighted essentially non-oscillatory schemes on triangular meshes, Journal of Computational Physics 150 (1) (1999) 97–127.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

APPROPRIATE COMBINATIONS OF CONTROLLER PARAMETERS FOR UNSTEADY FLOW SIMULATIONS WITH ADAPTIVE TIME STEP CONTROL

Kathrin Kožulović, Graham Ashcroft

Institute of Propulsion Technology, German Aerospace Center (DLR)
Linder Höhe, 51147 Cologne, Germany, http://www.dlr.de/at/
e-mail: kathrin.kozulovic@dlr.de, graham.ashcroft@dlr.de

Keywords: Unsteady Flow, Time Accurate, Runge-Kutta, High Order, Adaptive Time Stepping, Controller Parameter, Error Estimation

Abstract. *Compressible Navier-Stokes equations, which govern unsteady flow phenomena, have been numerically integrated with several high order ESDIRK schemes (Explicit first stage, Stiffly accurate Diagonally Implicit Runge-Kutta). This type of numerical scheme allows for embedded error estimation and, based on this error estimation, for adaptive time stepping, cf. [1, 2, 3]. In this paper different controller types have been applied for the automatic adaption of the time step size to a prescribed accuracy level: I-, PI-, PC- and PID-controller. Depending on the choice of corresponding controller parameters, the simulation time can be varied considerably. On the one hand, a conservative choice leads to slow time step adaption, small step size and long simulation time. On the other hand, an aggressive parameter setting can lead to quick step size adaption, but also to rejected computations, due to large time step size, which does not keep the prescribed accuracy. This also may result in long simulation time. Obviously, there must be appropriate combinations of controller parameters, which minimize the simulation duration. Hence, a systematic study on a large controller parameter space has been conducted and regions with appropriate combinations have been identified. Both generic and complex test cases have been investigated. Overall CPU times and number of time steps are compared among different ESDIRK schemes and controllers. For the generic test case, most schemes show a narrow space for the choice of suitable controller parameters, but only one scheme shows good simulation performance without large sensitivities to parameter combinations. For the complex test cases, all schemes and controllers are less sensitive to parameter combinations. Furthermore, the optimum combination varies strongly with the type of governing equations.*

1 INTRODUCTION

The time integration of unsteady flow simulations can be conducted by different schemes: explicit or implicit, multi-stage or multi-step. In the present work, ESDIRK schemes have been applied, due to their well-balanced combination of favourable features as stability, robustness, embedded error estimation and time step control, cf. [1], [2]. In particular, the ability to adapt the time step size automatically to some prescribed error tolerance is of large benefit for the simulation reliability and efficient use of computational resources. Previous work was focused on implementation and validation of different ESDIRK schemes of 3rd- and 4th-order of accuracy, together with a basic time step control, cf. [4], [5].

Now, the investigation of time step control has been considerably enhanced and the schemes are re-assessed in terms of their computational efficiency. Different time step controllers (I, PI, PID, PC) have been implemented and their parameters have been systematically varied among a very large design space, in order to identify the optimum settings. Two essential mechanisms affect the simulation time. On the one hand, a quick adaption of time step size leads to reduced run time. On the other hand, the simulation time is considerably prolonged if many iterations with rejected error estimates are conducted. The reason therefore is an excessively large time step, usually generated by an overshoot in the automatic time step size controller. In other words, the controller parameters should provide a quick time step size adaption, but prevent the overshoots. As will be shown later, both the type of equation system and of test case require a different controller type with a different combination of corresponding parameters, in order to achieve quick and reliable simulations.

Besides time step controllers, the accuracy order of the applied scheme (for time integration and error estimation) significantly affects the simulation duration. On the one hand, a higher scheme order can lead to larger time steps, hence reducing the runtime. On the other hand, the computational effort per iteration is larger for higher order schemes, which will increase the runtime. These effects, in combination with the aforementioned controller parameters, are the primary influence factors for the simulation duration and will be addressed in the subsequent sections.

2 FLOW SOLVER

The numerical simulations presented in this work are performed using the compressible Navier-Stokes solver TRACE [6], [7]. The TRACE code has been developed at DLR's Institute of Propulsion Technology in Cologne to model and investigate turbomachinery flows. The governing equations are solved using a multi-block approach for both structured and unstructured grids embedded in a parallel environment. The finite-volume method is adopted in generalized coordinates about the cell centers for space discretization. Upwind-biased spatial differencing in conjunction with Roe's flux-difference-splitting method is used to evaluate the inviscid fluxes, with limiters used to obtain smooth solutions in the vicinity of shocks. Viscous terms are discretized using second-order accurate central differences. Turbulence modeling is effected by a $k-\omega$ two-equation approach with turbomachinery-specific extensions. The discretization of the spatial operators in the Navier-Stokes equations results in a system of ordinary differential equations of the following form

$$\frac{d\mathbf{U}}{dt} = \mathbf{R}(t, \mathbf{U}(t)) \tag{1}$$

where $\mathbf{U}$ represents the vector of conservative variables and t denotes the physical time.

3 IMPLICIT RUNGE-KUTTA METHODS

The numerical solution $\mathbf{U}^{n+1}$ of the ordinary differential system 1 at time $t^{n+1} = t^n + \Delta t$ using an s-stage implicit Runge-Kutta scheme is

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t \sum_{j=0}^{s} b_j \mathbf{R}(t_n + c_j \Delta t, \mathbf{W}^i) \tag{2}$$

with the values of intermediate stages i

$$\mathbf{W}^i = \mathbf{U}^n + \Delta t \sum_{j=0}^{s} a_{ij} \mathbf{R}(t_n + c_j \Delta t, \mathbf{U}^j) \,, \quad i = 1, \ldots, s. \tag{3}$$

The values a_{ij}, b_j and c_j are called Butcher coefficients of the scheme, where the values c_j denote the evaluation time of the intermediate stages in the interval $t_n \rightarrow t_n + \Delta t$ and b_j the contribution of the individual stage to the final solution at the end of the Runge-Kutta step.

3.1 ERROR ESTIMATION

Implicit Runge-Kutta methods allow for error estimation using an embedded method in the same Butcher tableau with different order of accuracy $q = p \pm 1$, with q the order of the scheme advancing the solution and p the order of the error estimation scheme, respectively. The solution $\hat{\mathbf{U}}^{n+1}$ of the embedded scheme is

$$\hat{\mathbf{U}}^{n+1} = \mathbf{U}^n + \Delta t \sum_{j=0}^{s} \hat{b}_j \mathbf{R}(t_n + c_j \Delta t, \mathbf{W}^i), \tag{4}$$

where the values $\hat{b}_j$ are an additional set of solution coefficients appended to the Butcher tableau. The embedded method depends only on linear combinations of already present stage values $\mathbf{W}^i$ which makes the error estimation computationally cheap. The error estimate of embedded IRK methods is defined as

$$\Delta \mathbf{U}^{n+1} = \mathbf{U}^{n+1} - \hat{\mathbf{U}}^{n+1}. \tag{5}$$

Within this investigation the maximum error $||\Delta \mathbf{U}^{n+1}||_\infty = max(|\Delta \mathbf{U}_i^{n+1}|)$ is used.

3.2 ESDIRK SCHEMES

The solution of an n-dimensional differential equation system using an implicit Runge-Kutta method with s stages needs the evaluation of $n \cdot s$ non-linear equation sets per time step. Therefore several simplifications are useful for the successful implementation of these methods.

For diagonally implicit schemes the coefficient matrix a_{ij} is characterized by a lower triangular form, each stage of the scheme therefore depends only on previously solved stages and the system may thus be solved in s successive steps. An additional simplification for this type of scheme is the use of the same diagonal coefficient γ for each stage. For methods with explicit first stage ($a_{1j} = 0$) the stages are at least second-order accurate. The conditions for stiffly accurate methods are $a_{ij} = b_j, j = 1, \ldots, s$, which ensures that the solution of the last stage $\mathbf{U}^s$ is the solution at the next time step $\mathbf{U}^{n+1}$. A four stage example of the Butcher tableau for an Explicit first stage, Singly Diagonal Implicit Runge-Kutta (ESDIRK) scheme is given in Table 1.

	$c_1 = 0$	0	0	0	0
	c_2	a_{21}	γ	0	0
	c_3	a_{31}	a_{32}	γ	0
	$c_4 = 1$	b_1	b_2	b_3	γ
order of accuracy	q	b_1	b_2	b_3	γ
order of accuracy	$p \pm 1$	$\hat{b_1}$	$\hat{b_2}$	$\hat{b_3}$	$\hat{b_4}$

Table 1: Generic Butcher tableau of a four-stage embedded ESDIRK scheme

In this paper stiffly accurate embedded ESDIRK pairs from third-order up to fourth-order proposed by Alexander [1], Carpenter et. al. [2] and Kvaerno [3] are considered for the numerical solution of the compressible Navier-Stokes equations. The schemes are here denoted by authors name followed by the order of accuracy of the solution scheme and in brackets the order of the scheme for error prediction, e.g. Ale 3(2). The formula pairs presented by Kvaerno differ a little from the generic form given in Table 1, as the last two successive stages are evaluated at $c_s = c_{s-1} = 1$ and used either to advance the solution or to estimate the error. In that case the additional coefficients $\hat{b}$ are omitted. The Butcher tableaus of the methods used in the present work are given in Table 2- Table 4.

4 STEP SIZE CONTROL ALGORITHMS

The local error of the numerical solution is observed by the embedded ESDIRK scheme. To control the accuracy of the solution by holding the error within the desired tolerance bounds and to adapt the time step size Δt standard controller types are used. For embedded formula pairs with $q = p \pm 1$ time step control can be conducted using I-, PI-, PID- or PC-controllers [2], [8]. The algorithms to select the time step size for the different controllers are given by

$$\Delta t_I^{n+1} = dt_{fac}\, \Delta t^n \left(\frac{TOL}{||\Delta U^{n+1}||_\infty} \right)^{K_I}, \tag{6}$$

$$\Delta t_{PI}^{n+1} = dt_{fac}\, \Delta t^n \left(\frac{TOL}{||\Delta U^{n+1}||_\infty} \right)^{K_I} \left(\frac{||\Delta U^n||_\infty}{TOL} \right)^{K_P}, \tag{7}$$

$$\Delta t_{PID}^{n+1} = dt_{fac}\, \Delta t^n \left(\frac{TOL}{||\Delta U^{n+1}||_\infty} \right)^{K_I} \left(\frac{||\Delta U^n||_\infty}{TOL} \right)^{K_P} \left(\frac{TOL}{||\Delta U^{n-1}||_\infty} \right)^{K_D}, \tag{8}$$

where $dt_{fac} < 1$ is a safety factor which prevents rejected iterations.

Additionally the (so-called) predictive controller is considered in this investigation, which was designed by Gustafsson [8] for stiff equation systems. Stiff equations originate, among others, from separated space and time discretization, which is typical for the discretization of the Navier-Stokes equations. The predictive control algorithm

$$\Delta t_{PC}^{n+1} = dt_{fac}\, \Delta t^n \left(\frac{TOL}{||\Delta U^{n+1}||_\infty} \right)^{K_I} \left(\frac{||\Delta U^n||_\infty}{TOL} \right)^{K_D} \left[\frac{\Delta t^n}{\Delta t^{n-1}} \right]. \tag{9}$$

The choice of controller parameters K_I, K_P, K_D in the literature depends on the order of accuracy of the time integration scheme q and the order of accuracy of the embedded method p. Former investigations regarding ESDIRK schemes and their error estimation skills showed that for different methods the parameter choice depends not only on the order but also on the scheme itself (and test problem). Therefore the focus of this paper lies on the controller coefficients.

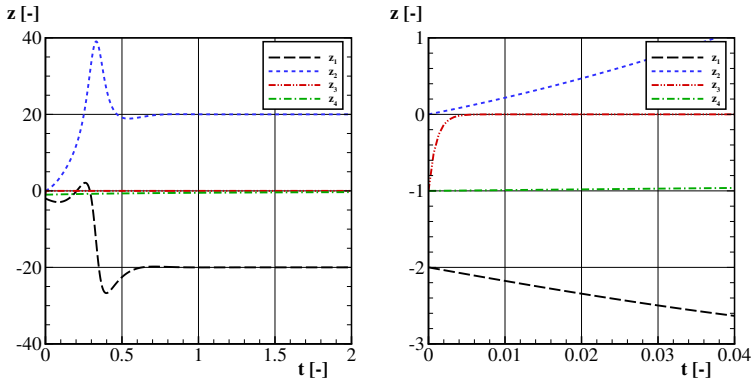


Figure 1: Analytical solution of E4 problem

Inappropriate parameter settings can lead to several controller failures: too aggressive time step adaption, alternating time step sizes, very small step sizes, decreasing step sizes with decreasing error or increasing step size with increasing error.

5 APPLICATION

5.1 Non-linear stiff ordinary differential equation system

For the basic validation of the ESDIRK error estimation the analytical test problem E4 given by Enright [9] was used, cf. [5]. The analytical solution for this system of ordinary differential equations allows an exact comparison with the numerical solution and enables a comparison of estimated and exact errors. Being non-linear and stiff, this test problem is representative of the discretized Navier-Stokes equations. The problem is given as follows

$$y = Uz, \;\; y(0) = (0, -2, -1, -1)^T \;\text{ with}$$

$$\begin{aligned} z_1' &= -(\beta_1 z_1 - \beta_2 z_2) + \tfrac{1}{2}(z_1^2 - z_2^2) \\ z_2' &= -(\beta_2 z_1 - \beta_1 z_2) + z_1 z_2 \\ z_3' &= -\beta_3 z_3 + z_3^2 \\ z_4' &= -\beta_4 z_4 + z_4^2 \end{aligned} \quad , \quad U = \tfrac{1}{2}\begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}, \quad \beta = \begin{bmatrix} 10 \\ -10 \\ 1000 \\ 0.001 \end{bmatrix}.$$

Here this test problem is used to investigate the time step control algorithms. The solution in the time interval $t = 0 - 0.2$ is shown in Figure 1 with a detailed view on the stiff starting region of component z_3. This time interval is considered throughout the whole study.

Error Estimation: A comparison of the analytical Δz_N and estimated error Δz_S obtained using $N = 100$ iterations to resolve the time period is shown in Figure 2. In order to obtain the error of a single Runge-Kutta step, each new step is initialized with the analytical solution. In general the various methods all provide a good estimate of the errors. In particular, however, the

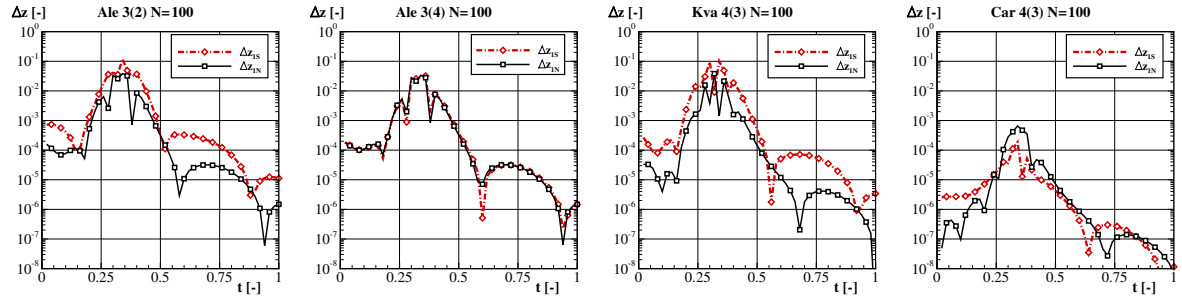


Figure 2: Error estimation Δz_S and analytical error Δz_N of component z_1 for different schemes

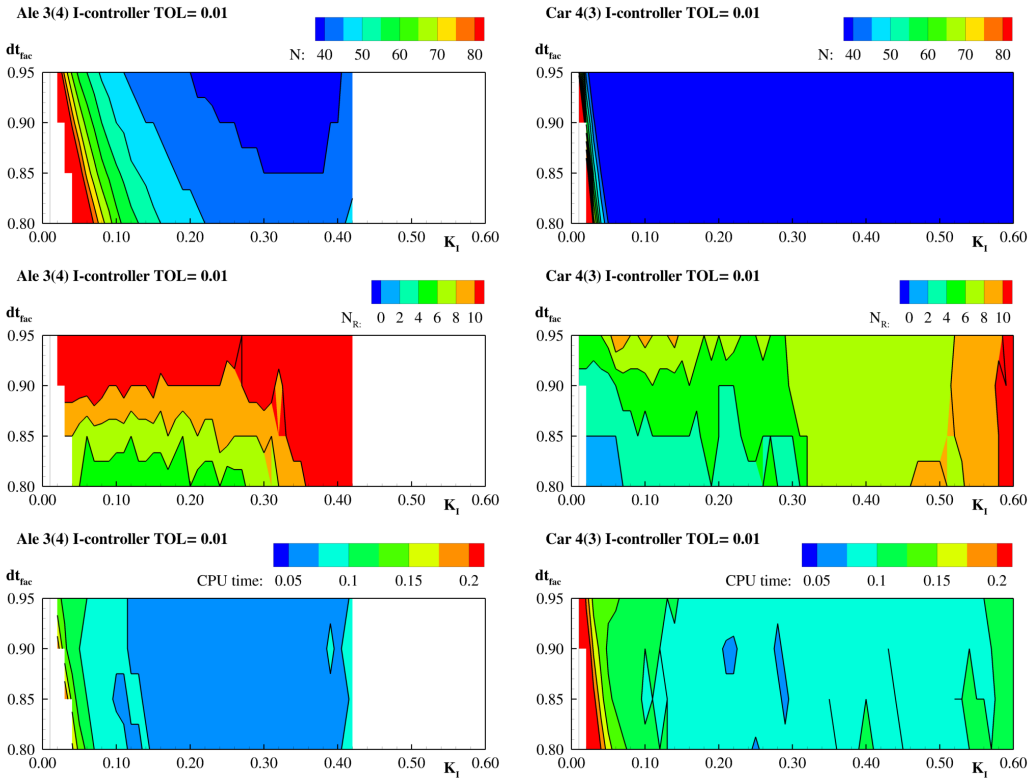


Figure 3: I-controller using Ale 3(4) (left) and Car 4(3) (right)

estimation scheme of Ale 3(4) produces the most accurate error trend and error level. In contrast Car 4(3) tends to underpredict the error. However, the error level is two orders of magnitude smaller due to the higher order of accuracy of the solution scheme.

To analyze the different controller types a systematic investigation of safety factor Δt_{fac} and parameters K_I, K_P, K_D is conducted. The influence of the safety factor was first tested using the I-Control and to reduce the effort for the following tests in consequence fixed to $dt_{fac} = 0.90$. Areas with any kind of controller failure are blanked in the figures concerning control parameter settings. Remember that this means not necessarily a total failure of the controller, also inappropriate parameters are blanked, e.g. producing long run times due to small step sizes.

I-controller: The resulting number of accepted time steps N for the integral element are given in Figure 3 for Ale 3(4) and Car 4(3), the number of rejected time steps N_R and the CPU time are also shown. Four values in the range of 0.80-0.95 are used to test the impact of dt_{fac}. The sampling of K_I in this test is much finer resolved using 60 points between $0.01 - 0.60$. Preliminary investigations showed that for most ESDIRK schemes the maximum allowable K_I is within this range. Here Ale 3(4) and Car 4(3) are shown exemplarily, which are the time integration methods with the highest sensitivity to control parameter choice, respectively lowest sensitivity. Car 4(3) is less sensitive to the parameter choice and performs over a broad parameter range very well. For all other methods the parameter range is limited. With increasing dt_{fac} the number of rejected steps slightly increases for Ale 3(4). On the other hand, together with the number of accepted steps, nearly the same number of overall iterations results. Δt_{fac} prevents for rejected time steps but is not the key factor to reduce overall number of time steps and therefore runtime.

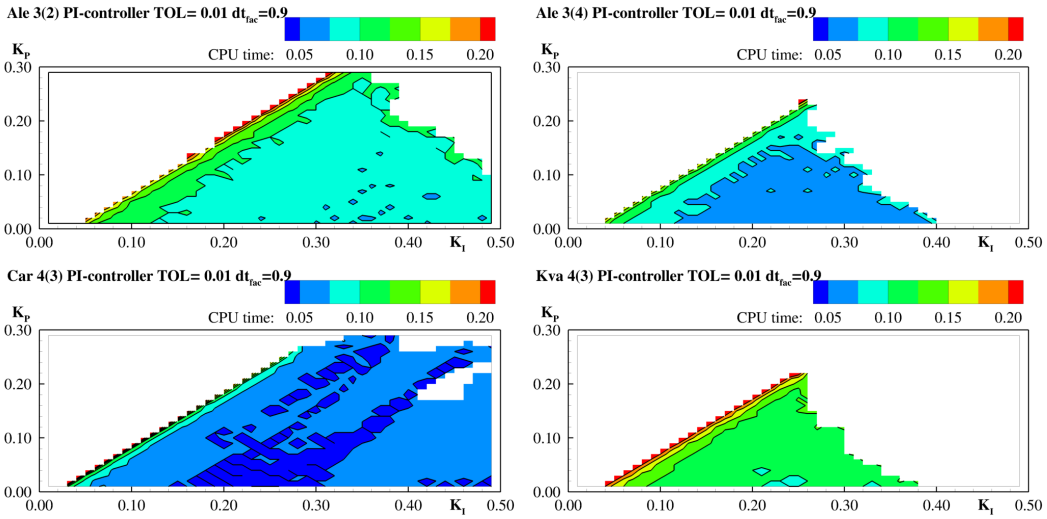


Figure 4: CPU time for PI-Control

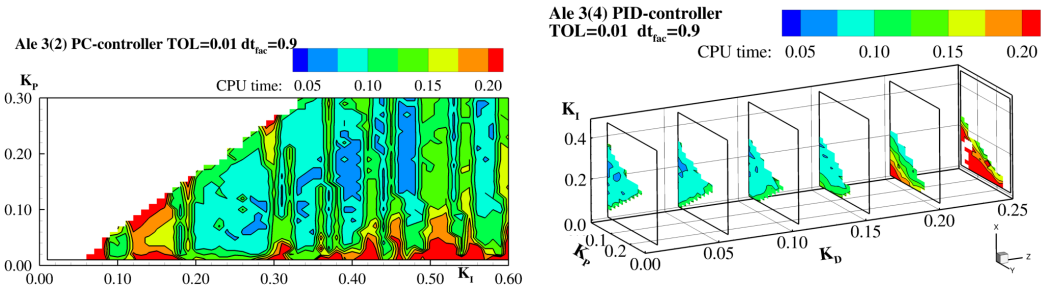


Figure 5: CPU time for Ale 3(2) using PC-Control (left) and Ale 3(4) using PID-Control (right)

PI-controller: Figure 4 compares the runtime for different ESDIRK schemes using Algorithm 7 for step size adaption. Here a fine sampling mesh with 29 points for K_I and 49 points for K_P is applied. The influence of the order of accuracy is clear lower order pairs such at Ale 3(2) tend to longer run times than methods with higher order e.g. Ale 3(4) or Car 4(3). Kva 4(3) contradicts this statement and is not well suited to this problem. Using this time step control algorithm the parameter range for Car 4(3) is nearly as limited as for the other integration schemes using this controller type. However, comparing I- and PI-Controllers it can be seen that the more sophisticated PI-Controller is able to produce lower runtimes, e.g. Car 4(3). This superiority is not that obvious for other methods.

PC-controller: Using the example of Ale 3(2) in Figure 5 (left) the time consumption for the predictive control algorithm (Equation 9) is shown. The results are scattered strongly within parameter space, small parameter changes cause large changes in runtime behaviour. Here this performance is only shown for one ESDIRK method, but it is a general result for all methods under consideration using this controller type.

PID-controller: The control parameter study for a controller of PID-type was performed using a sampling of 24 values for K_I and 14 values for K_P and K_D. In Figure 5 (right) results for Ale 3(4) are shown. With increasing values of K_D larger areas with increasing CPU time occur. An improvement of CPU time in comparison to PI-Control is not observed, however the risk of inappropriate parameter settings increases.

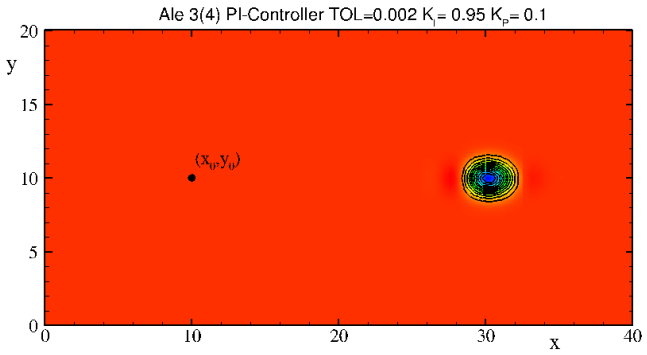


Figure 6: Computational domain and density contours at the end position

5.2 Advection of entropy disturbance

To investigate the implementation of adaptive time stepping in the context of the Navier-Stokes flow solver first a two-dimensional test problem is considered: the advection of an entropy disturbance. The simplicity of this problem facilitates a systematic controller parameter study in the Navier-Stokes solver. This is realized by initializing the density field as follows

$$\rho = \left[T_\infty - 0.001 e^{(1-r^2)}\right]^{\frac{1}{\gamma-1}}$$

with $r = \sqrt{(x-x_0)^2 + (y-y_0)^2}$ and $\gamma = 1.4$ (ratio of specific heats). The freestream values of axial velocity, pressure and temperature are $U_\infty = s135.6$ m/s, $P_\infty = 90500$ Pa and $T_\infty = 278.977$ K, respectively. The flow is uniform with $u = U_\infty$, $v = w = 0$ and $p = P_\infty$. The two-dimensional computational domain is a simple rectangle meshed with 12800 cells, split over 8 blocks, extended $40l$ in x-direction and $20l$ in y-direction, cf. Figure 6. The simulation starts at $t = 0$ with the entropy disturbance centered at $(x_0, y_0) = (10l, 10l)$ and ends about $(x_T, y_T) = (30l, 30l)$.

To analyze automatic time step adjustment and control mechanisms in the context of the flow solver a similar parameter study to that of the E4 problem with less sampling points is conducted. The prescribed tolerance of $TOL = 0.002$ is very stringent. Preliminary investigations, in which the tolerance targets were decreased from $0.01 - 0.001$, showed that this value is a threshold, which leads to strongly increased runtimes. Again the safety factor dt_{fac} is only considered for the I-controller and for all other controllers is fixed to $dt_{fac} = 0.96$.

I-controller: The resulting CPU time for all ESDIRK methods using the integral element are given in Figure 7. The number of rejected time steps N_R varies between 2 and 8, where two steps are needed to meet the prescribed tolerance in the first time step. Four values for $dt_{fac} = 0.87 - 0.96$ are studied and 10 values of $K_I = 0.05 - 0.95$. The valid parameter range is similar for all investigated time integration schemes, as well the areas with lowest and highest time consumption. A low value of dt_{fac} results in increasing CPU time, especially using Car 4(3) for time integration. In contrast to the results of E4 here the highest value of K_I yields the fastest runtime. Compared to the third-order accurate schemes, Car 4(3) does not reflect the superiority as shown for E4 problem. This scheme has six-stages while the other schemes have four or five stages. It becomes apparent that the number of stages influences the computational overhead of the flow solver. For this simple flow problem the higher order accuracy is not crucial for time consumption. In particular, with increasing number of rejected time steps (small dt_{fac}) the runtime increases dramatically.

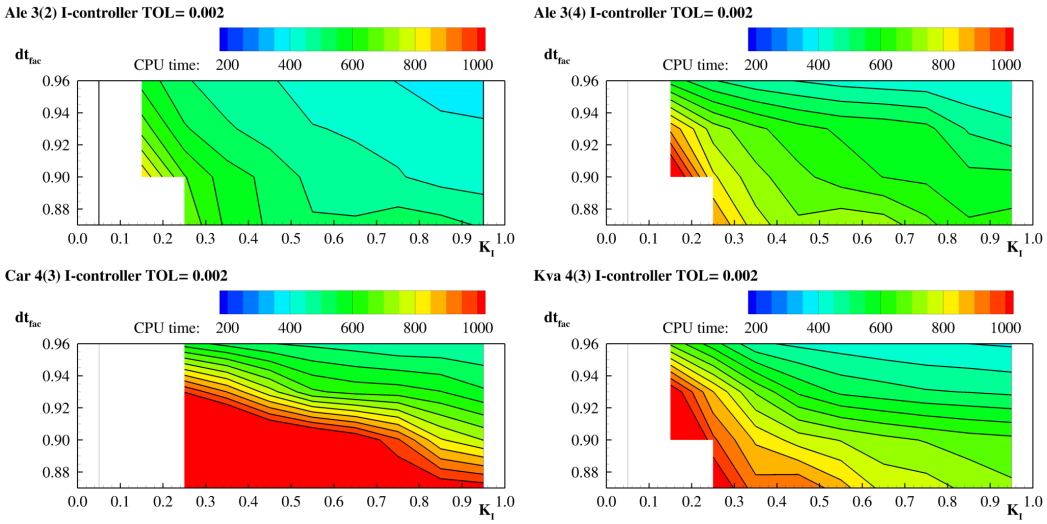


Figure 7: CPU time for advection problem using I-controller

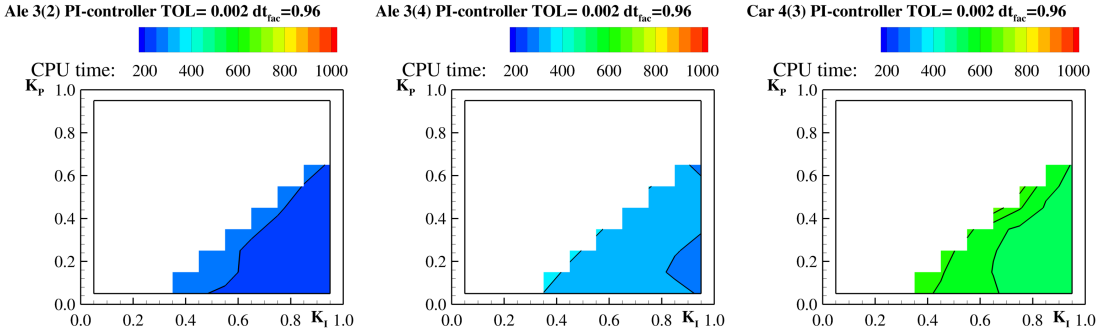


Figure 8: CPU time for advection problem using PI-Control

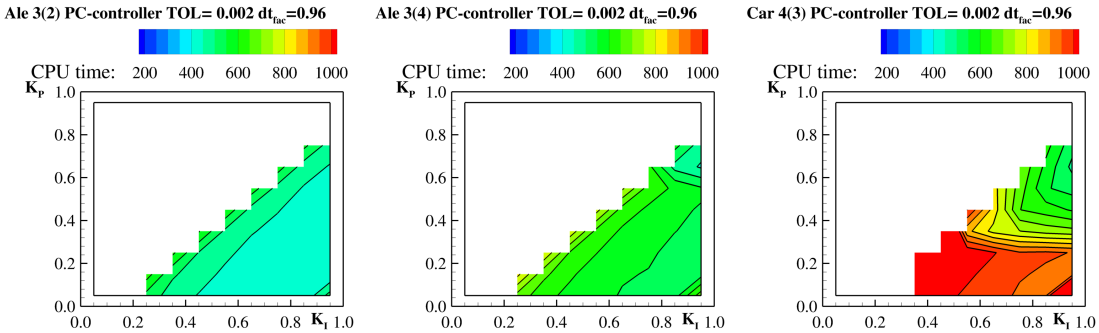


Figure 9: CPU time for advection problem using PC-Control

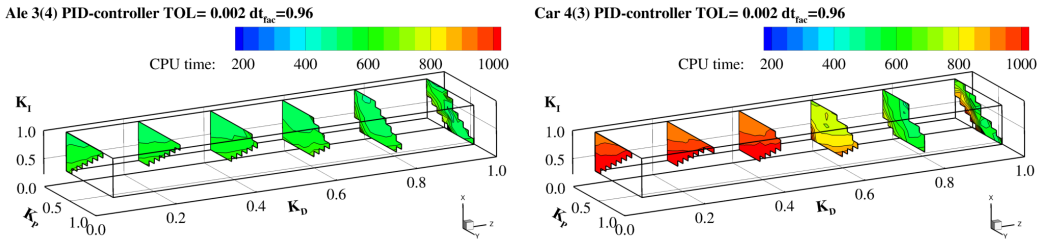


Figure 10: CPU time for advection problem using PID-Control

PI-controller: The proportional integral element is tested using 10 values for K_I and K_P in the range of $0.05 - 0.95$. The valid parameter range for PI-control is similar for all ESDIRK schemes, Figure 8. Results for Kva 4(3) are similar to Car 4(3) (not shown here). The proportional element, in addition to the integral element, smoothes the resulting runtimes over the whole parameter range. The fastest simulations are performed using high values of K_I and low values of K_P, e.g. Ale 3(4).

PC-controller: In the simulations performed using control algorithm 9 the same parameter range as for the PI type is considered, cf. Figure 9. Here longer run times are obtained. In particular, Car 4(3) shows significant higher simulation times in areas with low values of K_P. For this scheme a wide spread of the results with small parameter changes is observed, which makes it difficult to select a proper parameter combination. Kva 4(3) leads to similar CPU time to Ale 3(4) (not shown here). Although this controller type was specifically designed for stiff equation systems it does not seem to provide better results than the standard PI-controller in our investigations.

PID-controller: In Figure 10 resulting CPU times using Algorithm 8 for adaptive time stepping are shown. The parameter K_I ranges from $0.25 - 0.95$ using 8 samples and 10 samples for K_P, K_D in the range of $0.05 - 0.95$. With increasing K_D valid parameter combinations decrease. A remarkable feature of Ale 3(4) method is an almost constant CPU time for all parameter combinations. Furthermore, the region with $K_D = 0.6$ seems to be the preferred choice, due to largest area of valid combinations. In contrast, the Car 4(3) method shows a considerable decrease of simulation time with increasing K_D. Here, the region with $K_D = 0.8$ yields comparable runtime results as Ale 3(4), together with a larger region of valid combinations as for higher or lower values.

5.3 Flow around a circular cylinder

The basic investigations of controller coefficients and the implementation of embedded ESDIRK schemes in conjunction with adaptive time step control in the context of the Navier-Stokes solver were demonstrated using analytical problems. Now, as a final test case, a more realistic final problem is considered: the flow around a circular cylinder. As a further increase of complexity, the viscous effects are now captured by the two-equation k-ω turbulence modell of Wilcox [10]. The Reynolds number is 140.000, based on the cylinder diameter of 0.010137m. The free stream values for velocity and temperature are 21.20 m/s and $24°$C. The vortex shedding period is 0.02665 s, which corresponds to a Strouhal number of 0.179.

In contrast to preliminary test cases the computational effort for one simulation of the cylinder flow is much higher, therefore the number of control algorithms as well as the number of coefficients is reduced. Based on the results for the analytical problems, E4 and the advection of an entropy disturbance, here controllers of PI-type are used. For this flow problem a high value of $\Delta t_{fac} = 0.96$ is suitable.

As shown in the previous sections, the entropy advection problem, which is governed by Navier-Stokes equations, yields a region of appropriate controller parameters which is quite different than the corresponding region of the generic E4 test case, which is governed by an other differential equation system. Hence, the type of ODE system seems to have a major influence on the controller parameters. To test this assumption, another Navier-Stokes test case is investigated in this section. In other words, it is assumed that the same type of ODE system leads to the same range of suitable controller parameters. Due to the similar range of suitable K_I and K_P parameters for all methods in the previous Navier-Stokes test case, only Ale 3(4) method is investigated in this test case, and the corresponding results are considered

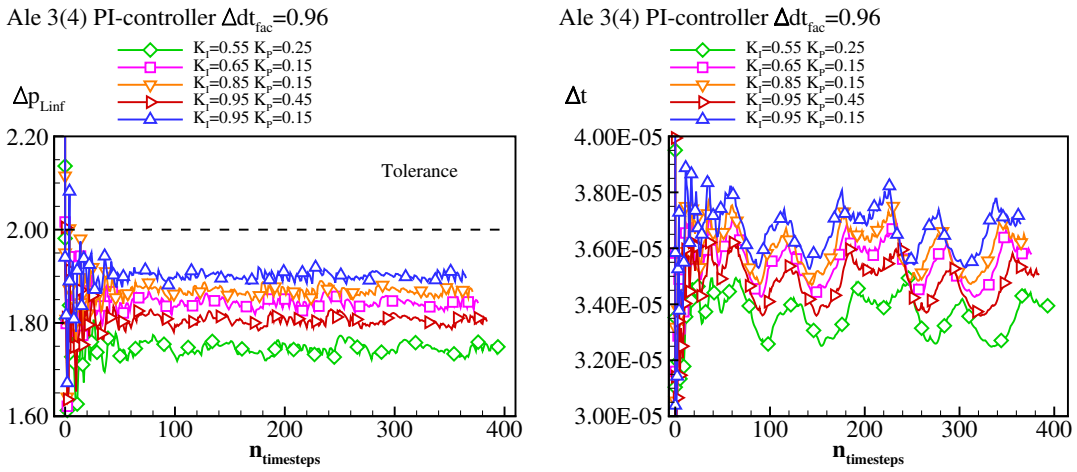


Figure 11: Cylinder flow using PI-Controller error of pressure Δp_{Linf}(right) and time step size Δt for Ale 3(4)

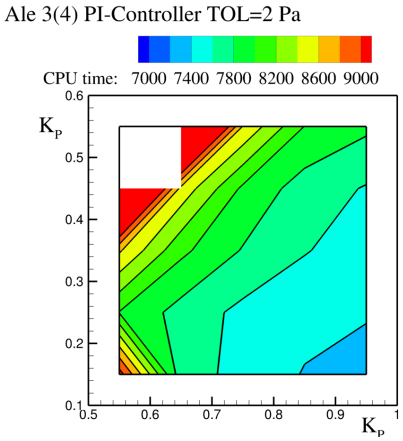


Figure 12: Cylinder flow using PI-controller for Ale 3(4)

as representative for all methods. For this problem a tolerance for the maximum pressure error of $\Delta p_{Linf} = 2\ Pa$ is prescribed.

Five simulations with different controller coefficients using Ale 3(2) are presented in Figure 11. As a beneficial result, it is found that all parameter combinations provide estimated errors, which are lower than the prescribed value. However, the simulation with small K_I and large K_P value provides a significantly smaller estimated error, making the simulation too conservative. This in turn leads to a large number of timesteps and a long simulation duration. In contrast, the combinations with large K_I and small K_P values run much closer to the prescribed tolerance, reducing the number of time steps by ca. $5\ \%$. This observation is in agreement with the time step size Δt results at the right hand side of the figure, where the largest values are found for the latter parameter combination.

Figure 12 confirms the assumption that the suitable parameters strongly depend on the governing equations. The region of valid parameter combinations, which simultaneously yields the shortest simulation duration, coincides very well with the results from the advection problem.

6 CONCLUSIONS

In this work different control algorithms for time step size control in unsteady flow simulations based on the compressible Navier-Stokes equations using embedded Runge-Kutta schemes have been investigated. Here a detailed analysis of parameter settings for time step size adaption using I-, PI-, PID- and PC-Control has been performed with the aim of calibrating the controllers for unsteady flow simulations.

Prior to the implementation in the Navier-Stokes solver a stiff analytical problem is considered to verify the precision of error prediction and for the basic understanding of different control algorithms. This test case is used for a detailed study of controller types and parameter range. In general the quality of error estimation for different ESDIRK schemes is very good. For example, the embedded ESDIRK method using the highest order scheme for error prediction, Ale 3(4), yields excellent results. However, in some cases, the estimated error is smaller than the analytical error by almost one order of magnitude.

The implementation of adaptive time step control in the flow solver is examined first with a simple advection problem, which allows for large controller parameter studies. As second test case the flow around a cylinder is used, at which only a selected number of controller coefficients is considered.

Even with the simplest controller type which contains only an integral element adaptive time stepping can be performed. The more sophisticated PC-Controller recommended by Gustafsson for stiff equations shows conflicting results, ranging from well suited for a simple advective flow problem to very poor performance for academic test problem. The PID-controller may provide region of well behaved combinations, but the resulting CPU time is much longer than for concurrent controllers. Overall, the PI-controller outperforms all other controllers for all investigated test cases. Finally, the region of suitable parameter combination depends strongly on the governing equations, which are to be solved. For the Navier-Stokes solver a high value of K_I and a low value of K_D are favourable.

REFERENCES

[1] R. Alexander. Design and implementation of DIRK integrators for stiff systems. *Applied Numerical Mathematics*, 46(1):1–17, 2003.

[2] C. A. Kennedy and M. H. Carpenter. Additive Runge-Kutta Schemes for Convection-Diffusion-Reaction Equations. *Applied Numerical Mathematics*, 44(1-2):139–181, 2003.

[3] A. Kvaerno. Singly Diagonally Implicit RungeKutta Methods with an Explicit First Stage. *BIT Numerical Mathematics*, 44:489–502, 2004.

[4] G. Ashcroft, C. Frey, K. Heitkamp, and C. Weckmüller. Advanced Numerical Methods for the Prediction of Tonal Noise in Turbomachinery - Part I: Implicit Runge-Kutta Schemes. *J. Turbomach.*, 136(2):021003–021003, 2013.

[5] K. Kozulovic and G. Ashcroft. Temporal Error Estimation and Adaptive Time Step Control in Unsteady Flow Simulations. In *WCCM XI, ECCM V, ECFD VI, July 20 - 25, 2014, Barcelona, Spain*, 2014.

[6] D. Nürnberger, F. Eulitz, and A. Zachcial. Recent Progress in the Numerical Simulation of Unsteady Viscous Multistage Turbomachinery Flow. In *15th International Symposium on Air Breathing Engines, Bangalore, India, Paper 2001-1082*, 2001.

[7] E. Kügeler. Numerical Method for the accurate analysis of cooling efficiency in film-cooled turbine-blades. Technical report, Institute of Propulsion Technology, German Aerospace Center, published in German, 2005. DLR Forschungsbericht 2005-11, Köln, 2005.

[8] K. Gustafsson. Control-theoretic techniques for stepsize selection in implicit runge-kutta methods. *ACM Trans. Math. Softw.*, 20(4):496–517, December 1994.

[9] W. H. Enright, T. E. Hull, and B. Lindberg. Comparing numerical methods for stiff systems of ODEs. *BIT Numerical Mathematics*, 15(1):10–48, 1975.

[10] David C. Wilcox. Multiscale model for turbulent flows. *AIAA Journal*, 26:1311–1320, 1988.

A BUTCHER TABLEAUS

$\gamma = 0.435866521508$

$b_1 = \frac{-18\gamma c_3+12\gamma^2 c_3+3c_3+12\gamma-12\gamma^2-2}{12\gamma c_3}$
$b_2 = \frac{2-3c_3+6\gamma c_3-6\gamma}{12\gamma(2\gamma-c_3)}$
$b_3 = \frac{6\gamma^2+1-6\gamma}{3c_3(c_3-2\gamma)}$

0	0			
2γ	γ	γ		
c_3	$\frac{6\gamma c_3-c_3^2-4\gamma^2}{4\gamma}$	$\frac{-c_3(c_3-2\gamma)}{4\gamma}$	γ	
1	b_1	b_2	b_3	γ
	b_1	b_2	b_3	γ
	$\hat{b_1}$	$\hat{b_2}$	$\hat{b_3}$	$\hat{b_4}$

Ale 3(2): $c_3 = \frac{1}{2} + \frac{\gamma}{4}$

$\hat{b_1} = b_1 - \delta b_2 - 0.096 + 0.284$
$\hat{b_2} = b_2 + \frac{-c_3 0.096+0.284}{2\gamma}$
$\hat{b_3} = b_3 + 0.096$
$\hat{b_4} = \gamma - 0.284$

Ale 3(4): $c_3 = \frac{18}{13}\gamma^2 - 2\gamma + \frac{14}{13}$

$\hat{b_1} = \frac{12\gamma c_3-4\gamma-2c_3+1}{24\gamma c_3}$
$\hat{b_2} = \frac{2c_3-1}{24\gamma(2\gamma-c_3)(2\gamma-1)}$
$\hat{b_3} = \frac{1-4\gamma}{12c_3(2\gamma-c_3)(c_3-1)}$
$\hat{b_4} = \frac{3+12\gamma c_3-4c_3-8\gamma}{12(2\gamma-1)(c_3-1)}$

Table 2: Ale 3(2) and Ale 3(4): Four-stage, 3rd-order ESDIRK method with embedded 2nd- or 4th-order scheme

$\gamma = 0.5728160625$

$a_{31} = \frac{144\gamma^5-180\gamma^4+81\gamma^3-15\gamma^2+\gamma}{(12\gamma^2-6\gamma+1)^2}$
$a_{32} = \frac{-36\gamma^4+39\gamma^3-15\gamma^2+2\gamma}{(12\gamma^2-6\gamma+1)^2)}$

0	0				
2γ	γ	γ			
$a_{31}+a_{32}+\gamma$	a_{31}	a_{32}	γ		
1	$\hat{b_1}$	$\hat{b_2}$	$\hat{b_3}$	γ	
1	b_1	b_2	b_3	b_4	γ

$\hat{b_1} = \frac{-144\gamma^5+396\gamma^4-330\gamma^3+117\gamma^2-18\gamma+1}{12\gamma^2(12\gamma^2-9\gamma+2)}$
$\hat{b_2} = \frac{72\gamma^4-126\gamma^3+69\gamma^2-15\gamma+1}{12\gamma^2(3\gamma-1)}$
$\hat{b_3} = \frac{(-6\gamma^2+6\gamma-1)(12\gamma^2-6\gamma+1)^2}{12\gamma^2(12\gamma^2-9\gamma+2)(3\gamma-1)}$

$b_1 = \frac{(288\gamma^4-312\gamma^3+120\gamma^2-18\gamma+1}{48\gamma^2(12\gamma^2-9\gamma+2)}$
$b_2 = \frac{24\gamma^2-12\gamma+1}{48\gamma^2(3\gamma-1)}$
$b_3 = \frac{-(12\gamma^2-6\gamma+1)^3}{48\gamma^2(3\gamma-1)(12\gamma^2-9\gamma+2)(6\gamma^2-6\gamma+1)}$
$b_4 = \frac{-24\gamma^3+36\gamma^2-12\gamma+1}{24\gamma^2-24\gamma+4}$

Table 3: Kva 4(3): Five-stage, 4th-order ESDIRK method with embedded 3rd-order scheme

0	0					
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$				
$\frac{83}{250}$	$\frac{8611}{62500}$	$-\frac{1743}{31250}$	$\frac{1}{4}$			
$\frac{31}{50}$	$\frac{5012029}{34652500}$	$-\frac{654441}{2922500}$	$\frac{174375}{388108}$	$\frac{1}{4}$		
$\frac{17}{20}$	$\frac{15267082809}{155376265600}$	$-\frac{71443401}{120774400}$	$\frac{730878875}{902184768}$	$\frac{2285395}{8070912}$	$\frac{1}{4}$	
1	$\frac{82889}{524892}$	0	$\frac{15625}{83664}$	$\frac{69875}{102672}$	$-\frac{2260}{8211}$	$\frac{1}{4}$
	$\frac{82889}{524892}$	0	$\frac{15625}{83664}$	$\frac{69875}{102672}$	$-\frac{2260}{8211}$	$\frac{1}{4}$
	$\frac{4586570599}{29645900160}$	0	$\frac{178811875}{945068544}$	$\frac{814220225}{1159782912}$	$-\frac{-3700637}{11593932}$	$\frac{61727}{225920}$

Table 4: Ken 4(3): Six-stage, 4th-order ESDIRK method with embedded 3rd-order scheme

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ADAPTIVE BEM-BASED FEM ON POLYGONAL MESHES FROM VIRTUAL ELEMENT METHODS

S. Weißer

Saarland University
Campus building E1.1, 66041 Saarbrücken, Germany
weisser@num.uni-sb.de

Keywords: BEM-based FEM, polygonal meshes, residual based error estimate, adaptivity

Abstract. *Polygonal meshes are especially suited for the discretization of boundary value problems in adaptive mesh refinement strategies. Such meshes are very flexible and incorporate hanging nodes naturally. But only a few approaches are available that handle polygonal discretizations in this context. The BEM-based Finite Element Method (FEM) and a residual based error estimate are reviewed in the presentation. This a posteriori error estimate is reliable and efficient on polygonal meshes and can be applied in adaptive FEM strategies. Furthermore, the BEM-based FEM is applicable on such general meshes and gains its flexibility by implicitly defined trial functions. They are given as solutions of local Dirichlet problems related to the global differential operator. These local problems are treated by means of Boundary Element Methods (BEM) in the realization. In the numerical experiments the test problems of the recent publication on adaptive Virtual Element Methods by L. Beirão da Veiga and G. Manzini [ESAIM Math. Model. Numer. Anal., 49(2):577–599, 2015] are considered in an adaptive BEM-based FEM simulation. The experiments show optimal rates of convergence for uniform and adaptive mesh refinement, where the latter one yields, in particular, very local mesh adaptation.*

1 Introduction

Polygonal and polyhedral meshes have attracted a lot of interest in the discretization of boundary value problems during the last few years. Especially in adaptive mesh refinement strategies such meshes are attractive, since they are very flexible and handle hanging nodes naturally. New methods have been developed and conventional approaches were mathematically revised to handle these general meshes. The most prominent representatives for the new approaches are the Virtual Element Method (VEM) [2] and the Weak Galerkin Method [17]. Strategies like discontinuous Galerkin [7] and the mimetic discretization techniques [3] are also considered on polygonal and polyhedral meshes. Additionally, there are the so called BEM-based Finite Element Methods (FEM) which have their roots in the boundary element domain decomposition approaches [6]. These methods make use of implicitly defined trial functions that are locally the solution of boundary value problems treated in the numerical realization by means of Boundary Element Methods (BEM).

The BEM-based FEM has been studied for high order approximations [13, 20], mixed formulations [9], convection-adapted trial functions [10], time dependent problems [21] and on general polyhedral meshes [14]. Additionally, the BEM-based FEM has shown its flexibility and applicability on adaptively refined polygonal meshes [18, 22, 23]. Especially adaptivity is an interesting topic in this context, but, there are only a few references to adaptive strategies and a posteriori error control on polygonal meshes. A posteriori error estimates for the discontinuous Galerkin method are given in [11]. To the best of our knowledge there is only one publication for the Virtual Element Method [4] and one for the Weak Galerkin Method [5]. The first mentioned publication deals with a residual a posteriori error estimate for a C^1-conforming approximation space, and the second one is limited to simplicial meshes. For the mimetic discretization technique there are also only few references which are limited to low order methods, see the recent work [1].

The aim of this publication is to review the findings for the adaptive BEM-based FEM and to present computational results obtained by the considered approach for the test problems studied in [4] for the Virtual Element Method. Beside of the high flexibility of the polygonal meshes, the numerical experiments show optimal rates of convergence on very locally adapted meshes. This is possible since no additional refinement in the neighbourhood of marked elements is necessary to keep the meshes admissible.

In Section 2, we give a model problem and discuss the preliminaries for the BEM-based FEM. The discretization is reviewed in Section 3, and in Section 4, the reliability as well as the efficiency of the residual based error estimate on polygonal meshes is stated. Section 5 describes the adaptive BEM-based FEM strategy and presents several numerical experiments. Finally, we give a conclusion in Section 6.

2 Preliminaries

Let $\Omega \subset \mathbb{R}^2$ be a connected, bounded, polygonal domain with unit outer normal field n_Ω. Furthermore, denote by $\Gamma = \Gamma_D \cup \Gamma_N$ the boundary of Ω, where $\Gamma_D \cap \Gamma_N = \emptyset$ and $|\Gamma_D| > 0$. We consider the diffusion equation with mixed boundary conditions

$$-\mathrm{div}(a\nabla u) = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma_D, \qquad a\nabla u \cdot n_\Omega = g_N \quad \text{on } \Gamma_N, \tag{1}$$

where, for simplicity, we restrict ourselves to piecewise constant and scalar valued diffusion $a \in L_\infty(\Omega)$ with $0 < a_{\min} \leq a \leq a_{\max}$ almost everywhere in Ω. The usual notation for Sobolev spaces and their norms is used.

The well known variational formulation of problem (1) reads

$$\text{Find } u \in H^1_D(\Omega): \quad b(u,v) = (f,v)_{L_2(\Omega)} + (g_N, v)_{L_2(\Gamma_N)} \quad \forall v \in H^1_D(\Omega), \tag{2}$$

for $g_N \in L_2(\Gamma_N)$ and $f \in L_2(\Omega)$, where

$$b(u,v) = (a\nabla u, \nabla v)_{L_2(\Omega)}$$

is bounded and coercive on

$$H^1_D(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}.$$

We decompose the domain Ω into a family of meshes $\mathcal{K}_h$ containing non-overlapping, polygonal elements $K \in \mathcal{K}_h$ such that

$$\overline{\Omega} = \bigcup \{x \in \overline{K} \,\big|\, K \in \mathcal{K}_h\}.$$

The mesh is called regular if all elements $K \in \mathcal{K}_h$ are star-shaped with respect to a circle of radius ρ_K such that the ratio of the element diameter h_K and ρ_K is uniformly bounded, i.e. $h_K/\rho_K < \sigma_{\mathcal{K}}$, and if the element diameter can be uniformly bounded by a constant times the smallest length of its edges, i.e. $h_K < c_{\mathcal{K}} h_E$ for $E \in \mathcal{E}(K)$. Here, $\mathcal{E}(K)$ denotes the set of edges of K. Furthermore, we denote by $\mathcal{N}_h$ and $\mathcal{E}_h$ the sets of all nodes and edges in the mesh. $\mathcal{E}_h$ is decomposed into the sets of edges which are in the interior of Ω, on the Dirichlet boundary Γ_D and on the Neumann boundary Γ_N, we write $\mathcal{E}_h = \mathcal{E}_{h,\Omega} \cup \mathcal{E}_{h,D} \cup \mathcal{E}_{h,N}$. Additionally, we assume without loss of generality that $h_K < 1$, $K \in \mathcal{K}_h$. This is always achievable by scaling the domain.

Finally, the conforming approximation space $V_h^k \subset H^1_D(\Omega)$ of order k with

$$V_h^k = \{v \in H^1_D(\Omega) : \Delta v|_K \in \mathcal{P}^{k-2}(K), K \in \mathcal{K}_h \text{ and } v|_E \in \mathcal{P}^k(E), E \in \mathcal{E}_h\}$$

is introduced. Here $\mathcal{P}^p(\cdot)$ denotes the space of polynomials of degree smaller or equal p over the elements and edges, respectively. In the case of the diffusion equation, which is studied here, the approximation space V_h^k for the BEM-based FEM is the same as for the Virtual Element Method. But, in contrast to the Virtual Element Method, the BEM-based FEM makes use of an explicit basis of V_h^k.

3 BEM-based FEM

The discrete space V_h^k is constructed by prescribing its basis functions that are subdivided into nodal, edge and element basis functions. Each of them is locally (element-wise) the unique solution of a boundary value problem. We give a brief review of the approach in [20], more details concerning the implementation can also be found in [13].

For each node we define the function ψ_z, $z \in \mathcal{N}_h$ such that

$$-\Delta \psi_z = 0 \quad \text{in } K \quad \text{for all } K \in \mathcal{K}_h,$$

$$\psi_z(x) = \begin{cases} 1 & \text{for } x = z, \\ 0 & \text{for } x \in \mathcal{N}_h \setminus \{z\}, \end{cases}$$

ψ_z is linear on each edge of the mesh.

For $E \in \mathcal{E}_h$, let $z_0, z_1 \in \mathcal{N}_h$ with $E = \overline{z_0 z_1}$. We set $p_{E,0} = \psi_{z_0}|_E$ and $p_{E,1} = \psi_{z_1}|_E$. Furthermore, let $\{p_{E,i} : i = 0, \dots, k\}$ be a basis of $\mathcal{P}^k(E)$. Then the edge basis functions $\psi_{E,i}$ for $i = 2, \dots, k$, $E \in \mathcal{E}_h$ are given by

$$\begin{aligned} -\Delta \psi_{E,i} &= 0 \quad \text{in } K \quad \text{for all } K \in \mathcal{K}_h, \\ \psi_{E,i} &= \begin{cases} p_{E,i} & \text{on } E, \\ 0 & \text{on } \mathcal{E}_h \setminus \{E\}. \end{cases} \end{aligned}$$

Last but not least, we have to incorporate the non-harmonic functions. Consequently, we define the element basis functions $\psi_{K,i,j}$ for $i = 0, \dots, k-2$ and $j = 0, \dots, i$, $K \in \mathcal{K}_h$ by

$$\begin{aligned} -\Delta \psi_{K,i,j} &= p_{K,i,j} \quad \text{in } K, \\ \psi_{K,i,j} &= 0 \qquad \text{else}, \end{aligned}$$

where $\{p_{K,i,j} : i = 0, \dots, k-2 \text{ and } j = 0, \dots, i\}$ is a basis of $\mathcal{P}^{k-2}(K)$.

All these functions are well defined. They are given locally as unique solutions of a boundary value problems and, due to the local Dirichlet boundary data, they are continuous over the whole domain Ω. It is easily seen that $\{\psi_z, \psi_{E,i}, \psi_{K,i,j}\}$ forms a basis of V_h^k, since each function $\psi \in V_h^k$ can be expressed locally as solution of the boundary value problem

$$\begin{aligned} -\Delta \psi &= p_K \quad \text{in } K, \\ \psi &= p_{\partial K} \quad \text{on } \partial K, \end{aligned} \tag{3}$$

with $p_K \in \mathcal{P}^{k-2}(K)$ and $p_{\partial K} \in \mathcal{P}^k_{\mathrm{pw}}(\partial K)$. Here, $\mathcal{P}^k_{\mathrm{pw}}(\partial K) = \mathcal{P}^k_{\mathrm{pw,d}}(\partial K) \cap C^0(\partial K)$ and

$$\mathcal{P}^k_{\mathrm{pw,d}}(\partial K) = \{p \in L_2(\partial K) : p|_E \in \mathcal{P}^k(E), E \in \mathcal{E}(K)\}.$$

The Galerkin approximation of (2) reads

$$\text{Find } u_h \in V_h^k : \quad b(u_h, v_h) = (f, v_h)_{L_2(\Omega)} + (g_N, v_h)_{L_2(\Gamma_N)} \quad \forall v_h \in V_h^k. \tag{4}$$

For $\psi \in V_h^k$, $v \in H^1_D(\Omega)$ and $a(\cdot) = a_K$ on each $K \in \mathcal{K}_h$, it is

$$b(\psi, v) = \sum_{K \in \mathcal{K}_h} a_K (\nabla \psi, \nabla v)_{L_2(K)} = \sum_{K \in \mathcal{K}_h} a_K \left\{ (\gamma_1^K \psi, \gamma_0^K v)_{L_2(\partial K)} - (\Delta \psi, v)_{L_2(K)} \right\},$$

where $\gamma_0^K v$ is the usual trace of v on ∂K and $\gamma_1^K \psi \in H^{-1/2}(\partial K)$ is the Neumann trace of ψ on ∂K, which is given for sufficient regular ψ by

$$\gamma_1^K \psi(x) = \lim_{K \ni \tilde{x} \to x} n_K(x) \cdot \nabla \psi(\tilde{x}) \quad \text{for } x \in \partial K.$$

Since the Neumann trace is unknown in general, we approximate $\gamma_1^K \psi$ by $\widetilde{\gamma_1^K \psi} \in \mathcal{P}^k_{\mathrm{pw,d}}(\partial K)$ using a Galerkin approximation for a boundary integral equation connecting the Dirichlet and Neumann trace. If ψ is given by (3) with $p_K = 0$, that is always achievable by homogenization with a polynomial, the utilized formulation of the Boundary Element Method (BEM) reads:

$$\text{Find } \widetilde{\gamma_1^K \psi} \in \mathcal{P}^k_{\mathrm{pw,d}}(\partial K) : \ \left(\mathbf{V}_K \widetilde{\gamma_1^K \psi}, q \right)_{L_2(\partial K)} = \left(\left(\tfrac{1}{2} \mathbf{I} + \mathbf{K}_K \right) p_{\partial K}, q \right)_{L_2(\partial K)} \ \forall q \in \mathcal{P}^{k-1}_{\mathrm{pw,d}}(\partial K).$$

This formulation involves the single layer potential $\mathbf{V}_K : H^{-1/2}(\partial K) \to H^{1/2}(\partial K)$ and the double layer potential $\mathbf{K}_K : H^{1/2}(\partial K) \to H^{1/2}(\partial K)$, which are boundary integral operators. Thus, the task to find an approximation of the Neumann trace is reduced to a one-dimensional problem on the boundary ∂K. Furthermore, we have the representation formula

$$\psi(x) = \int_{\partial K} U^*(x,y)\gamma_1^K\psi(y)\,ds_y - \int_{\partial K} \gamma_{1,y}^K U^*(x,y)\gamma_0^K\psi(y)\,ds_y \quad \text{for } x \in K, \tag{5}$$

where $U^*(x,y) = -\frac{1}{2\pi}\ln(|x-y|)$. More details can be found in the literature on boundary integral equations and Boundary Element Methods, see, e.g., [12, 15].

Next, an approximation to the variational formulation (4) is given. Let $\widetilde{v}$ be a piecewise polynomial and discontinuous approximation of $v \in H_D^1(\Omega)$ over the mesh $\mathcal{K}_h$, which is obtained by local averaging for example. As proposed in [22], the approximated discrete variational formulation is defined by

$$\text{Find } u_h \in V_h^k : \quad b_h(u_h, v_h) = (f, \widetilde{v}_h)_{L_2(\Omega)} + (g_N, v_h)_{L_2(\Gamma_N)} \quad \forall v_h \in V_h^k, \tag{6}$$

with

$$b_h(\psi, v) = \sum_{K\in\mathcal{K}_h} a_K \left\{ (\widetilde{\gamma_1^K\psi}, \gamma_0^K v)_{L_2(\partial K)} - (\Delta\psi, \widetilde{v})_{L_2(K)} \right\}, \quad \psi \in V_h^k, v \in H_D^1(\Omega),$$

where $\widetilde{\gamma_1^K\psi} \in \mathcal{P}_{\mathrm{pw,d}}^{k-1}(\partial K)$ is the BEM approximation and $\widetilde{v}$ is the polynomial approximation of v over each K. This form of the approximated bilinear form is especially suited for the a-posteriori error analysis since the second argument of $b_h(\cdot,\cdot)$ is only assumed to be in $H_D^1(\Omega)$. For the computational realization one might exploit the additional properties of $v_h \in V_h^k$ to treat the volume integral $(\Delta\psi, \widetilde{v}_h)_{L_2(K)}$, see [13, 20].

4 Residual based Error Estimate

In this section, the residual based error estimate is reviewed on polygonal meshes and the reliability as well as the efficiency are stated. This estimate bounds the error of the finite element computation in the energy norm, which is given by $\|v\|_{b,\omega}^2 = (a\nabla v, \nabla v)_{L_2(\omega)}$ over a subset $\omega \subset \Omega$. Due to the homogenious Dirichlet boundary data, $\|\cdot\|_{b,\Omega}$ is equivalent to $\|\cdot\|_{H^1(\Omega)}$ on $H_D^1(\Omega)$. The discrete jump of the approximation of the conormal derivatives over an internal edge $E \in \mathcal{E}_{h,\Omega}$ is defined by

$$[\![u_h]\!]_{E,h} = a_K\widetilde{\gamma_1^K u_h} + a_{K'}\widetilde{\gamma_1^{K'} u_h},$$

where $K, K' \in \mathcal{K}_h$ are the adjacent elements of E with $E \in \mathcal{E}(K) \cap \mathcal{E}(K')$.

Theorem 1 (Reliability)**.** *Let $\mathcal{K}_h$ be a regular mesh. Furthermore, let $u \in H_D^1(\Omega)$ and $u_h \in V_h^k$ be the solutions of* (2) *and* (6)*, respectively. Then, it is*

$$\|u - u_h\|_{b,\Omega} \le c\,\left\{\eta_R^2 + \delta_R^2\right\}^{1/2} \quad \textit{with} \quad \eta_R^2 = \sum_{K\in\mathcal{K}_h}\eta_K^2 \quad \textit{and} \quad \delta_R^2 = \sum_{K\in\mathcal{K}_h}\delta_K^2,$$

where the error indicators are defined by

$$\eta_K^2 = h_K^2\|R_K\|_{L_2(K)}^2 + \sum_{E\in\mathcal{E}(K)} h_E\|R_E\|_{L_2(E)}^2, \qquad \delta_K^2 = \|a_K\gamma_1^K u_h - a_K\widetilde{\gamma_1^K u_h}\|_{L_2(\partial K)}^2,$$

with the element residual

$$R_K = f + a_K \Delta u_h \quad \textit{for } K \in \mathcal{K}_h,$$

and the edge residual

$$R_E = \begin{cases} 0 & \textit{for } E \in \mathcal{E}_{h,D}, \\ g_N - a_K \widetilde{\gamma_1^K u_h} & \textit{for } E \in \mathcal{E}_{h,N} \textit{ with } E \in \mathcal{E}(K), \\ -\frac{1}{2}[\![u_h]\!]_{E,h} & \textit{for } E \in \mathcal{E}_{h,\Omega}. \end{cases}$$

The constant $c > 0$ only depends on the regularity parameters $\sigma_{\mathcal{K}}$, $c_{\mathcal{K}}$, the approximation order k and on the diffusion coefficient a.

The term δ_K measures the approximation error in the Neumann traces of the basis functions of V_h^k coming from the Boundary Element Method. To state the efficiency, we introduce the neighbourhood ω_K of an element $K \in \mathcal{K}_h$. Let ω_K be an open subset of Ω such that

$$\overline{\omega}_K = \bigcup \{x \in \overline{K'} : \mathcal{E}(K) \cap \mathcal{E}(K') \neq \emptyset, K' \in \mathcal{K}_h\}.$$

Theorem 2 (Efficiency)**.** *Under the assumptions of Theorem 1, the following local bound is fulfilled*

$$\begin{aligned} \eta_K \;\leq\; c \Bigg(& \|u - u_h\|_{b,\omega_K}^2 + h_K^2 \|f - \widetilde{f}\|_{L_2(\omega_K)}^2 + \sum_{E \in \mathcal{E}(K) \cap \mathcal{E}_{h,N}} h_E \|g_N - \widetilde{g}_N\|_{L_2(E)}^2 \\ & + \sum_{E \in \mathcal{E}(K)} \sum_{K' \subset \omega_E} h_E \|a_{K'} \gamma_1^{K'} u_h - a_{K'} \widetilde{\gamma_1^{K'} u_h}\|_{L_2(E)}^2 \Bigg)^{1/2}, \end{aligned}$$

where $\widetilde{f}$ and $\widetilde{g}_N$ are piecewise polynomial approximations of the data f and g_N, respectively. The constant $c > 0$ only depends on the regularity parameters $\sigma_{\mathcal{K}}$, $c_{\mathcal{K}}$, the approximation order k and on the diffusion coefficient a.

The terms involving the data approximation $\|f - \widetilde{f}\|_{L_2(\omega_K)}$ and $\|g_N - \widetilde{g}_N\|_{L_2(E)}$ are often called data oscillations. They are usually of higher order. Furthermore, the approximation of the Neumann traces by the Boundary Element Method appear in the right hand side. This term is related to δ_K.

5 Numerical experiments

For the numerical validation and verification, we consider the two test cases given in [4] for an adaptive Virtual Element Method (VEM). The boundary value problems are given on bounded polygonal domains Ω as

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D = \partial\Omega. \tag{7}$$

The Dirichlet data $g_D \in C(\Gamma_D)$ is treated in the usual way by a discrete extension into the domain Ω to obtain homogeneous data on Γ_D. Therefore, problem (7) takes the form (1). In the following, (7) is approximated on a sequence of uniformly and adaptively refined meshes, where the initial meshes coincide with the ones in [4]. For the refinement itself we chose a different procedure than in [4], namely the algorithm from [18].

In the uniform refinement strategy each element of the mesh is split into two new elements to obtain the next finer mesh. This splitting process is performed as described below. For the adaptive BEM-based FEM, we proceed in a common strategy, which involves a loop over the following four steps:

SOLVE The boundary value problem (1) is approximated by means of the BEM-based FEM on the current polygonal mesh using the approximation space V_h^k.

ESTIMATE The residual based error estimator η_R as well as the indicators η_K discussed in Section 4 are computed over each element of the discretization $\mathcal{K}_h$.

MARK A minimal subset $\mathcal{M}_h \subset \mathcal{K}_h$ of all elements are marked according to Dörflers strategy [8] such that

$$\left(\sum_{K \in \mathcal{M}_h} \eta_K^2 \right)^{1/2} \geq (1-\theta)\, \eta_R,$$

where $0 \leq \theta < 1$ is a user defined parameter. To obtain a minimal set $\mathcal{M}_h$, it is possible to sort the elements according to their indicators η_K and mark those with the largest indicators. Instead of that, we implemented the marking algorithm given in [8] to achieve linear complexity. Furthermore, we choose $\theta = 0.5$ in the numerical experiments.

REFINE Each marked element is refined and we consequently obtain a new mesh for the next cycle in the loop. For the refinement of an element K, we bisect K through its barycenter $\bar{x}$ orthogonal to its characteristic direction that is given as eigenvector to the largest eigenvalue of the matrix

$$M_{\mathrm{Cov}} = \int_K (x-\bar{x})(x-\bar{x})^\top dx, \qquad \bar{x} = \frac{1}{|K|} \int_K x\, dx.$$

For more details see [18, 19]. Furthermore, we check the regularity of the mesh and refine additional elements if the condition $h_K < c_{\mathcal{K}} h_E$ for $E \in \mathcal{E}(K)$ is violated with a user defined parameter $c_{\mathcal{K}}$.

The adaptive mesh refinement process is kept very local. Only the marked and degenerated elements are bisected during the refinement. It is not necessary to resolve hanging nodes and keep the mesh admissible as for example in the *red-blue-green* refinement procedure for triangular meshes, see [16]. This advantage is due to the polygonal meshes with very flexible elements. The local refinement character can be seen in the following test problems in the Figures 1 and 3.

To analyse the experiments, we compute the relative error in the energy norm over each mesh in the convergence process. In contrast to the Virtual Element Method, it is possible to evaluate the approximation u_h inside of the elements by means of the representation formula (5). Therefore, we approximate $\|u - u_h\|_{b,\Omega}$ by Gaussian quadrature over a fine auxiliary triangulation of the domain that is aligned with the polygonal mesh. The mesh size $h_{\max} = \max\{h_K : K \in \mathcal{K}_h\}$ does not uniformly tend to zero on adaptively refined meshes. Thus, the relative errors $\|u-u_h\|_{b,\Omega}/\|u\|_{b,\Omega}$ are plotted with respect to the number of degrees of freedom (DoF) in logarithmic scale. Since $\mathrm{DoF} = \mathcal{O}(h_{\max}^{-2})$ for a sequence of uniformly refined meshes, we expect that the error in the energy norm behaves like $\mathcal{O}(\mathrm{DoF}^{-k/2})$ for an optimal method with approximation order k.

5.1 Test case 1: singular solution

Let $\Omega = ((-1,1)\times(-1,1)) \setminus ([0,1]\times[-1,0])$ be a L-shaped domain and $f = 0$ in (7). The Dirichlet data g_D is chosen such that

$$u(r\cos\varphi, r\sin\varphi) = r^{2/3} \sin\left(\tfrac{2}{3}\varphi\right), \quad x = (r\cos\varphi, r\sin\varphi) \in \mathbb{R}^2$$

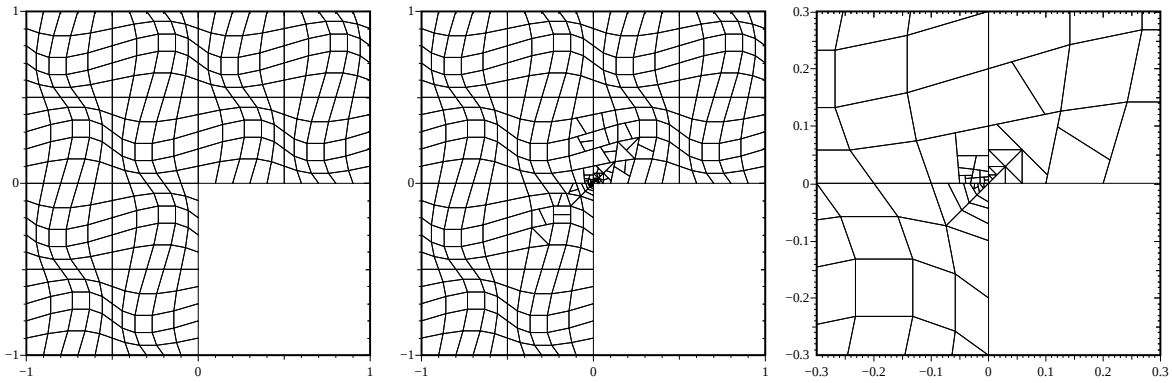


Figure 1: Test 1: initial mesh (left), adaptively refined mesh for $k = 1$ and $\|u - u_h\|_{b,\Omega}/\|u\|_{b,\Omega} \approx 0.02$ (middle), zoom of adaptively refined mesh for $k = 3$ and $\|u - u_h\|_{b,\Omega}/\|u\|_{b,\Omega} \approx 0.002$ (right).

is the exact solution of (7), where (r, φ) are polar coordinates. The initial mesh is visualized in Figure 1 (left). Furthermore, the adaptive refined meshes are given in this figure for the first order method with V_h^1 and a relative error of approximately 0.02 as well as for the third order method with V_h^3 and a relative error of approximately 0.002. In the right most picture a zoom of the mesh for V_h^3 is visualized since the adaptive refinement only affect elements close to the reentrant corner of the domain, where the singularity of the solution is located. The presented meshes were achieved after 11 and 16 refinement steps.

In Figure 2, the convergence graphs are given for the first, second and third order method and for the uniform as well as the adaptive strategy. In all three cases the adaptive BEM-based FEM yields optimal rates of convergence, namely a slope of $-k/2$ in the logarithmic plots. Since $u \in H^{5/3}(\Omega)$ is not sufficient regular, the convergence on uniform meshes slows down. The theory predicts a behaviour of the error independent of the approximation order k like $\mathcal{O}(h_{\max}^{2/3})$. This corresponds to $\mathcal{O}(\mathrm{DoF}^{-1/3})$, which is recovered in the convergence graphs in Figure 2.

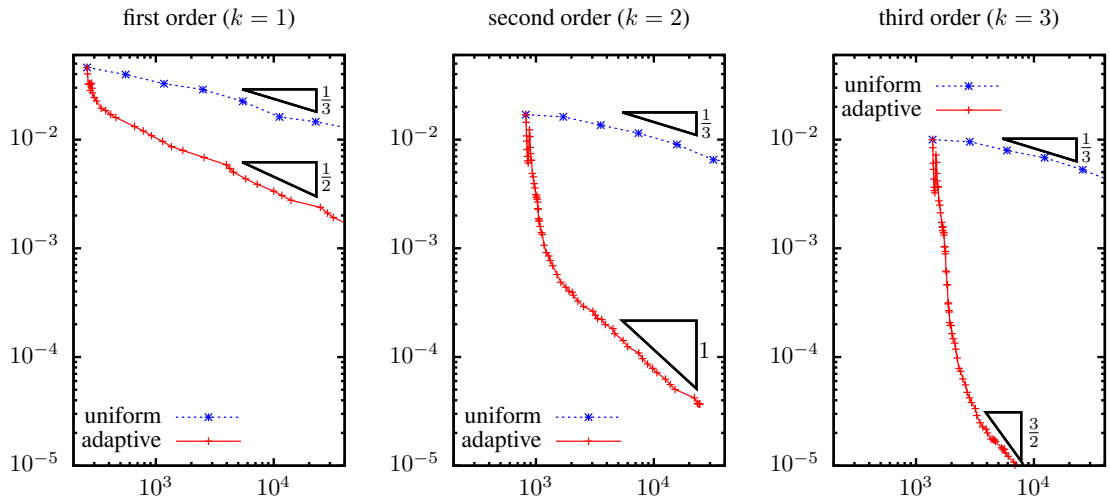


Figure 2: Test 1: convergence of the relative energy error $\|u - u_h\|_{b,\Omega}/\|u\|_{b,\Omega}$ with respect to the number of degrees of freedom for the approximation orders $k = 1, 2, 3$ on uniformly and adaptively refined meshes.

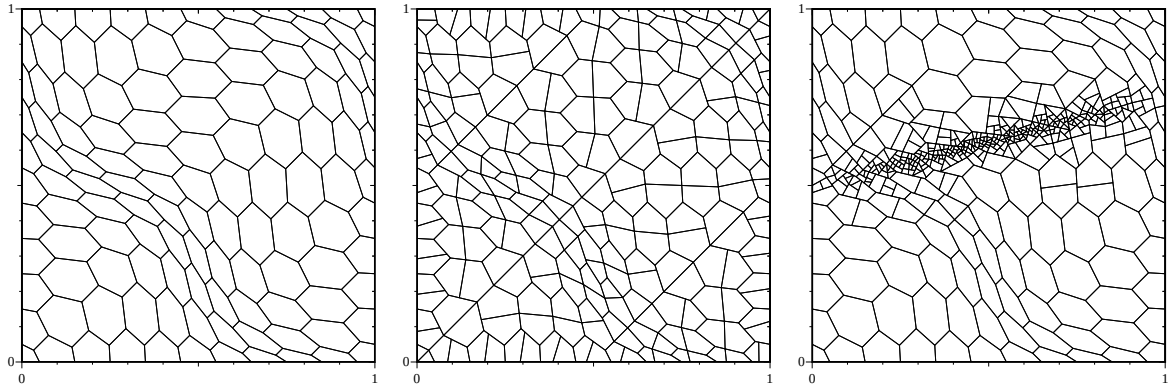


Figure 3: Test 2: initial mesh (left), uniformly refined mesh (middle), adaptively refined mesh for $k = 1$ and $\|u - u_h\|_{b,\Omega}/\|u\|_{b,\Omega} \approx 0.2$ (right).

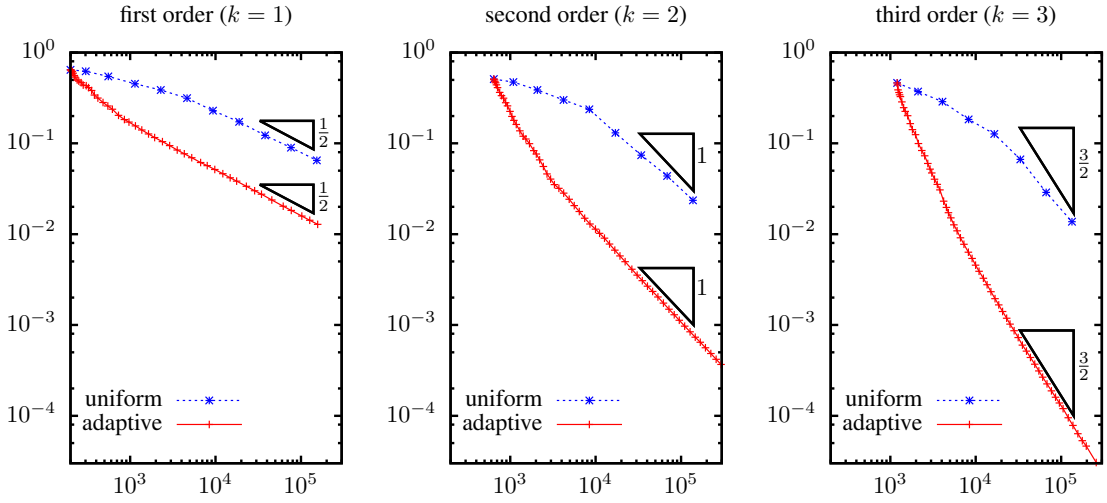


Figure 4: Test 2: convergence of the relative energy error $\|u - u_h\|_{b,\Omega}/\|u\|_{b,\Omega}$ with respect to the number of degrees of freedom for the approximation orders $k = 1, 2, 3$ on uniformly and adaptively refined meshes.

5.2 Test case 2: strong internal layer

Let $\Omega = (0, 1)^2$, $g_D = 0$ and f be chosen such that

$$u(x) = 16x_1(1 - x_1)x_2(1 - x_2)\arctan(25x_1 - 100x_2 + 50), \quad x = (x_1, x_2) \in \mathbb{R}^2$$

is the exact solution of (7). Since u is arbitrary smooth, we expect optimal rates of convergence in the case of uniform mesh refinement in an asymptotic regime. Although the solution u is smooth, it has a strong internal layer along the line $x_2 = 1/2 + x_1/4$. The initial mesh is visualized in Figure 3 (left). Furthermore, the first uniform refined mesh is given in the middle of Figure 3. Here, one recognizes that each polygonal cell is bisected in the refinement process. In the right most picture of Figure 3 the adaptively refined mesh for V_h^1 and a relative error of approximately 0.2 is presented. This mesh was achieved after 19 refinement steps. It is seen that the adaptive strategy refines along the internal layer of the exact solution.

In Figure 4, we give the convergence graphs for the first, second and third order method and for the uniform as well as the adaptive strategy. In all cases we recover the optimal convergence

rates which correspond to a slope of $-k/2$. But, for the uniform refinement, the internal layer has to be resolved sufficiently before the optimal rates are achieved. Since the adaptive strategy resolves the layer automatically, the adaptive BEM-based FEM is much more accurate for the same number of unknowns.

6 Conclusion

The numerical experiments have shown that polygonal meshes are appealing in adaptive FEM strategies. Their natural incorporation of hanging nodes allows very local mesh adaptations and facilitates the refinement process. The possibility to evaluate the approximation u_h inside of the elements in the BEM-based FEM turned out to be a useful feature. Nevertheless, the developments in this area are very recent and need further investigations to exploit their full potential.

Acknowledgements The Author would like to thank L. Beirão da Veiga and G. Manzini for providing the initial meshes used in the numerical tests.

REFERENCES

[1] P. F. Antonietti, L. Beirão da Veiga, C. Lovadina, and M. Verani. Hierarchical a posteriori error estimators for the mimetic discretization of elliptic problems. *SIAM J. Numer. Anal.*, 51(1):654–675, 2013.

[2] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.*, 23(01):199–214, 2013.

[3] L. Beirão da Veiga, K. Lipnikov, and G. Manzini. Arbitrary-order nodal mimetic discretizations of elliptic problems on polygonal meshes. *SIAM J. Numer. Anal.*, 49(5):1737–1760, 2011.

[4] L. Beirão da Veiga and G. Manzini. Residual *a posteriori* error estimation for the virtual element method for elliptic problems. *ESAIM Math. Model. Numer. Anal.*, 49(2):577–599, 2015.

[5] L. Chen, J. Wang, and X. Ye. A posteriori error estimates for weak Galerkin finite element methods for second order elliptic problems. *J. Sci. Comput.*, 59(2):496–511, 2014.

[6] D. Copeland, U. Langer, and D. Pusch. From the boundary element domain decomposition methods to local Trefftz finite element methods on polyhedral meshes. In *Domain decomposition methods in science and engineering XVIII*, volume 70 of *Lect. Notes Comput. Sci. Eng.*, pages 315–322. Springer, Berlin Heidelberg, 2009.

[7] V. Dolejší, M. Feistauer, and V. Sobotíková. Analysis of the discontinuous Galerkin method for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 194(25-26):2709–2733, 2005.

[8] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996.

[9] Y. Efendiev, J. Galvis, R. Lazarov, and S. Weißer. Mixed FEM for second order elliptic problems on polygonal meshes with BEM-based spaces. In I. Lirkov, S. Margenov, and J. Waśniewski, editors, *Large-Scale Scientific Computing*, Lecture Notes in Computer Science, pages 331–338. Springer, Berlin Heidelberg, 2014.

[10] C. Hofreither, U. Langer, and S. Weißer. Convection adapted BEM-based FEM. *ArXiv e-prints*, 2015. arXiv:1502.05954.

[11] O. A. Karakashian and F. Pascal. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. *SIAM J. Numer. Anal.*, 41(6):2374–2399 (electronic), 2003.

[12] W. C. H. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge, 2000.

[13] S. Rjasanow and S. Weißer. Higher order BEM-based FEM on polygonal meshes. *SIAM J. Numer. Anal.*, 50(5):2357–2378, 2012.

[14] S. Rjasanow and S. Weißer. FEM with Trefftz trial functions on polyhedral elements. *J. Comput. Appl. Math.*, 263:202–217, 2014.

[15] O. Steinbach. *Numerical approximation methods for elliptic boundary value problems: finite and boundary elements*. Springer, New York, 2007.

[16] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[17] J. Wang and X. Ye. A weak Galerkin mixed finite element method for second order elliptic problems. *Math. Comp.*, 83(289):2101–2126, 2014.

[18] S. Weißer. Residual error estimate for BEM-based FEM on polygonal meshes. *Numer. Math.*, 118(4):765–788, 2011.

[19] S. Weißer. *Finite Element Methods with local Trefftz trial functions*. PhD thesis, Universität des Saarlandes, Saarbrücken, Germany, September 2012.

[20] S. Weißer. Arbitrary order Trefftz-like basis functions on polygonal meshes and realization in BEM-based FEM. *Comput. Math. Appl.*, 67(7):1390–1406, 2014.

[21] S. Weißer. BEM-based finite element method with prospects to time dependent problems. In E. Oñate, J. Oliver, and A. Huerta, editors, *Proceedings of the jointly organized WCCM XI, ECCM V, ECFD VI, Barcelona, Spain, July 2014*, pages 4420–4427. International Center for Numerical Methods in Engineeering (CIMNE), 2014.

[22] S. Weißer. Residual based error estimate and quasi-interpolation on polygonal meshes for high order BEM-based FEM. *ArXiv e-prints*, 2015. arXiv:1511.08993.

[23] S. Weißer. Residual Based Error Estimate for Higher Order Trefftz-Like Trial Functions on Adaptively Refined Polygonal Meshes. In A. Abdulle, S. Deparis, D. Kressner, F. Nobile, and M. Picasso, editors, *Numerical Mathematics and Advanced Applications - ENUMATH 2013*, volume 103 of *Lecture Notes in Computational Science and Engineering*, pages 233–241. Springer International Publishing, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

VEM AND TOPOLOGY OPTIMIZATION ON POLYGONAL MESHES

P.F. Antonietti[1], M. Bruggi[2], S. Scacchi[3] and M. Verani[1]

[1] MOX-Department of Mathematics
Politecnico di Milano
Piazza Leonardo da Vinci, 32 20133 Milano, Italy
e-mail: paola.antonietti@polimi.it, marco.verani@polimi.it
[2] Department of Civil and Environmental Engineering DICA
Politecnico di Milano
Piazza Leonardo da Vinci, 32 20133 Milano, Italy
e-mail: matteo.bruggi@polimi.it
[3] Department of Mathematics
University of Milan
Via Saldini, 50 20133 Milano, Italy
e-mail: simone.scacchi@unimi.it

Keywords: Topology Optimization, Virtual Element Method

Abstract. *Topology optimization is a fertile area of research that is mainly concerned with the automatic generation of optimal layouts to solve design problems in Engineering. The classical formulation addresses the problem of finding the best distribution of an isotropic material that minimizes the work of the external loads at equilibrium, while respecting a constraint on the assigned amount of volume. This is the so-called minimum compliance formulation that can be conveniently employed to achieve stiff truss-like layout within a two-dimensional domain. A classical implementation resorts to the adoption of four node displacement-based finite elements that are coupled with an elementwise discretization of the (unknown) density field. When regular meshes made of square elements are used, well-known numerical instabilities arise, see in particular the so-called checkerboard patterns. On the other hand, when unstructured meshes are needed to cope with geometry of any shape, additional instabilities can steer the optimizer towards local minima instead of the expected global one. Unstructured meshes approximate the strain energy of truss-like members with an accuracy that is strictly related to the geometrical features of the discretization, thus remarkably affecting the achieved layouts. In this paper we will consider several benchmarks of truss design and explore the performance of the recently proposed technique known as the Virtual Element Method (VEM) in driving the topology optimization procedure. In particular, we will show how the capability of VEM of efficiently approximating elasticity equations on very general polygonal meshes can contribute to overcome the aforementioned mesh-dependent instabilities exhibited by classical finite element based discretization techniques.*

1 INTRODUCTION

Topology optimization is a design tool that is extensively exploited in problems related to Engineering and its industrial applications. The classical formulation searches for the best distribution of a prescribed amount of isotropic material that minimizes the so–called compliance, e.g., twice the work of the external loads at equilibrium [7]. This framework is based on the adoption of a suitable penalization of the mechanical properties of the elastic body depending on the local values of the density field, see e.g. the well–known SIMP (Solid Isotropic Material with Penalization) [8]. A classical approach to the solution of a topology optimization problem resorts to the discretization of the density field to cope with the minimization setting, along with the adoption of finite element methods to tackle the approximation of the equilibrium equation. Methods of mathematical programming are adopted to solve the arising minimization problem, see [24]. In general, low–order displacement–based finite elements are coupled with an element–wise density discretization, being affected by numerical instabilities as checkerboard and mesh dependence. The former drawback depends on the choice of the displacement shape functions with respect to the adopted density approximation, whereas the latter is due to an inherent ill-posedness of the continuous problem [22]. Checkerboard and mesh dependence are solved, in general, via ad hoc filtering procedures that preserve the advantages of the above mentioned discrete approach. However, this approach suffers from drawbacks that are peculiar to the adopted low–order discretization of the state equation. It is worth mentioning, among the others, a lack of accuracy in the approximation of the displacement and stress fields and the arising of undesired locking phenomena when considering incompressible media. Variations of the basic approach have been investigated to solve the first issue, see among the others [15, 25] focusing on adaptivity. Robust discretizations to cope with incompressible materials have also been addressed, see the u-p formulation [23] and mixed finite elements [14]. Moreover, when unstructured meshes are needed to deal with geometry of any shape, additional instabilities can steer the optimizer towards local minima, instead of the expected global one, when using the classical approach. Unstructured meshes approximate the strain energy of optimal truss–like members with an accuracy that is strictly related to the geometrical features of the approximation, thus remarkably affecting the achieved layouts.

The aim of this paper is to investigate whether the recently proposed Virtual Element Method (VEM), cf. [3], can be effectively employed for topology optimization problems. Following the pioneering work in [17], several benchmarks of truss design will be investigated to explore the performance the Virtual Element Method when employed to drive the topology optimization algorithm. More precisely, we will demonstrate that the inherited flexibility of VEM to deal with general polygonal meshes can contribute to overcome the mesh-dependent instabilities featured by classical discretization techniques.

The Virtual Element Method (see, e.g., [3] for an introduction to the method and [4] for the details of its practical implementation) is characterized by the capability of dealing with very general polygonal/polyedral meshes and by the possibility of easily implementing highly regular discrete spaces. Indeed, by avoiding the explicit construction of the local basis functions, the VEM can easily handle general polygons/polyhedrons without complex integrations on the element. In addition, thanks to this added flexibility, it was discovered [12, 6] that virtual elements can also be used to build global discrete spaces of arbitrary regularity (C^1 and more) that are quite simple in terms of degrees of freedom and coding. So far, VEM has been applied to

a variety of different problems [11, 1, 5, 9, 10, 16, 18, 20, 21, 2]. However, apart from [17] a systematic study of the impact of VEM on topology optimization problems is still in its infancy. The aim of this paper is to contribute in filling this gap.

The outline of the paper is as follows. Section 2.1 recalls the classical topology optimization problem, whereas Section 2.2 provides fundamentals of the Virtual Element Method and introduces the VEM discrete volume–constrained minimum compliance problem. Section 2.3 addresses numerical instabilities, i.e. checkerboard and mesh dependence. Section 3 presents numerical simulations performed on benchmark examples that are based on the adoption of structured and unstructured meshes of both polygonal and quadrangular elements. Section 4 provides final remarks and outlines the ongoing research.

2 GOVERNING EQUATIONS

2.1 *Continuous problem*

Topology optimization for the maximization of the structural stiffness is herein addressed, according to the well–known *minimum compliance* formulation. The problem of distributing a given amount of linear elastic isotropic material is solved, such that the work of the external load against the corresponding displacement at equilibrium is minimized.

Let $\Omega \subset \mathbb{R}^2$ be an open, bounded domain of $\mathbb{R}^2$ with Lipschitz boundary $\Gamma = \partial\Omega$. Let $\rho \in \mathcal{Q}_{ad} = \{\rho \in L^\infty(\Omega) : \quad 0 < \rho_{min} \leq \rho \leq 1 \text{ a.e. in } \Omega\}$ be a bounded function representing the material density in Ω, where ρ_{min} is some positive lower bound that is already introduced to avoid any singularity in the analysis of the further derived discrete problem. Let $\mathbf{C} = \mathbf{C}(\rho(x))$ be a fourth order elasticity tensor depending on the material density at the point $x \in \Omega$. According to the well–known SIMP model, see e.g. [8], we assume

$$\mathbf{C}(\rho(x)) = \rho(x)^p \mathbf{C}^0 \quad p > 1 \,,$$

where

$$C^0_{ijhk} = \frac{E\nu}{(1+\nu)(1-2\nu)}\delta_{ij}\delta_{hk} + \frac{E}{2(1+\nu)}(\delta_{ih}\delta_{jk} + \delta_{ik}\delta_{jh}) \tag{1}$$

is the stiffness tensor for an assigned isotropic medium with Young modulus E and Poisson's ratio ν, whereas p is a penalization parameter that is usually assumed to be equal to 3, see e.g. [7].

Let $\mathbf{u}_d$ and $\mathbf{f}_t$ denote the displacement and traction prescribed on $\Gamma_d \subset \Gamma$ and $\Gamma_t = \Gamma \setminus \Gamma_d$, respectively. For a given density material distribution ρ, the displacement field $\mathbf{u}$ solves the following state problem

$$\begin{aligned}
\text{div } \sigma + \mathbf{g} &= \mathbf{0} && \text{in } \Omega && \text{(2a)}\\
\sigma &= \mathbf{C}(\rho)\epsilon(\mathbf{u}) && && \text{(2b)}\\
\mathbf{u} &= \mathbf{u}_d && \text{on } \Gamma_d && \text{(2c)}\\
\sigma \cdot \mathbf{n} &= \mathbf{f}_t && \text{on } \Gamma_t, && \text{(2d)}
\end{aligned}$$

where $\epsilon(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + \nabla^t \mathbf{u}) \equiv \nabla^s \mathbf{u}$. According to a frequent assumption in topology optimization, no body load will be considered in the sequel, meaning that $\mathbf{g} = \mathbf{0}$.

Let us introduce the semi-linear form $a(\rho;\cdot,\cdot) : \mathcal{Q}_{ad} \times [H^1(\Omega)]^2 \times [H^1(\Omega)]^2 \to \mathbb{R}$

$$a(\rho;\mathbf{u},\mathbf{v}) := \int_{\mathbf{\Omega}} \mathbf{C}(\rho)\epsilon(\mathbf{u}) : \epsilon(\mathbf{v})\ \mathbf{dx}$$

and the linear functional $\mathcal{F}(\cdot) : [H^1(\Omega)]^2 \to \mathbb{R}$

$$\mathcal{F}(\mathbf{v}) = \int_{\mathbf{\Gamma_t}} \mathbf{f_t} \cdot \mathbf{v}\ \mathbf{dx}.$$

We introduce the following spaces

$$\begin{aligned} \mathcal{V}_0 &= \{\mathbf{u} \in [H^1(\Omega)]^2 : \quad \mathbf{u} = \mathbf{0} \text{ on } \Gamma_d\} \\ \mathcal{V} &= \{\mathbf{u} \in [H^1(\Omega)]^2 : \quad \mathbf{u} = \mathbf{u_d} \text{ on } \Gamma_d\}\ . \end{aligned}$$

Then the weak formulation of the problem (2a)-(2d) reads as follows: for a given $\rho \in \mathcal{Q}_{ad}$ find $\mathbf{u} \in \mathcal{V}$ such that

$$a(\rho;\mathbf{u},\mathbf{v}) = \mathcal{F}(\mathbf{v}) \tag{3}$$

for all $\mathbf{v} \in \mathcal{V}_\mathbf{0}$.

According to the Clapeyron theorem, the continuous formulation of the topology optimization problem for minimum compliance may be therefore written as:

$$\begin{cases} \min\limits_{\rho\in\mathcal{Q}_{ad}} \quad \mathcal{C}(\rho,\mathbf{u}) = \displaystyle\int_{\Gamma_t} \mathbf{f}_t \cdot \mathbf{u}\ dx = \int_\Omega \rho^p \mathbf{C}_0\ \epsilon(\mathbf{u}) : \epsilon(\mathbf{u})\ dx \\ \text{s.t.} \qquad a(\rho;\mathbf{u},\mathbf{v}) = \mathcal{F}(\mathbf{v})\ \ \forall \mathbf{v} \in \mathcal{V}_\mathbf{0} \\ \qquad\quad \dfrac{1}{V}\displaystyle\int_\Omega \rho dx\ \leq V_f, \end{cases} \tag{4}$$

being V_f the available amount of material as a fraction of the whole domain $V = \int_\Omega 1 dx$. Minimizing the compliance $\mathcal{C}$ of a structure acted upon by a prescribed set of assigned forces means minimizing the work of external loads, i.e. looking for a stiff structure.

2.2 Virtual Element discretization

In this section we introduce the Virtual Element discretization of problem (4). From now on, we will assume that Ω is a polygonal domain in $\mathbb{R}^2$. Let $\mathcal{T}_h$ represent a decomposition of Ω into general, possibly non-convex, polygonal elements E with $\text{diam}(E) = h_E$, where $\text{diam}(E) = \max_{x,y\in E} \|x - y\|$. In the following, we will denote by e the straight edges of the mesh $\mathcal{T}_h$ and, for all $e \in \partial E$, $\mathbf{n}_E^e$ will denote the unit normal vector to e pointing outward to E. We will use the symbol $\mathbb{P}_k(\omega)$ to denote the space of polynomials of degree less than or equal to k living on the set $\omega \subseteq \mathbb{R}^2$.

Let us first introduce the finite dimensional space of piecewise constant admissible controls

$$\mathbb{Q}_{ad} = \{\rho_h \in \mathcal{Q}_{ad} :\ \rho_{h|E} \in \mathbb{P}_0(E)\ \ \forall E \in \mathcal{T}_h\}\ .$$

Clearly, a function $\rho_h \in \mathbb{Q}_{ad}$ is uniquely determined by its value ρ_E in each polygon $E \in \mathcal{T}_h$. Hence, the dimension of $\mathbb{Q}_{ad}$ equals the cardinality of $\mathcal{T}_h$.

Moreover, following [5], it is possible to introduce the low-order discrete VEM spaces $\mathbf{V}_{0,h} \subset \mathcal{V}_0$ and $\mathbf{V}_h \subset \mathcal{V}$, a discrete form $a_h(\rho_h; \mathbf{u}_h, \mathbf{v}_h)$ approximating $a(\rho; \mathbf{u}, \mathbf{v})$ and a discrete functional $\mathcal{F}_h(\mathbf{v}_h)$ approximating $\mathcal{F}(\mathbf{v})$ such that the VEM discretization of (3) reads as: given $\rho_h \in \mathbb{Q}_{ad}$ find $\mathbf{u}_h \in \mathcal{V}_h$ such that

$$a_h(\rho_h; \mathbf{u}_h, \mathbf{v_h}) = \mathcal{F}_\mathbf{h}(\mathbf{v_h}) \tag{5}$$

for any $\mathbf{v_h} \in \mathcal{V}_{\mathbf{0},\mathbf{h}}$. In particular, the global VEM spaces $\mathbf{V}_{0,h}$ and $\mathbf{V}_h$ are obtained by gluing suitable local discrete VEM spaces, denoted by $\mathcal{V}_h(E)$, whose elements are uniquely identified by the values at the vertices of the polygon E and contain linear polynomials, i.e $\mathbb{P}_1(E) \subset \mathcal{V}_h(E)$. It is immediate to verify that the dimension of $\mathbf{V}_{0,h}$ (the same happens for $\mathbf{V}_h$) equals the number of the interior vertices of the partition $\mathcal{T}_h$ plus those belonging to Γ_t, having fixed the values at vertices belonging to Γ_d to incorporate Dirichlet boundary conditions. The construction of the global form $a_h(\rho_h; \mathbf{u}_h, \mathbf{v_h})$ hinges upon the construction of local forms $a_h^E(\mathbf{u}_h, \mathbf{v}_h) : \mathcal{V}_h(E) \times \mathcal{V}_h(E) \to \mathbb{R}$ approximating the integrals $\int_E \mathbf{C}_0\, \epsilon(\mathbf{u}) : \epsilon(\mathbf{v})\, dx$ so that

$$a_h(\rho_h; \mathbf{u}_h, \mathbf{v}_h) = \sum_{E \in \mathcal{T}_h} \rho_E^p a_h^E(\mathbf{u}_h, \mathbf{v}_h) \tag{6}$$

where we employed the fact that $\rho_h|_E \in \mathbb{P}_0(E)$.

It is important to remark that the the linear system stemming from (5) can be assembled employing only the values of u_h and v_h at the vertices of $\mathcal{T}_h$ (see [5] for more details) and the values of ρ_h on each polygon. Using (6) it is immediate to see that, with a slight abuse of notation in denoting the vector of densities, the algebraic version of (5) reads as

$$\mathbf{K}(\rho_h)\, \mathbf{U} \;=\; \sum_{E \in \mathcal{T}_h} \rho_E^p \mathbf{K}_E^0 \mathbf{U}_E \;=\; \mathbf{F}, \tag{7}$$

where:

- ρ_h is the vector of the element densities, whose components are ρ_E;
- $\mathbf{K}$ is the global stiffness matrix depending on ρ_h;
- $\mathbf{K}_E^0$ is the element stiffness matrix for the virgin material, i.e. associated to $a_h^E(\cdot, \cdot)$;
- $\mathbf{U}$ is the generalized displacement vector;
- $\mathbf{U}_E$ is the element displacement vector;
- $\mathbf{F}$ is the load vector.

Employing (5) and (7) the discrete version of the topology optimization problem (4) reads as follows

$$\begin{cases} \min\limits_{\rho_{min} \le \rho_E \le 1} & \mathcal{C}_h(\rho_h, \mathbf{u}_h) := \mathbf{U}^T \mathbf{K} \mathbf{U} = \sum\limits_{E \in \mathcal{T}_h} \rho_E^p \mathbf{U}_e^T \mathbf{K}_e^0 \mathbf{U}_e \\ \text{s.t.} & \mathbf{K}(\mathbf{x})\, \mathbf{U} \;=\; \mathbf{F}, \\ & \mathcal{W} \,/\, \mathcal{W}_0 \;\le\; V_f, \end{cases} \tag{8}$$

In the above equation, the objective function is the structural compliance $\mathcal{C}_h$, Eqn. $(8)_2$ enforces the discrete equilibrium equation discussed in Section 2.1 and Eqn. $(8)_3$ enforces the volume

constraint. The weight $\mathcal{W}$ is computed multiplying the element density ρ_E by the area of the polygonal elements of $\mathcal{T}_h$, while $\mathcal{W}_0$ stands for the volume of the whole design region. A lower bound $\rho_{min} > 0$ is enforced on each density unknown ρ_E to avoid singularity of Eqn. $(8)_2$.

It must be remarked that the simple forms in Eqns. (7) and (8) are due to the adoption of piecewise constant discretization adopted for the density field.

2.3 *Mesh dependence and checkerboard numerical instabilities*

The Method of Moving Asymptotes (MMA) [24], which is an approach of mathematical programming, is herein adopted to iteratively solve the discrete problem in Eqn. (8). The algorithm searches the unknowns over the element–wise density discretization. It is well–known that this discrete scheme is affected by numerical instabilities, such as the arising of checkerboard patterns and mesh dependence, see e.g. [7]. Many strategies are available nowadays to solve both problems [22]. Following [19] and the various applications in stress–based optimal design, see e.g. [13], a density filter approach is herein adopted, instead of applying the filter to the objective function and its sensitivities, as done in most of cases. The original design variables ρ_E are transformed in a new set of physical unknowns $\tilde{\rho}_E$ as:

$$\tilde{\rho}_E = \frac{1}{\sum_{E' \in \mathcal{T}_h} H_{E,E'}} \sum_{E' \in \mathcal{T}_h} H_{E,E'} \rho_{E'}, \qquad H_{E,E'} = \sum_{E' \in \mathcal{T}_h} \max(0, r_{min} - \mathrm{dist}(E, E')). \quad (9)$$

In the above equation $\mathrm{dist}(E, E')$ is the distance between the centroids of the elements E and E', whereas $r_{min} > d_m$ is the filter radius, d_m being the square root of the area of each polygon in $\mathcal{T}_h$. The assumption $r_{min} = 1,5 d_m$ allows to avoid the arising of undesired checkerboard patterns. Larger values of r_{min} provide control on the minimum thickness of any member of the design.

3 NUMERICAL SIMULATIONS

A set of preliminary numerical simulations are presented in this section, adopting the proposed VEM procedure described above. Both structured and unstructured polygonal grids have been employed, see e.g. Figure 1 for an example of grids consisting of 501 elements.. A refer-

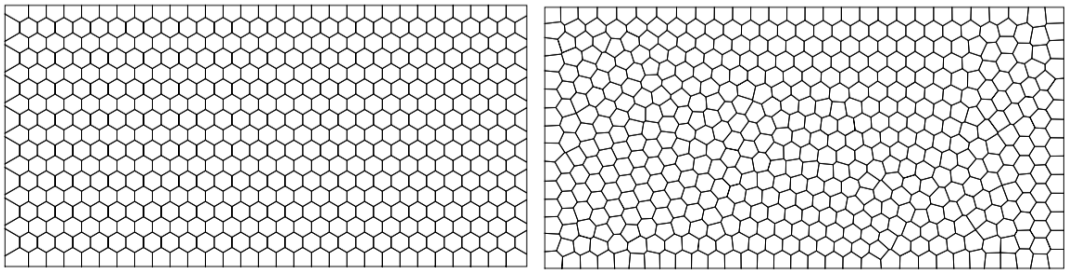


Figure 1: Examples of structured (left) and unstructured (right) polygonal grids consisting of 501 elements.

ence linear elastic isotropic material with Young modulus $E = 1$ and Poisson's ratio $\nu = 0.3$ is adopted. The volume fraction of available material is $V_f = 0.3$ for the whole set of simulations. Moreover, to assess the proposed procedure with respect to well–known benchmark problems we consider different values of the filtering radius.

For the sake of comparison, in the following we compare our results with the analogous ones obtained by employing the classical bilinear displacement–based finite elements on Cartesian meshes, see [7].

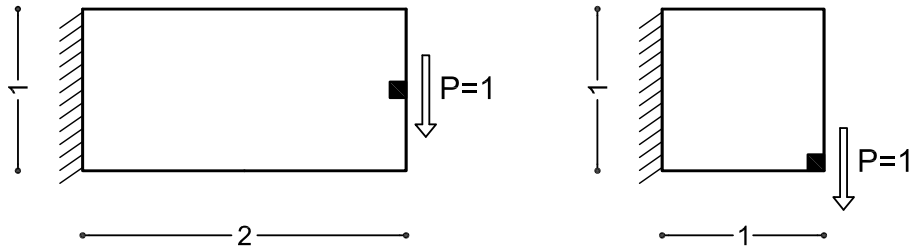


Figure 2: Geometry and boundary conditions for the numerical simulations: Example 1 (left) and Example 2 (right). In each case $P = 1$ is the intensity of the unitary traction $\mathbf{f}_t$ applied as a nodal force oriented as indicated by the arrow.

3.1 Example 1. A rectangular cantilever

A first investigation is performed on the rectangular cantilever depicted in Figure 2(left). The reference solution is shown in Figure 3(b) and has been obtained with bilinear displacement–based finite elements on a Cartesian grid consisting of 8192 squares (2^6 elements lie along the thickness of the cantilever). The filter radius is $r_{min} = 3.0d_m$. As expected, a truss–like structure arises where inclined members carry shear forces, whereas horizontal ones cope with bending actions. Figure 3(a) shows the optimal design obtained by means of the VEM on a structured mesh consisting of 7990 polygonal elements (2^6 elements lie along the thickness of the cantilever) and employing the same filter radius $r_{min} = 3.0d_m$. The achieved result is approximately the same found by the classical approach.

An additional set of simulations is performed adopting 2^5 elements along the thickness of the cantilever, but preserving the same filter radius implemented in the previous investigations. Figure 4(a) shows the optimal design obtained with VEM on a structured mesh consisting of 2006 polygonal elements, whereas Figure 4(b) shows the optimal design achieved through the bilinear displacement–based approach for a regular mesh of 2048 square elements. Although the main layout of Figure 3 is recovered in both pictures, the displacement–based design provides an unexpected variation in the inclination of the thinner braces of the arising truss–like structure. The low available amount of material ($V_f = 0.3$) and the rough mesh of square elements make the optimizer find a 45–degree inclination, which means a local optimum instead of the expected global one. On the other hand, the VEM succeeds in finding the expected layout even in case of unstructured meshes, see the result shown in Figure 5 obtained on a mesh consisting of 2048 elements. This simulation employs the same filter radius r_{min} as above.

Finally, Figure 6 shows a comparison between the VEM approach (a) and the displacement–based one (b) for regular meshes of 32028 and 32768 elements, respectively (2^7 elements lie along the thickness of the cantilever). Despite the adopted smaller filter radius $r_{min} = 1.5d_m$, the achieved results are almost identical.

3.2 Example 2. A square cantilever

In the second example we consider the square cantilever depicted in Figure 2(right). Our VEM-based topology optimization method has been implemented employing unstructured grids of polygonal elements as those shown in Figure 1

First, an unstructured meshes with 2^6 elements along the thickness of the cantilever is used. The overall number of elements is 4096 and the filter radius is $r_{min} = 1.5d_m$ where, as before, d_m is the square root of the average area of the polygonal elements in the unstructured grid.

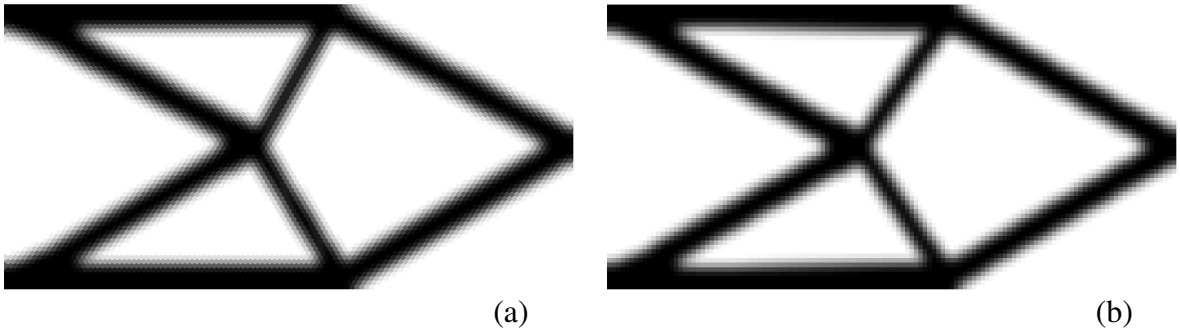


Figure 3: Example 1. Optimal topologies computed on on structured meshes with 2^6 elements along the thickness of the cantilever: proposed VEM sed formulation (a), bilinear displacement–based formulation (b).

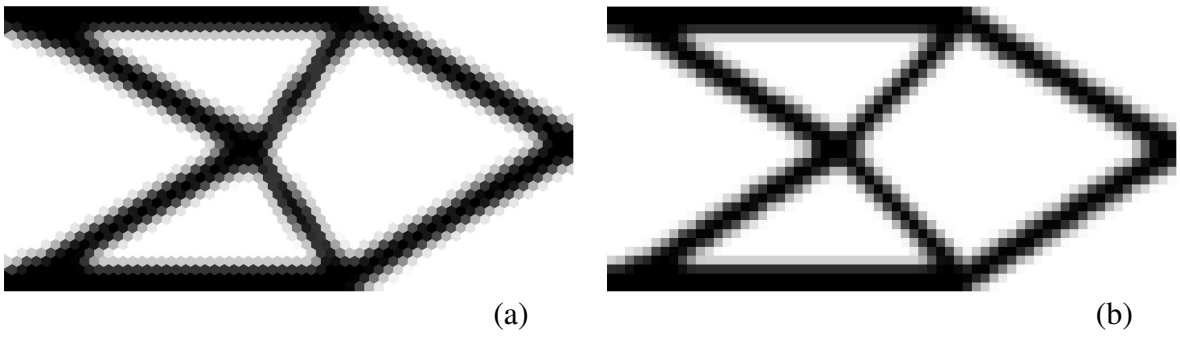


Figure 4: Example 1. Optimal topologies computed on structured meshes with 2^5 elements along the thickness of the cantilever: proposed VEM formulation (a), bilinear displacement–based formulation (b).

The computed result is shown in Figure 7. As expected, a truss–like structure arises where the central node receives two major tensile–stressed bars and two minor compressive–stressed trusses. Figure 8 shows the computed optimal configuration for a stricter prescription on the filter radius, i.e. $r_{min} = 3.0d_m$. The minimum member thickness enforced through the updated filter setup provides a simpler design with two ties and one big strut, in full agreement with the well–known solution of this benchmark problem, see [7]. Finally, in Figure 9 we show the optimal design computed when a finer mesh with 2^7 elements along the thickness of the cantilever is employed. The overall number of polygonal elements is 16384 and the filter radius is $r_{min} = 6.0d_m$, that is approximately the same value used for the result presented in Figure 8. The same result is obtained, suggesting that no mesh dependence affects the proposed VEM formulation.

4 CONCLUSIONS

A set of preliminary numerical simulations has been presented to investigate the performance of the Virtual Element Method in driving a topology optimization algorithm. A classical volume–constrained minimum compliance problem has been tested, adopting an element–wise constant discretization for the unknown density field. Both structured and unstructured grids made of polygonal elements have been employed to solve two benchmark problems. For the sake of comparison, we have also presented the corresponding results obtained with bilinear displacement–based finite elements on Cartesian grids. Employing a sufficiently fine mesh, the same optimal layout is recovered by both methods. However, in case of a coarse mesh and

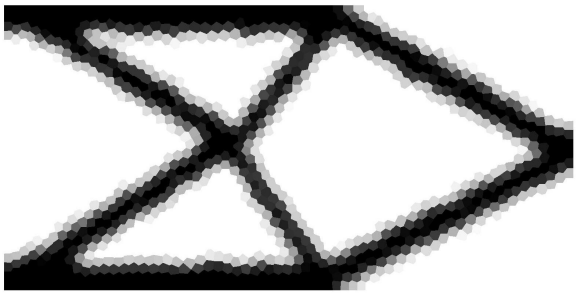
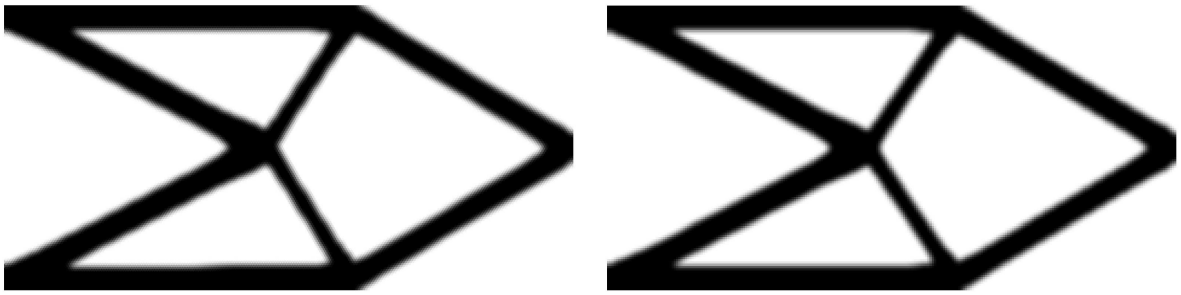


Figure 5: Example 1. Optimal topology achieved through the proposed VEM formulation for an unstructured mesh with 2^5 elements along the thickness of the cantilever.



(a) (b)

Figure 6: Example 1. Optimal topologies computed on structured meshes with 2^7 elements along the thickness of the cantilever: proposed VEM formulation (a), bilinear displacement–based formulation (b).

low fraction of available material, the bilinear displacement–based approach is prone to convergence to a local optimum. Indeed, the orientation of the members is highly affected by the geometrical features of the underlying mesh of square elements. On the contrary, employing either structured or unstructured grids polygonal grids with approximately the same number of elements this issue is overcome and the expected global optimum is obtained (see Example 1). Additional numerical simulations assess that the proposed VEM seems to be fully mesh independent. Indeed, by using classical filtering techniques, the same layout is obtained independently of the number of mesh elements (see Example 2). The ongoing research is mainly concerned with investigations focusing on employing the VEM for the optimal design of incompressible materials thanks to the the full stability properties of the scheme for $\nu \to 0.5$.

ACKNOWLEDGEMENTS

Paola F. Antonietti has been partially supported by SIR Project n. RBSI14VT0S *"PolyPDEs: Non-conforming polyhedral finite element methods for the approximation of partial differential equations"* funded by MIUR. Marco Verani has been partially supported by GNCS-INDAM and by the Italian research grant Prin 2012 n. 2012HBLYE4 *"Metodologie innovative nella modellistica differenziale numerica"*.

REFERENCES

[1] P. F. Antonietti, L. Beirão da Veiga, D. Mora, and M. Verani, *A stream virtual element*

Figure 7: Example 2. Optimal topology computed with the proposed VEM formulation on an unstructured mesh with 2^6 elements along the thickness of the cantilever and filter radius $r_{min} = 1.5d_m \approx 1.5/2^6$ (left) and $r_{min} = 3.0d_m \approx 1.5/2^5$ (right).

Figure 8: Example 2. Optimal topology computed with the proposed VEM formulation on an unstructured mesh with 2^6 elements along the thickness of the cantilever and filter radius $r_{min} = 3.0d_m \approx 1.5/2^5$.

formulation of the Stokes problem on polygonal meshes, SIAM Journal on Numerical Analysis, **52** (2014), 386–404.

[2] P. F. Antonietti, L. Beirão da Veiga, S. Scacchi, and M. Verani, *A C^1 virtual element method for the Cahn-Hilliard equation with polygonal meshes*, SIAM Journal on Numerical Analysis, doi:10.1137/15M1008117.

[3] L. Beirao da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo, *Basic principles of virtual element methods*, Mathematical Models and Methods in Applied Sciences, **23** (2103), 199–214.

[4] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo, *The hitchhiker's guide to the virtual element method*, Mathematical Models and Methods in Applied Sciences, **24** (2014), 1541–1573.

[5] L. Beirao da Veiga, F. Brezzi, and L. D. Marini, *Virtual elements for linear elasticity problems* SIAM Journal on Numerical Analysis, **51** (2013), 794–812.

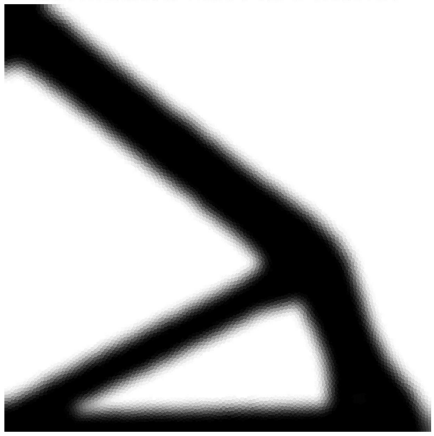


Figure 9: Example 2. Optimal topology computed with the proposed VEM formulation on an unstructured mesh with 2^7 elements along the thickness of the cantilever and filter radius $r_{min} = 6.0d_m \approx 1.5/2^5$.

[6] L. Beirão da Veiga and G. Manzini, *A virtual element method with arbitrary regularity*, IMA Journal of Numerical Analysis, **34** (2014), 759–781.

[7] M.P. Bendsøe, O. Sigmund, *Topology optimization theory, methods and applications*, New York, Springer, 2003.

[8] M.P. Bendsøe, *Optimal shape design as a material distribution problem*, Structural Optimization, **1** (1989), 193–202.

[9] M. F. Benedetto, S. Berrone, S. Pieraccini, and S. Scialò, *The virtual element method for discrete fracture network simulations*, Computer Methods in Applied Mechanics and Engineering, **280** (2014), 135–156.

[10] M. F. Benedetto, S. Berrone and S. Scialò, *A globally conforming method for solving flow in discrete fracture networks using the Virtual Element Method*, Finite Elements In Analysis and Design, **109** (2016), 23–36.

[11] F. Brezzi, R.S. Falk, and L.D. Marini, *Basic principles of mixed Virtual Element Methods*, ESAIM. Mathematical Modelling and Numerical Analysis, **48** (2014), 1227–1240.

[12] F. Brezzi and L. D. Marini, *Virtual element methods for plate bending problems*, Computer Methods in Applied Mechanics and Engineering, **253** (2013), 455–462.

[13] M. Bruggi, P. Duysinx, *A stress-based approach to the optimal design of structures with unilateral behavior of material or supports*, Structural and Multidisciplinary Optimization, **48** (2013), 311–326.

[14] M. Bruggi, P. Venini, *Eigenvalue–based optimization of incompressible media using mixed finite elements with application to isolation devices*, Computer Methods in Applied Mechanics and Engineering, **197** (2008), 1262–1279.

[15] M. Bruggi, M. Verani, *A fully adaptive topology optimization algorithm with goal–oriented error control*, Computers and Structures, **89** (2011), 1481–1493.

[16] A. Cangiani, G. Manzini, A. Russo, and N. Sukumar, *Hourglass stabilization and the virtual element method*, International Journal for Numerical Methods in Engineering, **102** (2015), 404–436.

[17] A.L. Gain, G.H. Paulino, L.S. Duarte, and I.F.M. Menezes, *Topology optimization using polytopes virtual elements for linear elasticity problems*, Computer Methods in Applied Mechanics and Engineering, **293** (2015), 411–430.

[18] A.L. Gain, C. Talischi, G.H. Paulino, *On the Virtual Element Method for three-dimensional linear elasticity problems on arbitrary polyhedral meshes*, Computer Methods in Applied Mechanics and Engineering, **282** (2014), 132–160.

[19] C. Le, J. Norato, T.E. Bruns, C. Ha, D.D. Tortorelli, *Stress–based topology optimization for continua*,Structural and Multidisciplinary Optimization, **41** (2010), 605–620.

[20] G. Manzini, A. Russo, and N. Sukumar, *New perspectives on polygonal and polyhedral finite element methods*, Mathematical Models and Methods in Applied Sciences, **24** (2014), 1665–1699.

[21] D. Mora, G. Rivera, and R. Rodríguez, *A virtual element method for the Steklov eigenvalue problem*, Mathematical Models and Methods in Applied Sciences **25** (2015), 1421–1445.

[22] O. Sigmund O, J. Petersson, *Numerical instabilities in topology optimization: a survey on procedures dealing with checkerboards, mesh-dependencies and local minima*, Structural Optimization, **16** (1998), 68–75.

[23] O. Sigmund, P.M. Clausen, *Topology optimization using a mixed formulation: an alternative way to solve pressure load problems*, Computer Methods in Applied Mechanics and Engineering, **196** (2017), 1874–1889.

[24] K. Svanberg, *Method of moving asymptotes - a new method for structural optimization*, International Journal for Numerical Methods in Engineering, **24** (1897), 193–202.

[25] Y. Wang, Z. Kang, Q. He, *Adaptive topology optimization with independent error control for separated displacement and density fields*,Computers and Structures, **135** (2014), 50–61.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THE VIRTUAL ELEMENT METHOD FOR DISCRETE FRACTURE NETWORK FLOW AND TRANSPORT SIMULATIONS

Matías Fernando Benedetto, Stefano Berrone, Andrea Borio, Sandra Pieraccini, and Stefano Scialo

Politecnico di Torino
Corso Duca degli Abruzzi, 24 – 10129, Torino (TO), Italy
e-mail: {matias.benedetto,stefano.berrone,andrea.borio,sandra.pieraccini,stefano.scialo}@polito.it

Keywords: Discrete Fracture Network, Virtual Element Method, Underground flow, Fractured media.

Abstract. *We discuss several issues concerning the application of the Virtual Element Method (VEM) to the flow in fractured media modeled by the Discrete Fracture Network (DFN) model. Due to the stochastic nature of the computational domains, several geometrical complexities make the computations very challenging. The geometrical flexibility provided by the Virtual Element Method can be exploited to mutually couple local problems, either by resorting to a Mortar approach, or by allowing for the global conformity of the local meshes, while keeping the computational cost under control. We describe these two approaches in detail and we test them on a realistic test case, showing the viability of the two approaches.*

1 Introduction

Subsurface fluid flow has applications in a wide range of fields, including e.g. oil/gas recovery, gas storage, pollutant percolation, water resources monitoring. Underground fluid flow in fractured media is a heterogeneous multi-scale phenomenon that involves complex geological configurations; a possible approach for modeling the phenomenon is given by Discrete Fracture Networks (DFNs), which are complex sets of polygonal intersecting fractures randomly generated from known distributions for geometrical features (such as orientation in the three dimensional space, position, dimensions) and hydro-geological properties. Geological fractured media are therefore characterized by a very challenging geometrical complexity, which is one of the major difficulties to be tackled when performing flow simulations.

In this work we recall some results concerning the application of the Virtual Element Method [4, 3, 2] to the steady state simulation of the flow in DFNs [1, 22, 27, 30, 24, 32, 23, 15, 16, 17, 18, 19, 9, 10, 8, 13]. In this approach we can exploit the flexibility of VEM in order to tackle the geometrical complexity. Indeed, a crucial issue in DFN flow simulations is the need to provide on each fracture a good quality mesh [24, 20, 29, 30] on any randomly generated configuration. Namely, if classical triangular or quadrilateral meshes on the fractures are required to be conforming to the traces (fracture intersections), and also conforming each other, the meshing process for each fracture is not independent of the others, thus yielding in practice a quite demanding computational effort for the mesh generation process. In some cases, the meshing process may even result infeasible so that some authors propose to modify the DFN removing problematic fractures [24].

Here, the VEM will be used within several possible approaches to the problem: in conjunction with a totally conforming polygonal mesh [10] and with a Mortar approach [8]. Indeed, taking advantage from the great flexibility of VEM in allowing the use of rather general polygonal mesh elements, a suitable mesh for representing the solution and imposing matching conditions between the solutions on different fractures can be easily obtained, starting from an arbitrary triangular mesh independently built on each fracture, and independent of the trace disposition. Robustness and efficiency of the approach are of great importance also in the framework of Uncertainty Quantification analysis applied to DFNs, see [14].

The paper is organized as follows: after introducing some notation about the DFNs in Section 2, we describe the more common geometrical complexities present in DFNs flow simulations in Subsection 2.1. The problem considered and the formulation used are described in Section 3. Section 4 is devoted to discuss the VEM formulation and some issues concerning its implementation. Subsection 4.3 recalls the elements of a hybrid mortar approach to the matching conditions, whereas Subsection 4.4 recalls the basic ideas of a globally conforming approach. Finally, we propose in Section 5 some numerical results obtained with the two approaches.

2 Discrete Fracture Networks

A DFN is a possible model for a fractured medium in which the surrounding rock matrix can be assumed to be impervious. In a DFN, fractures in the underground medium are represented as bi-dimensional open polygons. The intersections between fractures are called *traces*, and we assume, for the sake of simplicity, that precisely two fractures meet at each trace. For a DFN Ω, we will indicate by F_i the generic fracture, with $i \in \mathcal{I} = \{1, \dots, N\}$, while its boundary will be $\partial\Omega = \cup_{i\in\mathcal{I}}\partial F_i$. Traces will be denoted by Γ_m, with $m \in \mathcal{M} = \{1, \dots, M\}$. Without loss of generality, we assume that the set $\bar{\Omega}$ is connected. Finally, we introduce the following notation:

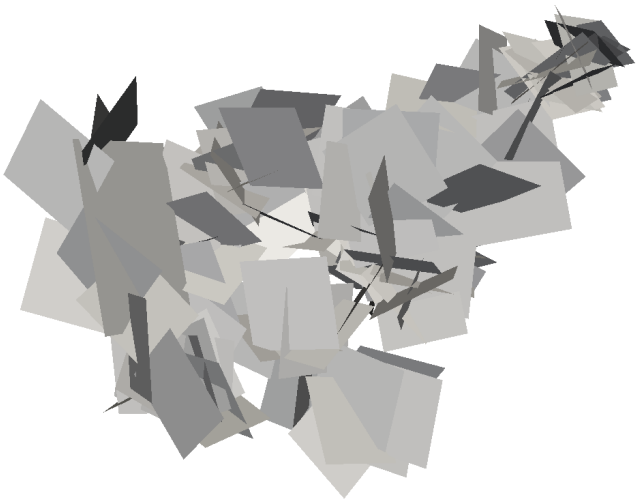
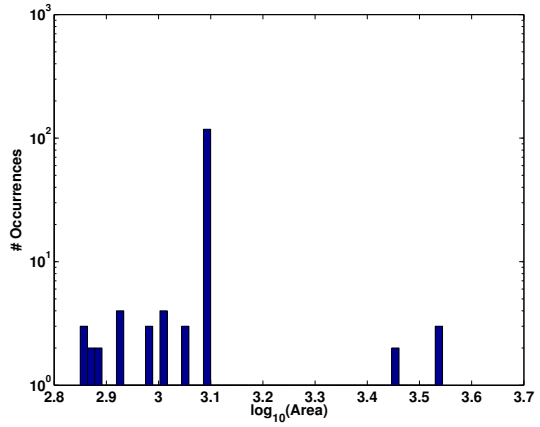


Figure 1: DFN with 134 fractures

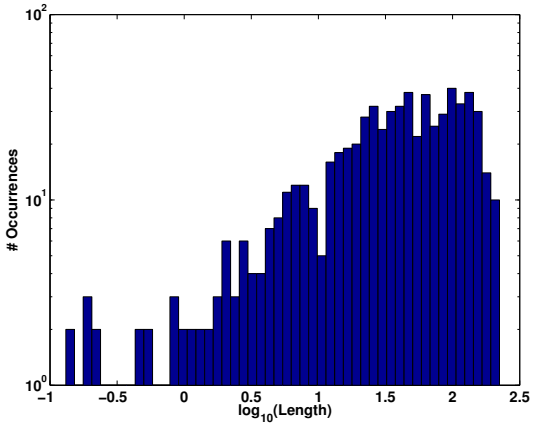
- $\forall i \in \mathcal{I}$, let $\mathcal{M}_i \subset \mathcal{M}$ be the subset of trace indices corresponding to traces lying on F_i; each subset $\mathcal{M}_i$ is assumed to be ordered, and we will denote by $\mathcal{M}_i(k)$ the k-th index of a trace in $\mathcal{M}_i$;
- $\forall m \in \mathcal{M}$, let $\mathcal{I}_m = (i, j)$ be the ordered couple of indices such that $\Gamma_m = F_i \cap F_j$, with $i < j$;
- for each $i \in \mathcal{I}$ and each $m \in \mathcal{M}_i$, we fix a unit vector $\hat{n}^i_m$ normal to Γ_m on F_i.

2.1 Geometrical issues

When dealing with DFN flow simulations, the greatest obstacle is to devise a robust and efficient meshing process, while enforcing some kind of conformity of the mesh polygons to the traces. This can be required locally on each fracture, or globally, asking that polygons on different fractures meeting on the same trace share either one point or a whole side. These constraints can make the meshing process infeasible using traditional simplicial elements, because traces may intersect with very small angles, may have very different lengths or can be very close to each other without intersections. To illustrate how common are problematic configurations, let us consider a quite simple, although realistic, DFN containing 134 fractures; the DFN is displayed in Figure 1. Even though fractures in this DFN are of comparable size (see Figure 2a) the trace lengths span several orders of magnitude (Figure 2b), thus creating problems when mesh edges are required either to lie entirely on traces or have an intersection with them with a null measure. Furthermore, in Figure 3a we show the global distribution of angles between intersecting traces, while Figure 3b shows that a considerable amount of them is very small, making it very hard to build good quality triangular elements close to the corresponding intersections. Finally, another geometrical issue which is worth to analyze is the presence of very close and non – intersecting traces. In Figure 4 we see that there is a significant amount of them, which would yield very small elements, in case a regular triangulation is built.

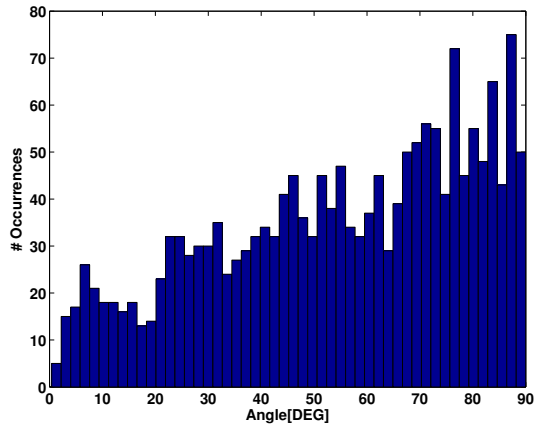


(a) Distribution of fracture areas.

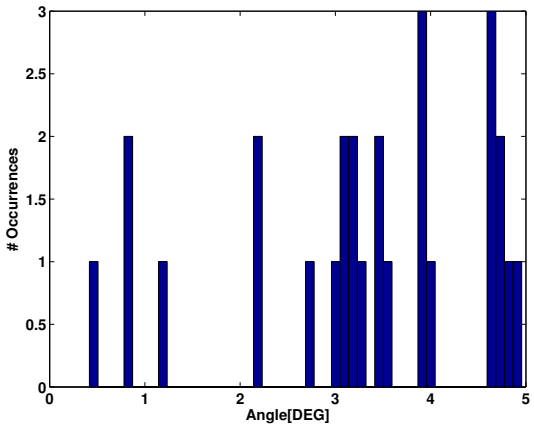


(b) Distribution of trace lengths.

Figure 2: Distribution of fracture areas and traces lengths of the DFN in Figure 1.

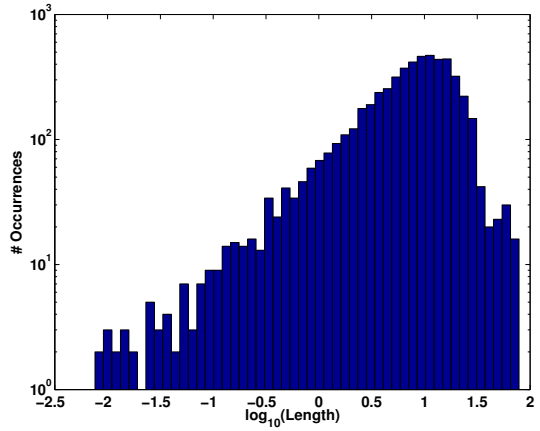


(a) Full distribution.

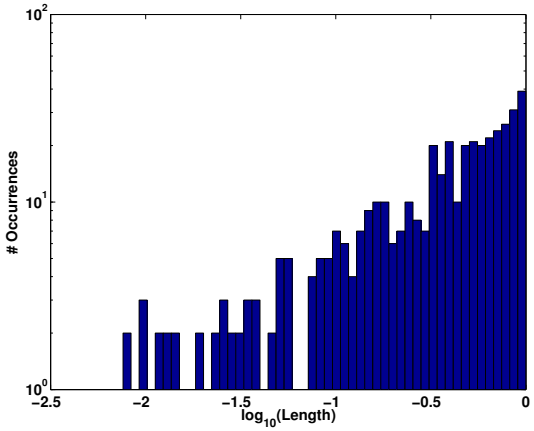


(b) Occurrences of very small angles.

Figure 3: Distribution of angles between traces for the DFN in Figure 1.



(a) Full distribution.



(b) Occurrences of very small distances.

Figure 4: Distribution of distance between non intersecting traces for the DFN in Figure 1

3 Problem formulation

Given an open bounded domain ω, let $(\cdot,\cdot)_\omega$ and $\|\cdot\|_\omega$ denote the $\mathrm{L}^2(\omega)$ scalar product and norm, respectively, and $(\cdot,\cdot)_{\alpha,\omega}$ and $\|\cdot\|_{\alpha,\omega}$ denote the $\mathrm{H}^\alpha(\omega)$ scalar product and norm, respectively. In general, a subscript i will denote the restriction of a function to the fracture F_i.

For any segment $\sigma \subset F_i$, $i \in \mathcal{I}$, we introduce the trace operator $\gamma_\sigma \colon \mathrm{H}^1(F_i) \to \mathrm{H}^{\frac{1}{2}}(\sigma)$ and the notation

$$\langle \mu, \beta \rangle_\sigma := {}_{\mathrm{H}^{-\frac{1}{2}}(\sigma)}\langle \mu, \beta \rangle_{\mathrm{H}^{\frac{1}{2}}(\sigma)}, \quad \forall \mu \in \mathrm{H}^{-\frac{1}{2}}(\sigma), \beta \in \mathrm{H}^{\frac{1}{2}}(\sigma),$$

to denote the duality product between $\mathrm{H}^{-\frac{1}{2}}(\sigma)$ and $\mathrm{H}^{\frac{1}{2}}(\sigma)$. Let $v \in \mathrm{H}^1(F_i)$, in order to simplify the notation, it is convenient to introduce the vectors $\gamma_{\mathcal{M}_i}(v)$, $\forall i \in \mathcal{I}$, the k-th element of $\gamma_{\mathcal{M}_i}(v)$ being $\gamma_{\Gamma_{\mathcal{M}_i(k)}}(v)$. Furthermore, we introduce the jump across a trace Γ_m as

$$[\![v]\!]_{\Gamma_m} := \gamma_{\Gamma_m}(v_i) - \gamma_{\Gamma_m}(v_j), \quad \text{if } \mathcal{I}_m = (i,j),$$

and we introduce the symbols $[\![v]\!]_{\mathcal{M}}$ and $[\![v]\!]_{\mathcal{M}_i}$ to denote the vectors of jumps of v across all traces in the network, and across traces on F_i, respectively. With the same purpose, $\forall i \in \mathcal{I}$, we introduce the notation

$$\langle \mu, \beta \rangle_{\mathcal{M}_i} := \sum_{m \in \mathcal{M}_i} \langle \mu_m, \beta_m \rangle_{\Gamma_m}, \ \forall \mu \in \prod_{m \in \mathcal{M}_i} \mathrm{H}^{-\frac{1}{2}}(\Gamma_m), \beta \in \prod_{m \in \mathcal{M}_i} \mathrm{H}^{\frac{1}{2}}(\Gamma_m).$$

Finally, for any $\beta, \lambda \in \prod_{m \in \mathcal{M}_i} \mathrm{H}^{\frac{1}{2}}(\Gamma_m)$ we denote

$$(\beta, \lambda)_{\mathcal{M}_i} := \sum_{m \in \mathcal{M}_i} (\beta, \lambda)_{\Gamma_m}.$$

We are interested in computing the hydraulic head $h = \pi/(\rho g) + z$, where π is the fluid pressure, g the gravitational acceleration, ρ the fluid density and z the elevation. The hydraulic head, on each fracture F_i, is modeled by means of the Darcy law as follows.

Let K_i denote the transmissivity on F_i, which we assume to be constant, and $f_i = f_i(x)$ denote the source term on F_i; notice that both K_i and f_i are functions of the local tangential coordinate system. Let $\Gamma^D \subseteq \partial\Omega$ be the Dirichlet boundary, and let h^D be the Dirichlet boundary condition defined on Γ^D. We define the functional spaces:

$$\begin{aligned}
V_i &:= \left\{ v \in \mathrm{H}^1(F_i) : \gamma_{\Gamma^D}(v) = 0 \right\} && \forall i \in \mathcal{I}, \\
V_i^D &:= \left\{ v \in \mathrm{H}^1(F_i) : \gamma_{\Gamma^D}(v) = h_i^D \right\} && \forall i \in \mathcal{I}, \\
V &:= \left\{ v \colon v_i \in V_i \quad \forall i \in \mathcal{I} \right\}, \\
V^D &:= \left\{ v \colon v_i \in V_i^D \quad \forall i \in \mathcal{I} \right\}.
\end{aligned}$$

On each fracture F_i, $i \in \mathcal{I}$, we want to find $h_i \in V_i^D$ such that, $\forall v_i \in V_i$,

$$a_i(h_i, v_i) := (\mathrm{K}_i \nabla h_i, \nabla v_i)_{F_i} = (f_i, v_i)_{F_i} + \left\langle h_i^N, \gamma^i_{\Gamma_i^N}(v_i) \right\rangle_{\Gamma_i^N} + \left\langle \left[\!\left[\frac{\partial h_i}{\partial \hat{n}_{\mathcal{M}_i}} \right]\!\right]_{\mathcal{M}_i}, \gamma_{\mathcal{M}_i}(v_i) \right\rangle_{\mathcal{M}_i} \quad (1)$$

where $\left[\!\left[\frac{\partial h_i}{\partial \hat{n}^i_m} \right]\!\right]_{\Gamma_m}$ is the jump of the co-normal derivative $\frac{\partial h_i}{\partial \hat{n}^i_m} = \mathrm{K}_i \nabla h_i \cdot \hat{n}^i_m$ along $\hat{n}^i_m$; furthermore, $\Gamma_i^N \subseteq \partial F_i$ is the Neumann boundary on F_i and $h_i^N \in \mathrm{H}^{-\frac{1}{2}}\left(\Gamma_i^N\right)$ is the Neumann

boundary condition. For future reference, we set $\Gamma^N = \cup_{i\in\mathcal{I}}\Gamma_i^N \subset \partial\Omega$ and define h^N such that h_i^N is the restriction of h^N to F_i. We couple the problems on each fracture by imposing the continuity of the solution and balance of incoming and outgoing fluxes at each trace: $\forall m \in \mathcal{M}$, with $\mathcal{I}_m = (i,j)$, we have

$$\begin{aligned}\left[\!\left[h\right]\!\right]_{\Gamma_m} &= \gamma_{\Gamma_m}(h_i) - \gamma_{\Gamma_m}(h_j) = 0,\\ \left[\!\left[\frac{\partial h_i}{\partial \hat{n}_m^i}\right]\!\right]_{\Gamma_m} &+ \left[\!\left[\frac{\partial h_j}{\partial \hat{n}_m^j}\right]\!\right]_{\Gamma_m} = 0.\end{aligned} \tag{2}$$

3.1 Saddle point formulation of the DFN problem

The DFN problem (1)-(2) can be easily re-formulated as a saddle point problem. To this aim, let us first define, for each trace $m \in \mathcal{M}$, and recalling that $\mathcal{I}_m = (i,j)$, the function s_{Γ_m} such that:

$$s_{\Gamma_m}(i) = 1, \quad s_{\Gamma_m}(j) = 0,$$

and, for each fracture, a bilinear form $b_i\colon M_i := \prod_{m\in\mathcal{M}_i} \mathrm{H}^{-\frac{1}{2}}(\Gamma_m) \times V_i \to \mathbb{R}$ defined by

$$b_i(v,\psi) := \sum_{m\in\mathcal{M}_i} (-1)^{s_{\Gamma_m}(i)} \left\langle \psi_m, \gamma_{\Gamma_m}(v_i)\right\rangle_{\Gamma_m}.$$

Moreover, let us define

$$b(v,\psi) := \sum_{i\in\mathcal{I}} b_i(v,\psi) = \sum_{m\in\mathcal{M}} \left\langle \psi_m, \left[\!\left[v\right]\!\right]_{\Gamma_m}\right\rangle_{\Gamma_m}.$$

These bilinear forms are used in the definition of the functional $\mathcal{F}\colon V\times M := \prod_{m\in\mathcal{M}} \mathrm{H}^{-\frac{1}{2}}(\Gamma_m)$, given as a sum of contributions from the different fractures:

$$\begin{aligned}\mathcal{F}(v,\mu) := \sum_{i\in\mathcal{I}} \frac{1}{2}(\mathrm{K}_i\nabla v_i, \nabla v_i)_{F_i} - (f_i, v_i)_{F_i} - \left\langle h_i^N, \gamma_{\Gamma_i^N}^i(v_i)\right\rangle_{\Gamma_i^N}\\ + \left(\nabla\mathcal{R}_i\left(h_i^D\right), \nabla v_i\right)_{F_i} + b_i(v,\mu) + b_i\left(\mathcal{R}_i\left(h_i^D\right),\mu\right),\end{aligned}$$

where $\mathcal{R}_i$ is the lifting operator from $\mathrm{H}^{\frac{1}{2}}(F_i)$ to $\mathrm{H}^1(F_i)$, $i \in \mathcal{I}$.

The solution $(h,\lambda) \in V^D \times M$ such that $h_i = h_i^0 + \mathcal{R}_i\left(h^D\right)$, $h_i^0 \in V_i$, $i \in \mathcal{I}$ and

$$\mathcal{F}\left(h^0,\lambda\right) = \min_{v\in V}\max_{\psi\in M} \mathcal{F}(v,\psi), \tag{3}$$

is equivalent to the unique solution to the problem: $h = h^0 + \mathcal{R}\left(h^D\right)$, with $h^0 \in V$ and $\lambda \in M$ such that,

$$\begin{cases} a(h^0,v) + b(v,\lambda) = (f,v) + \left(h^N,v\right)_{\Gamma^N} - a(\mathcal{R}\left(h^D\right),v) & \forall v \in V,\\ b(h^0,\psi) = -b\left(\mathcal{R}\left(h^D\right),\psi\right) & \forall \psi \in M,\end{cases} \tag{4}$$

that provides the solution to the problem (1)-(2), as it can be proven following classical arguments (see e.g. [31]). Moreover, we have

$$\lambda_m = \left[\!\left[\frac{\partial h_i}{\partial \hat{n}_m^i}\right]\!\right]_{\Gamma_m} = -\left[\!\left[\frac{\partial h_j}{\partial \hat{n}_m^j}\right]\!\right]_{\Gamma_m},$$

with $\mathcal{I}_m = (i,j)$.

4 The discrete DFN problem

First introduced in [4] and extended in [5, 6, 3, 21, 2, 26], the Virtual Element Method allows the use of quite general non-degenerate and star-shaped polygons to mesh the spatial domain, even including the possibility of straight angles. In the present framework, we take advantage from this flexibility to easily build a mesh which, on each fracture, is locally or globally conforming to the traces. In the following of this section, we review the use of VEM, focusing on the framework of DFN simuations.

4.1 Construction of the mesh

Let a fracture F_i be fixed. To obtain a locally conforming mesh, we first introduce on F_i a triangular mesh built independently of trace positions; the triangles are then cut into polygons by the traces, possibly prolonging the trace segment up to the nearest mesh edge if it happens to end in the interior of a triangle. Note that in this latter case the trace tip is kept as a node of the discretization and the trace is not modified, a new node is created at the intersection between the prolongation of the trace segment and the mesh edge, and therefore two edges are created, with a $180°$ angle between them. Let $\mathcal{T}_{\delta i}^{\text{lc}}$ be the resulting local mesh. We refer to Figure 5 for a possible mesh configuration and to Figure 6a for a locally conforming mesh on a fracture. Let $\mathcal{T}_{\delta}^{\text{lc}} = \cup_{i\in\mathcal{I}}\mathcal{T}_{\delta i}^{\text{lc}}$. We will use the symbols $\mathcal{E}_{\delta i}^{\text{lc}}$ and $\mathcal{V}_{\delta i}^{\text{lc}}$ to denote the sets of edges and vertices on fracture F_i, respectively, and define the sets of the mesh edges and vertices of the whole DFN as $\mathcal{E}_{\delta}^{\text{lc}} = \cup_{i\in\mathcal{I}}\mathcal{E}_{\delta i}^{\text{lc}}$, $\mathcal{V}_{\delta}^{\text{lc}} = \cup_{i\in\mathcal{I}}\mathcal{V}_{\delta i}^{\text{lc}}$, respectively. The mesh built in this way can be used to couple VEM discretizations on each fracture with a Mortar approach, as described in subsection 4.3.

Another possible meshing process, that aims at building a globally conforming mesh $\mathcal{T}_{\delta}^{\text{gc}}$, can be devised as follows, starting from the above described mesh $\mathcal{T}_{\delta}^{\text{lc}}$. Let us consider an arbitrary trace Γ_m, with $m \in \mathcal{M}$ and $\mathcal{I}_m = (i, j)$. Then, we add to $\mathcal{T}_{\delta i}^{\text{lc}}$ the nodes generated by $\mathcal{T}_{\delta j}^{\text{lc}}$ on Γ_m, and viceversa. Some polygons belonging to mesh $\mathcal{T}_{\delta i}^{\text{lc}}$ ($\mathcal{T}_{\delta j}^{\text{lc}}$, respectively) having an edge lying on Γ_m, will possibly have such edges split by the new nodes, the new edges forming a straight angle at their intersection. Again, we refer the reader to Figure 5 for a visualization of a rather intricate trace configuration and to Figure 6b for a resulting globally conforming mesh on a fracture. All sets of geometrical objects relative to this globally conforming mesh will have a superscript "gc" in the following. This spatial discretization will be used in subsection 4.4 to allow the construction of VEM spaces containing globally continuous functions.

To further illustrate this process, we show in Figure 6 a possible situation in which the number of nodes on traces in $\mathcal{T}_{\delta}^{\text{gc}}$ is larger than the number of nodes in $\mathcal{T}_{\delta}^{\text{lc}}$. In the two subfigures, traces are drawn in red: as we can see the globally conforming mesh in Figure 6b presents some nodes that do not correspond to intersections between polygons and traces, but are induced by polygons on another fracture that generate the same trace. These nodes are not treated as hanging nodes, but as vertices of a polygon whose edges meeting there form a flat angle. The locally conforming mesh, instead, uses as nodes only the intersections between local polygons and traces. Finally, notice that in both cases the trace tip is added as a node.

Remark. All the elements created with the above procedures are convex.

4.2 The VEM setting

Let us now briefly sketch the main ideas at the basis of the use of the VEM in the context of flow simulations in complex networks of fractures, referring to the specific literature (see e.g. [4, 6, 2]) for a deeper insight about the method.

Let $k \in \mathbb{N}$ be a fixed polynomial degree, corresponding to the desired polynomial accuracy

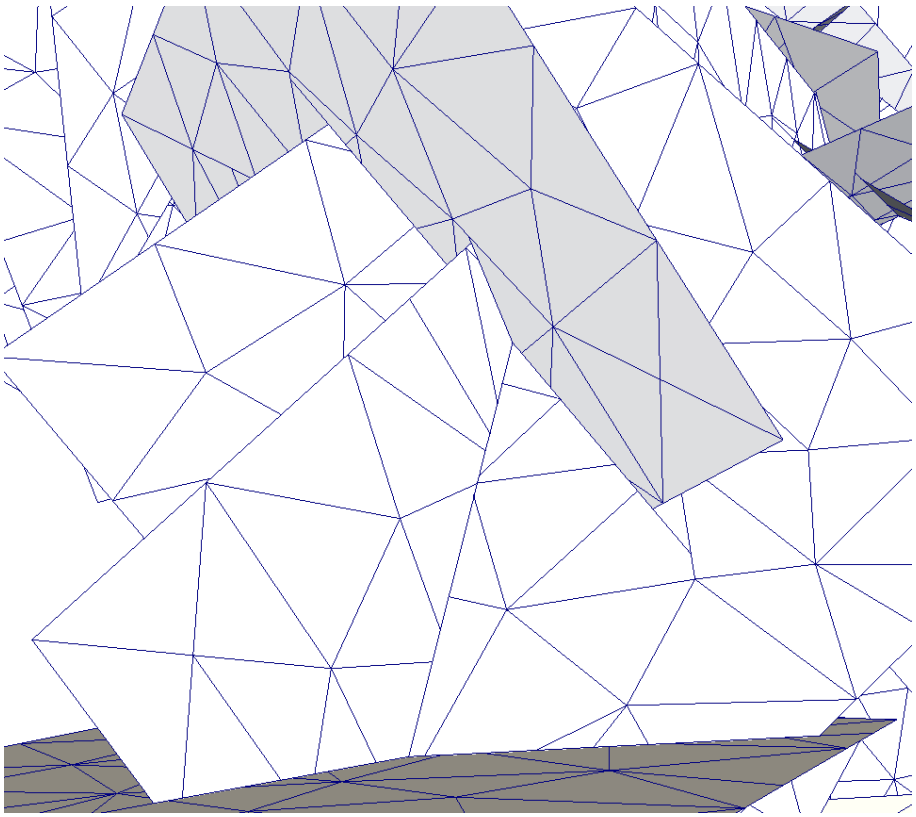
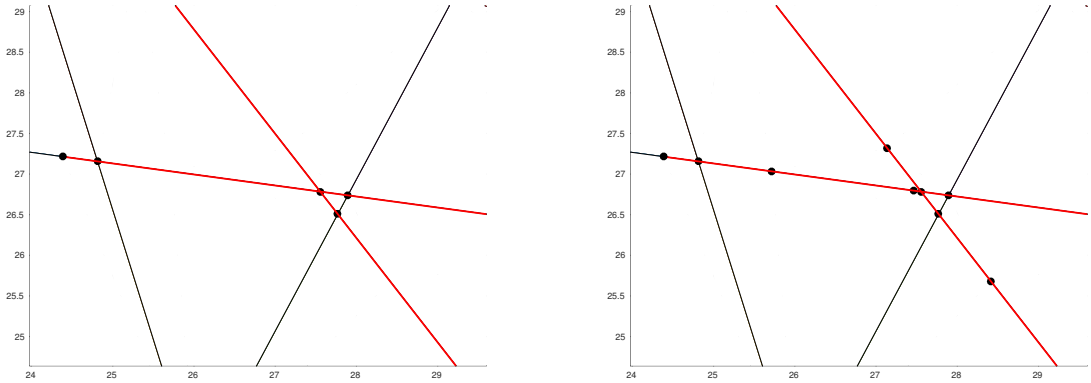


Figure 5: A detail of the mesh around traces on the DFN in Figure 1



(a) Locally conforming mesh $\mathcal{T}_\delta^{\text{lc}}$ (b) Globally conforming mesh $\mathcal{T}_\delta^{\text{gc}}$

Figure 6: Details of the vertices of the discretizations on a trace

for the VEM approximation, and let E be a generic polygonal element of the mesh $\mathcal{T}^\star_{\delta i}$ for $i \in \mathcal{I}$ and $\star =' lc'$ or $'gc'$, as in subsection 4.1. Let us define on E the discrete functional space

$$V^E_{\delta i} := \left\{ v \in \mathrm{H}^1(E) : \gamma_e(v) \in \mathbb{P}_k(e) \ \forall e \subset \partial E,\ \Delta v|_E \in \mathbb{P}_{k-2}(E) \right\}, \tag{5}$$

and, on each fracture F_i the space

$$V^\star_{\delta i} := \left\{ v_\delta \in \mathbb{C}^0(F_i) : v \in V^E_{\delta i} \quad \forall E \in \mathcal{T}^\star_{\delta i} \right\}.$$

Then the discrete subspace of V with the VEM is:

$$V^\star_\delta := \left\{ v_\delta \in V : v \in V^\star_{\delta i} \quad \forall i \in \mathcal{I} \right\}, \tag{6}$$

spanned by basis functions $\phi^\star_\ell$, $\ell = 1, \dots, N_\ell$, being N_ℓ its dimension. The following set of degrees of freedom (DOFs) is introduced to uniquely define a function $v_\delta \in V^\star_{\delta i}$ [4]:

- the values of v_δ at each vertex in $\mathcal{V}^\star_\delta$;

and, if $k > 1$ also

- the values of v_δ at $k-1$ internal points (e.g. internal Gauss-Lobatto quadrature nodes) on each edge $e \in \mathcal{E}^\star_\delta$;
- the moments $\frac{1}{|E|}\int_E v_h m_{\boldsymbol{\alpha}}$ where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \mathbb{N}^2$,

$$m_{\boldsymbol{\alpha}}(x, y) := \left(\frac{x - x_E}{h_E} \right)^{\alpha_1} \left(\frac{y - y_E}{h_E} \right)^{\alpha_2},$$

  $\forall E \in \mathcal{T}^\star_\delta$ and $\forall \boldsymbol{\alpha}$ such that $|\,\boldsymbol{\alpha}\,| \leq k-2$,

where (x_E, y_E) and h_E are the barycenter and the diameter of the element E, respectively. The chosen basis functions for $V^\star_\delta$ are Lagrangian with respect to this set of DOFs. We remark that functions ϕ_ℓ are not explicitly known in the interior of each element.

Assuming that the fracture transmissivity K_i is a constant function on each fracture, we introduce, for each $i \in \mathcal{I}$ and $E \in \mathcal{T}^\star_{\delta i}$, the operator $\Pi^\nabla_E \colon V^E_{\delta i} \to \mathbb{P}_k(E)$ defined as follows:

$$\begin{cases} \left(\mathrm{K}_i \nabla \Pi^\nabla_E \phi, \nabla p\right)_E = \left(\mathrm{K}_i \nabla \phi, \nabla p\right)_E & \forall p \in \mathbb{P}_k(E), \\ \sum_{V \in \mathcal{V}^{\mathrm{E}}_{\delta i}} \Pi^\nabla_E \phi(V) = \sum_{V \in \mathcal{V}^{\mathrm{E}}_{\delta i}} \phi(V) & \text{if } k = 1, \\ \int_E \Pi^\nabla_{E,k} v_\delta = \int_E v_\delta & \text{if } k > 1. \end{cases}$$

that can be computed by means of the above listed degrees of freedom, and the symmetric bilinear form $S^E \colon V^E_{\delta i} \times V^E_{\delta i} \to \mathbb{R}$, such that $\ker S^E \cap \ker \Pi^\nabla_E = \{0\}$ and

$$c_* \left(\mathrm{K}_i \nabla \phi, \nabla \phi\right)_E \leq S^E(\phi, \phi) \leq c^* \left(\mathrm{K}_i \nabla \phi, \nabla \phi\right)_E, \quad \forall \phi \in V^E_{\delta i},\ \text{s.t. } \Pi^\nabla_E \phi = 0 \tag{7}$$

for two positive constants c_* and c^*, independent of E and of the fracture F_i. The discrete counterpart of the bilinear form $a_i : V_i \times V_i \to \mathbb{R}$ in equation (1) is $a_{\delta i} \colon V^\star_{\delta i} \times V^\star_{\delta i} \to \mathbb{R}$, defined by

$$a_{\delta i}(v, w) := \sum_{E \in \mathcal{T}_{\delta i}} a^E_\delta(v, w),$$

where $a_\delta^E \colon V_\delta^\star \times V_\delta^\star \to \mathbb{R}$ is such that $\forall v, w \in V_\delta^\star$,

$$a_\delta^E(v,w) := \left(\mathrm{K}_i \nabla \Pi_E^\nabla v, \nabla \Pi_E^\nabla w\right)_E + S^E\left(v - \Pi_E^\nabla v, w - \Pi_E^\nabla w\right) .$$

Thanks to the definition of S^E and to property (7), the coercivity of the discrete bilinear form can be easily proven and it can be shown that it scales like $(\mathrm{K}_i \nabla v, \nabla v)_{F_i}$, i.e.:

$$\exists \alpha_*, \alpha^* > 0 \colon \alpha_* \left(\mathrm{K}_i \nabla v, \nabla v\right)_{F_i} \le a_{\delta i}(v,v) \le \alpha^* \left(\mathrm{K}_i \nabla v, \nabla v\right)_{F_i} . \tag{8}$$

For the computation of the discrete counterpart of the scalar product (f, v_δ) at the right-hand-side of equation (1), when v_δ is a function of the VEM space not known in the interior of each mesh element, we introduce the following discrete scalar product

$$(f_i, v_{\delta i})_{\delta, F_i} := \left(f_i, \tilde{\Pi}_k^0 v_{\delta i}\right)_{F_i} \quad \forall i \in \mathcal{I} ,$$

where the pseudo-projection $\tilde{\Pi}_k^0 : V_{\delta i} \to \mathbb{P}_k$ is defined, as in [3], by local projections, using $\Pi_{E,k}^\nabla v_{\delta i}$ in place of $v_{\delta i}$ to compute the moments of order $k-1$ and k:

$$\forall E \in \mathcal{T}_{\delta i}, \begin{cases} \left(\tilde{\Pi}_k^0 v_{\delta i}, p\right)_E = (v_{\delta i}, p)_E & \forall p \in \mathbb{P}_{k-2}(E) , \\ \left(\tilde{\Pi}_k^0 v_{\delta i}, p\right)_E = \left(\Pi_{E,k}^\nabla v_{\delta i}, p\right)_E & \forall p \in \mathbb{P}_k(E) \setminus \mathbb{P}_{k-2}(E) . \end{cases}$$

We finally introduce the following global discrete operators:

$$a_\delta(v,w) := \sum_{i \in \mathcal{I}} a_{\delta i}(v,w) \qquad \forall v, w \in V_\delta^\star,$$

$$(f, v_\delta)_\delta := \sum_{i \in \mathcal{I}} (f_i, v_{\delta i})_{\delta, F_i} \qquad \forall v \in V_\delta^\star.$$

Remark. A possible choice for the term S^E in the context of the simulation of the flow in DFNs is proposed in [9] and is given by the scalar product between the vectors containing the degrees of freedom of the two arguments on the element ([4, 6]). This choice guarantees property (7) under some basic regularity assumptions on the triangulation.

4.3 The locally conforming approach with Hybrid Mortar Virtual Elements

Let us now consider the locally conforming mesh $\mathcal{T}_\delta^{\mathrm{lc}}$ defined in subsection 4.1. Following [8], we will use the mortar method [11, 7, 12] to weakly enforce continuity of the solution h_δ across the traces. Let $m \in \mathcal{M}$ be a trace index and let i be the fracture index such that $(i,j) = \mathcal{I}_m$, we construct a discretization of Γ_m induced by the nodes of $\mathcal{T}_{\delta i}^{\mathrm{lc}}$, i.e. the nodes on Γ_m coincide with the vertices of the elements in $\mathcal{T}_{\delta i}^{\mathrm{lc}}$ lying on Γ_m. On this mesh the finite dimensional space $M_{\delta m}^{\mathrm{lc}} \subset \mathrm{L}^2(\Gamma_m)$ is defined, and we also set $M_\delta^{\mathrm{lc}} = \prod_{m \in \mathcal{M}} M_{\delta m}^{\mathrm{lc}}$. In the present context the space $M_{\delta m}^{\mathrm{lc}}$ contains piecewise (discontinuous) polynomials of degree k^λ in the interior of the trace and the polynomial functions of degree $k^\lambda - 1$ on the first and last intervals of the discretization, [11, 33, 8]. The discrete version of problem (3) within this framework is: find $h_\delta = h_\delta^0 + \mathcal{R}_\delta\left(h^D\right)$, with $h_\delta^0 \in V_\delta^{\mathrm{lc}}$ and $\lambda_\delta \in M_\delta^{\mathrm{lc}}$ such that,

$$\begin{cases} a_\delta\left(h_\delta^0, v_\delta\right) + b\left(v_\delta, \lambda_\delta\right) = (f, v_\delta)_\delta + \left(h^N, v_\delta\right)_{\Gamma^N} \\ \qquad\qquad - a_\delta\left(\mathcal{R}_\delta\left(h^D\right), v_\delta\right) & \forall v_\delta \in V_\delta^{\mathrm{lc}} , \\ b\left(h_\delta^0, \psi_\delta\right) = -b\left(\mathcal{R}_\delta\left(h^D\right), \psi_\delta\right) & \forall \psi_\delta \in M_\delta^{\mathrm{lc}} , \end{cases} \tag{9}$$

where the bilinear form $b(.,.)$ on each trace is evaluated as an $\mathrm{L}^2(\Gamma_m)$ scalar product.

Well posedness of (9) is proven in [8] under a quite common regularity assumption on the mesh. As a consequence the following *inf-sup* condition holds:

$$\inf_{\psi_\delta \in M_\delta^{\mathrm{lc}}} \sup_{v_\delta \in V_\delta^{\mathrm{lc}}} \frac{b(v_\delta, \psi_\delta)}{\|v_\delta\|_{V_\delta^{\mathrm{lc}}} \|\psi_\delta\|_{M_\delta^{\mathrm{lc}}}} \geq \beta\,,$$

for a constant $\beta > 0$ independent of δ. We remark that, in the present context the solution $\lambda_{\delta m}$ represents a piecewise polynomial approximation of $\left[\!\left[\frac{\partial h_i}{\partial \hat{n}_m^i}\right]\!\right]_{\Gamma_m}$. If we collect the DOFs for h_δ^0 and λ_δ in the vectors $\mathbf{h} \in \mathbb{R}^{N_h^{\mathrm{lc}}}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{N_\lambda^{\mathrm{lc}}}$, respectively, the discrete solution is obtained solving the following linear system:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{h} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{d} \end{pmatrix},$$

where $A \in \mathbb{R}^{N_h^{\mathrm{lc}} \times N_h^{\mathrm{lc}}}$ is the block-diagonal matrix of the fracture-local stiffness matrices

$$(A_i)_{kl} = a_{\delta i}\left(\phi_{ki}^{\mathrm{lc}}, \phi_{li}^{\mathrm{lc}}\right).$$

The matrix $B \in \mathbb{R}^{N_\lambda^{\mathrm{lc}} \times N_h^{\mathrm{lc}}}$ collects the terms of the form:

$$B_{lk} := b\left(\phi_k^{\mathrm{lc}}, \mu_l^{\mathrm{lc}}\right) \quad \forall l \in \{1, \dots, N_\lambda^{\mathrm{lc}}\},\ k \in \{1, \dots, N_h^{\mathrm{lc}}\},$$

being ϕ_k^{lc} the k-th basis function of V_δ^{lc} and μ_l^{lc} the l-th basis function of M_δ^{lc}. Finally we have

$$\begin{aligned} f_k &:= \left(f, \phi_k^{\mathrm{lc}}\right)_\delta + \left(h^N, \phi_k^{\mathrm{lc}}\right)_{\Gamma^N} - a_\delta\left(\mathcal{R}_\delta\left(h^D\right), \phi_l^{\mathrm{lc}}\right) && \forall k \in \{1, \dots, N_h^{\mathrm{lc}}\}, \\ d_l &:= -b\left(\mathcal{R}_\delta\left(h^D\right), \mu_l^{\mathrm{lc}}\right) && \forall l \in \{1, \dots, N_\lambda^{\mathrm{lc}}\}. \end{aligned}$$

4.4 A globally conforming approach

A second approach, based on a globally conforming discretization of the DFN is proposed in in [10], where the matching conditions at the traces are strongly enforced by means of Lagrange multipliers. We set, on each fracture F_i, $i \in \mathcal{I}$, and on each trace Γ_m, $m \in \mathcal{M}_i$, the discretization induced by $\mathcal{T}_{\delta i}^{\mathrm{gc}}$, excluding the two extreme points (tips) of the trace, and we build, on this mesh the finite dimensional space

$$M_{\delta m,i}^{\mathrm{gc}} = \mathrm{span}\left\{\mu_{ki}^m,\ k = 1, \dots, N_{\Gamma_m}^{\mathrm{gc}}\right\},$$

where $N_{\Gamma_m}^{\mathrm{gc}}$ is the number of interior nodes on Γ_m and μ_{ki}^m is a continuous linear operator such that

$$\left\langle \mu_{ki}^m, \gamma_{\Gamma_m}(v_{\delta j}) \right\rangle_{\Gamma_m} = \delta_{ij} v_{\delta i}(\mathbf{x}_k^m) \quad \forall v_\delta \in V_\delta^{\mathrm{gc}}, \tag{10}$$

being δ_{ij} the Kroneker delta and $\mathbf{x}_k^m$ the k-th node on trace Γ_m. We then set

$$M_{\delta m}^{\mathrm{gc}} := \left\{\mu_k^m \colon \mu_k^m = \mu_{ki}^m - \mu_{kj}^m \text{ if } \mathcal{I}_m = (i,j),\ k = 1, \dots, N_{\Gamma_m}^{\mathrm{gc}}\right\},$$

and the discrete subspace $M_\delta^{\mathrm{gc}} \subset M$ is $M_\delta^{\mathrm{gc}} = \prod_{m \in \mathcal{M}} M_{\delta m}^{\mathrm{gc}}$. Observe that the continuity of functions $v_\delta \in V_\delta^\star$ across al the traces, i.e. the condition $[\![v_\delta]\!]_{\mathcal{M}} = 0$, is equivalent to an orthogonality condition of jumps across the traces with respect to the space M_δ^{gc}

$$b(v_\delta, \psi_\delta) = 0 \quad \forall \psi_\delta \in M_\delta^{\mathrm{gc}} \iff [\![v_\delta]\!]_{\mathcal{M}} = 0\,,$$

being for any $m \in \mathcal{M}$, if $\mathcal{I}_m = (i,j)$, $\forall k \in \{1,\dots,N^{\mathrm{gc}}_{\Gamma_m}\}$,

$$\left\langle \mu^m_k, [\![v_\delta]\!]_{\Gamma_m} \right\rangle_{\Gamma_m} = \left\langle \mu^m_{ki}, \gamma_{\Gamma_m}(v_{\delta i}) \right\rangle_{\Gamma_m} - \left\langle \mu^m_{kj}, \gamma_{\Gamma_m}(v_{\delta j}) \right\rangle_{\Gamma_m} = v_{\delta i}(\mathbf{x}^m_k) - v_{\delta j}(\mathbf{x}^m_k).$$

Setting $h_\delta = h^0_\delta + \mathcal{R}_\delta\left(h^D\right)$, with $h^0_\delta \in V^{\mathrm{gc}}_\delta$ and $\mathcal{R}_\delta\left(h^D\right)$ the lifting of the boundary conditions, the discrete solution to (3) $(h^0_\delta, \lambda_\delta) \in V^{\mathrm{gc}}_\delta \times M^{\mathrm{gc}}_\delta$, such that

$$\mathcal{F}\left(h^0_\delta, \lambda_\delta\right) = \min_{v_\delta \in V^{\mathrm{gc}}_\delta} \max_{\psi_\delta \in M^{\mathrm{gc}}_\delta} \mathcal{F}(v_\delta, \psi_\delta) \ ,$$

is given by the solution of:

$$\begin{cases} a_\delta\left(h^0_\delta, v_\delta\right) + b\left(v_\delta, \lambda_\delta\right) = (f, v_\delta)_\delta + \left\langle h^N, \gamma_{\Gamma^N}(v_\delta)\right\rangle_{\Gamma^N} \\ \qquad\qquad\qquad\qquad + a_\delta\left(\mathcal{R}_\delta\left(h^D\right), v_\delta\right) & \forall v_\delta \in V^{\mathrm{gc}}_\delta, \\ b\left(h^0_\delta, \psi_\delta\right) = -b\left(\mathcal{R}_\delta\left(h^D\right), \psi_\delta\right) & \forall \psi_\delta \in M^{\mathrm{gc}}_\delta. \end{cases}$$

This problem is well posed, as it can be easily proven, observing that, given the space

$$W^{\mathrm{gc}}_\delta := \{v_\delta \in V^{\mathrm{gc}}_\delta \colon b(v_\delta, \psi_\delta) = 0 \quad \forall \psi_\delta \in M^{\mathrm{gc}}_\delta\} = \left\{v_\delta \in V^{\mathrm{gc}}_\delta \colon [\![v_\delta]\!]_{\Gamma_m} = 0 \quad \forall m \in \mathcal{M}\right\} \ ,$$

$v_\delta \mapsto \sum_{i\in\mathcal{J}} (\mathrm{K}_i \nabla v_{\delta i}, \nabla v_{\delta i})_{F_i}$ is a norm on W^{gc}_δ and then a_δ is coercive on W^{gc}_δ thanks to (8), and to the fact that

$$\forall \psi_\delta \in M^{\mathrm{gc}}_\delta, \quad \sup_{v_\delta \in V^{\mathrm{gc}}_\delta} \frac{b(v_\delta, \psi_\delta)}{\|v_\delta\|_{V^{\mathrm{gc}}_\delta}} = \|\psi_\delta\|_{M^{\mathrm{gc}}_\delta} \ .$$

The discrete solution in the present framework can be obtained solving the linear system

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{h} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{d} \end{pmatrix} , \tag{11}$$

where $\mathbf{h}$ is the vector collecting all the DOFs for h_δ, $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers, $\mathbf{f}$ is the vector containing the right-hand-side terms, $\mathbf{d}$ the vector of nodal values of h^D on the traces. Matrix A is again the block diagonal matrix obtained collecting the stiffness matrices A_i related to the VEM discrete bilinear forms on the globally conforming mesh of each fracture:

$$(A_i)_{kl} = a_{\delta i}\left(\phi^{\mathrm{gc}}_{ki}, \phi^{\mathrm{gc}}_{li}\right) \ .$$

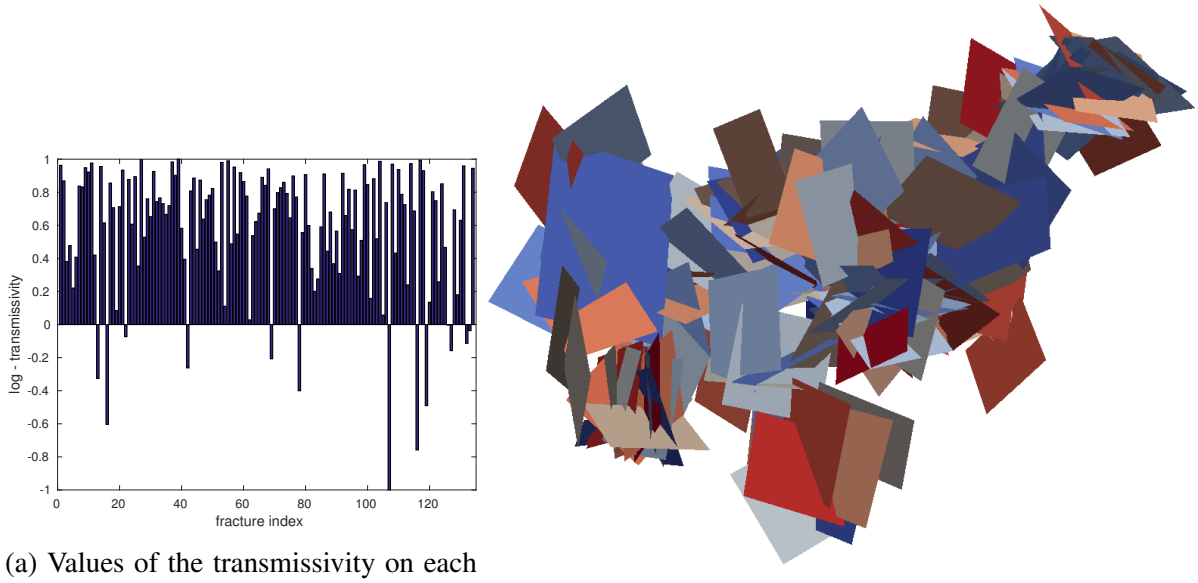
Matrix B is related to the bilinear form $b(v_\delta, \psi_\delta)$ in the following way. Let us observe that

$$b_i\left(\phi^{\mathrm{gc}}_l, \mu^m_k\right) = \begin{cases} (-1)^{s_{\Gamma_m}(i)} & \text{if } \mathbf{x}_l = \mathbf{x}^m_k, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{x}_l$ are the coordinates of the mesh vertex such that $\phi_l(\mathbf{x}_l) = 1$. After introducing a global numbering for the degrees of freedom on all the traces according to trace numbering, we introduce, for each trace index $m \in \mathcal{M}$, the row vector B_m such that $(B_m)_k = 1$ if $\mu^m_k \in M^{\mathrm{gc}}_{\delta m,i}$ and $(B_m)_k = -1$ if $\mu^m_k \in M^{\mathrm{gc}}_{\delta m,j}$ with $(i,j) = \mathcal{I}_m$. Finally

$$B := \begin{pmatrix} B_1 \\ \vdots \\ B_M \end{pmatrix} \ .$$

The solution to (11) is unique, as it can be proven using classical results (see e.g. [28]). Further, within this framework it is possible to use preconditioning techniques borrowed from domain decomposition method, as for example the one-level FETI preconditioner [25], as shown in [10].



(a) Values of the transmissivity on each fracture (the vertical axis is in log-scale).

(b) Plot of the distribution of transmissivity.

Figure 7: The distribution of transmissivity chosen for the numerical test.

5 Numerical results

For numerical results about the convergence behaviour of the described methods we refer the reader to [10, 8]. Here we focus on the application of the globally conforming method and the VEM-Mortar approach to a almost realistic DFN.

Let us consider the 134 Fractures DFN already depicted in Figure 1. We impose a non-null Neumann boundary condition (an incoming flux) on one side of three source fractures, a homogeneous Dirichlet boundary condition on one side of a sink fracture, and a homogeneous Neumann boundary condition (no flux) on all other boundaries. The forcing term is the null function, i.e. we assume there are no internal sources or sinks inside the fractures; this corresponds to neglect the effect on the fractures of the surrounding rock matrix. We assume a constant transmissivity on each fracture, with values randomly distributed between 10^{-1} and 10, displayed in Figure 7a. As we can see in Figure 7b, there are significant jumps of transmissivity between intersecting fractures. See Figure 8 for the discrete hydraulic head obtained with the globally conforming method and Figure 10 for a visualization of the solutions obtained by the two methods using the same base mesh, on two particular fractures.

In practical applications, the most important quantity to be evaluated is the flux through the traces. This is obtained as a direct solution (the Lagrange multiplier) of the system if the VEM-Mortar method is used, whereas in the globally conforming case it has to be computed as a post-processing of the discrete VEM solution, first projecting the latter on the space of polynomials of degree k and then computing the jump of the co-normal derivative of the projection in correspondence of traces.

The two considered approaches yield very similar results, as we can see in Figure 9, where we show the discrete fluxes computed with VEM of order 1 and the results for the VEM-Mortar method are obtained using continuous piecewise linear Lagrange multipliers. We remark that the oscillations of the discrete flux coming from the Mortar approach are justified by the fact that convergence is proved in the $\mathrm{H}^{-\frac{1}{2}}(\Gamma_m)$ norm, $\forall m \in \mathcal{M}$, which is a weaker norm than the $\mathrm{L}^2(\Gamma_m)$ norm.

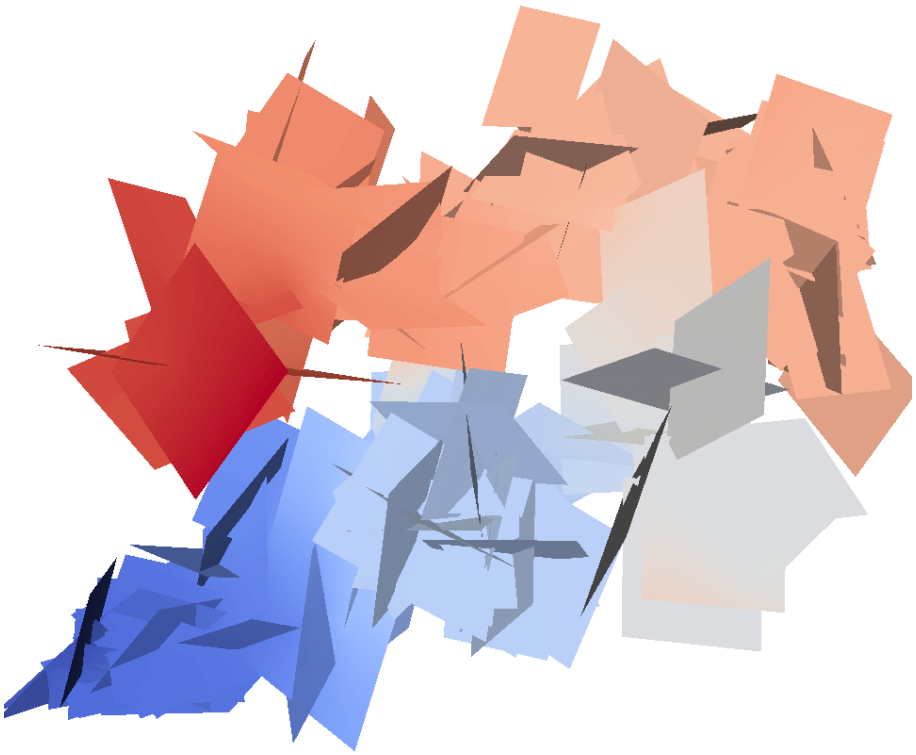


Figure 8: A discrete solution of the problem.

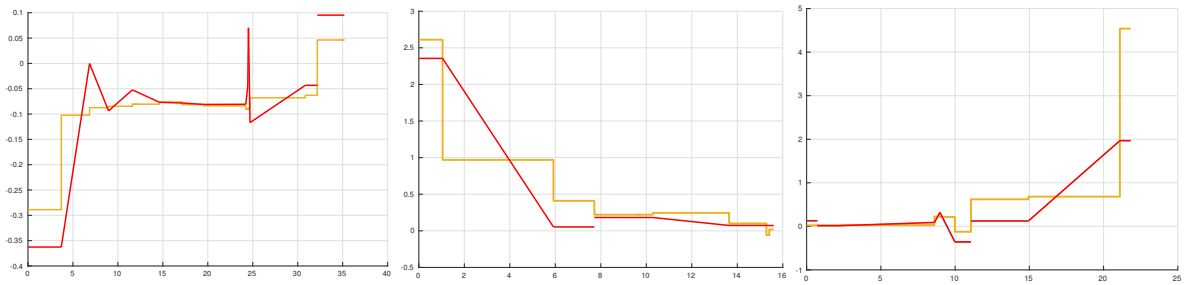
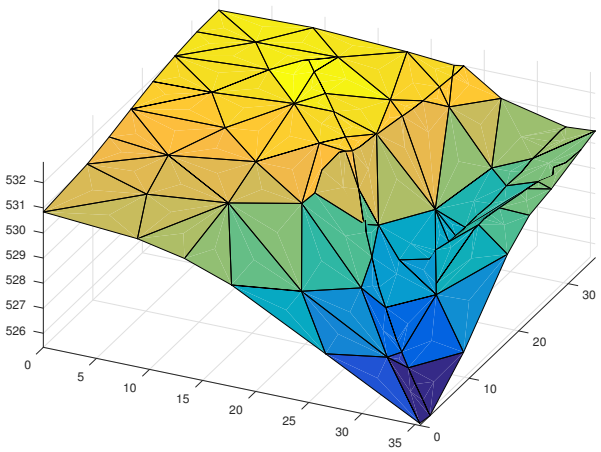
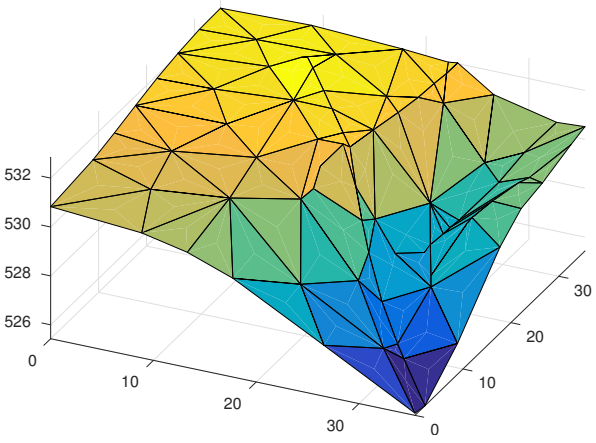


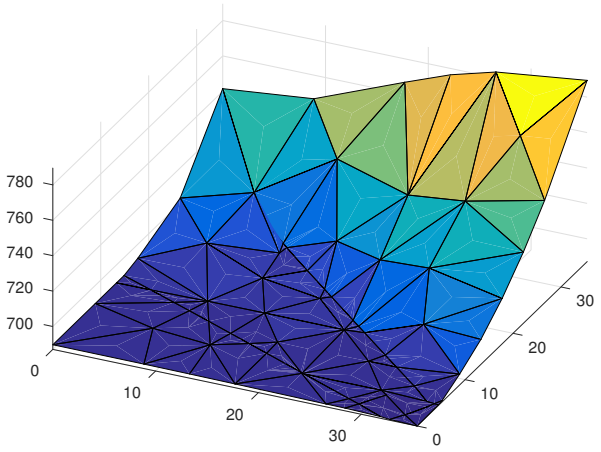
Figure 9: Comparison of fluxes computed by the two methods on three selected traces. Yellow: globally conforming VEM. Red: VEM-Mortar.



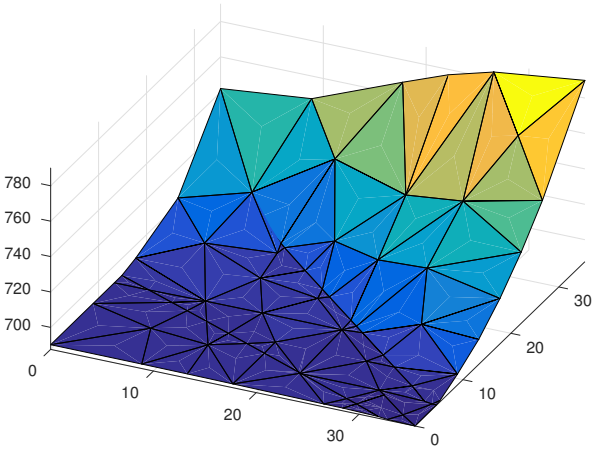
(a) Globally conforming method



(b) VEM-Mortar method



(c) Globally conforming method



(d) VEM – Mortar method

Figure 10: Solutions given by the two approaches on two selected fractures

6 Conclusions

We have shown that the capability of the VEM to handle a large class of polygons enables to easily construct functional spaces defined on local meshes on each of the fractures that are locally or globally conforming to traces. This allows for the use of standard domain decomposition approaches, coupling local problems by a Mortar method or, if meshes are globally conforming, by resorting to global continuity. Numerical results on the computed discrete fluxes at traces show that the two approaches are viable and sufficiently relaibale.

7 Acknowledgement

This work has been supported by the Italian MIUR through PRIN research grant 2012HBLYE4_001 "Metodologie innovative nella modellistica differenziale numerica" and by INdAM-GNCS.

Fernando Matìas Benedetto was supported by the European Commission through the Erasmus Mundus Action 2-Strand1 ARCOIRIS programme, Politecnico di Torino.

Stefano Berrone also acknowledges the partial financial support received from the European Unions Seventh Framework Programme (FP7/ 20142016) under Grant Agreement Number 607626 (Safeciti). Project title: "Simulation Platform for the Analysis of Crowd Turmoil in Urban Environments with Training and Predictive Capabilities". This publication reflects the views only of the authors and the Commission cannot be held responsible for any use which may be made of the information here contained.

REFERENCES

[1] P. M. Adler. *Fractures and Fracture Networks*. Kluwer Academic, Dordrecht, 1999.

[2] B. Ahmad, A. Alsaedi, F. Brezzi, L. D. Marini, and A. Russo. Equivalent projectors for virtual element methods. *Comput. Math. Appl.*, 66(3):376–391, September 2013.

[3] F. Beirão da Veiga, L. Brezzi, L. D. Marini, and A. Russo. The hitchhikers guide to the virtual element method. *Mathematical Models and Methods in Applied Sciences*, 24(08):1541–1573, 2014.

[4] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.*, 23(1):199–214, 2013.

[5] L. Beirão da Veiga, F. Brezzi, and L. D. Marini. Virtual elements for linear elasticity problems. *SIAM J. Numer. Anal.*, 51(2):794–812, 2013.

[6] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. Virtual element method for general second-order elliptic problems on polygonal meshes. *Mathematical Models and Methods in Applied Sciences*.

[7] F.B. Belgacem. The mortar finite element method with lagrange multipliers. *Numerische Mathematik*, 84(2):173–197, 1999.

[8] M.F. Benedetto, S. Berrone, A. Borio, S. Pieraccini, and S. Scialò. A hybrid mortar virtual element method for discrete fracture network simulations. *J. Comput. Phys.*, 306:148–166, 2016.

[9] M.F. Benedetto, S. Berrone, S. Pieraccini, and S. Scialò. The virtual element method for discrete fracture network simulations. *Comput. Methods Appl. Mech. Engrg.*, 280(0):135 – 156, 2014.

[10] M.F. Benedetto, S. Berrone, and S. Scialò. A globally conforming method for solving flow in discrete fracture networks using the virtual element method. *Finite Elem. Anal. Des.*, 109:23–36, 2016.

[11] C. Bernardi, Y. Maday, and A. T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In *Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. XI (Paris, 1989–1991)*, volume 299 of *Pitman Res. Notes Math. Ser.*, pages 13–51. Longman Sci. Tech., Harlow, 1994.

[12] C. Bernardi, Y. Maday, and F. Rapetti. Basics and some applications of the mortar element method. *GAMM-Mitt.*, 28(2):97–123, 2005.

[13] S. Berrone, A. Borio, and S. Scialò. A posteriori error estimate for a PDE-constrained optimization formulation for the flow in DFNs. *SIAM J. Numer. Anal.*, 54(1):242–261, 2016.

[14] S. Berrone, C. Canuto, S. Pieraccini, and S. Scialò. Uncertainty quantification in discrete fracture network models: stochastic fracture transmissivity. *Comput. Math. Appl.*, 70(4):603–623, 2015.

[15] S. Berrone, S. Pieraccini, and S. Scialò. A PDE-constrained optimization formulation for discrete fracture network flows. *SIAM J. Sci. Comput.*, 35(2):B487–B510, 2013.

[16] S. Berrone, S. Pieraccini, and S. Scialò. On simulations of discrete fracture network flows with an optimization-based extended finite element method. *SIAM J. Sci. Comput.*, 35(2):A908–A935, 2013.

[17] S. Berrone, S. Pieraccini, and S. Scialò. An optimization approach for large scale simulations of discrete fracture network flows. *J. Comput. Phys.*, 256:838–853, 2014.

[18] S. Berrone, S. Pieraccini, and S. Scialò. Towards effective flow simulations in realistic discrete fracture networks. *J. Comput. Phys.*, 310:181–201, 2016.

[19] S. Berrone, S. Pieraccini, S. Scialò, and F. Vicini. A parallel solver for large scale DFN flow simulations. *SIAM J. Sci. Comput.*, 37(3):C285–C306, 2015.

[20] D. Blessent, R. Therrien, and C.W. Gable. Large-scale numerical simulation of groundwater flow and solute transport in discretely-fractured crystalline bedrock. *Advances in Water Resources*, 34(12):1539 – 1552, 2011.

[21] F. Brezzi, R.S. Falk, and D.L. Marini. Basic principles of mixed virtual element methods. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(04):1227–1240, 2014.

[22] W. S. Dershowitz and C. Fidelibus. Derivation of equivalent pipe networks analogues for three-dimensional discrete fracture networks by the boundary element method. *Water Resource Res.*, 35:2685–2691, 1999.

[23] L. Formaggia, A. Fumagalli, A. Scotti, and P. Ruffo. A reduced model for Darcy's problem in networks of fractures. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48:1089–1116, 7 2014.

[24] J.D. Hyman, C.W. Gable, S.L. Painter, and N. Makedonska. Conforming delaunay triangulation of stochastically generated three dimensional discrete fracture networks: A feature rejection algorithm for meshing strategy. *SIAM Journal on Scientific Computing*, 36:A1871–A1894, 2014.

[25] A. Klawonn. FETI domain decomposition methods for second order elliptic partial differential equations. *GAMM-Mitteilungen*, 29(2):319–341, 2006.

[26] D. Mora L. Beirão da Veiga, C. Lovadina. A virtual element method for elastic and inelastic problems on polytope meshes. *ArXiv*, 2015.

[27] V. Lenti and C. Fidelibus. A *bem* solution of steady-state flow problems in discrete fracture networks with minimization of core storage. *Computers & Geosciences*, 29(9):1183 – 1190, 2003.

[28] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, Berlin, 1999.

[29] G. Pichot, J. Erhel, and J. de Dreuzy. A mixed hybrid mortar method for solving flow in discrete fracture networks. *Applicable Analysis*, 89:1629 – 643, 2010.

[30] G. Pichot, J. Erhel, and J. de Dreuzy. A generalized mixed hybrid mortar method for solving flow in stochastic discrete fracture networks. *SIAM Journal on scientific computing*, 34:B86 – B105, 2012.

[31] P.A. Raviart and J.M. Thomas. Primal hybrid finite element methods for 2nd order elliptic equations. *Mathematics of computation*, 31(138):391–413, 1977.

[32] M. Vohralík, J. Maryška, and O. Severýn. Mixed and nonconforming finite element methods on a system of polygons. *Applied Numerical Mathematics*, 51:176–193, 2007.

[33] B.I. Wohlmuth. A mortar finite element method using dual spaces for the lagrange multiplier. *SIAM journal on numerical analysis*, 38(3):989–1012, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

COMPOSITE FEM MODELS FOR LIMIT AND SHAKEDOWN ANALYSIS.

L. Leonetti[2], G. Garcea[2], and H. Nguyen-Xuan[3]

[1]Dipartimento di Modellistica per l'Ingegneria, Università della Calabria
87030 Rende (Cosenza), Italy
e-mail: domenico.magisano@hotmail.it

[2] Dipartimento di Modellistica per l'Ingegneria, Università della Calabria
87030 Rende (Cosenza), Italy
e-mail: {leonardo.leonetti,giovanni.garcea}@unical.it

[3] Interdisciplinary Research in Technology (CIRTech), Ho Chi Minh University of Technology (HUTECH), Vietnam
e-mail: ngx.hung@hutech.edu.vn

Keywords: Smoothed FEM, Plasticity, Limit Analysis, Shakedown

Abstract. *The paper improved S-FEMs formulations with an enriched displacement field, making use of modified Allman's shape functions. This mixed interpolation is the natural context in performing lower bound strategy for shakedown, limit analysis and elastoplastic analysis. The model takes advantages from the simplicity and few addressed requirements for good performances in nonlinear analysis. The simple assumption made for the stress field regards the convenience of using self-equilibrated stress interpolations in Cartesian coordinates. In the proposed composite elements the stress is discontinuous on the element and across their sides and the mesh of the elements is coincident with the discretization of the geometry. This stress interpolation is able to address the discontinuities in the plastic strain and, in such a way, to define in their description a finer mesh with respect the basic grid.*

1 Introduction

Plastic analysis plays an important role in the evaluation of structural safety of a wide class of structures. From mechanical point of view the plastic behaviour is characterised by the localization of the plastic deformations and by continuous displacements. Typically very fine meshes are required to address the plastic mechanism when standard finite elements are used. In particular the analysis doesn't take advantage from the use of high order FEM models. On the contrary lower-order finite elements shows many drawbacks and locking phenomena in many context of analysis. The goal is to improve the behaviour of lower-order finite elements due to its low computational cost when moderately fine meshes are required. In particular, linear triangular discretization shows a great capability to well describe complicated data and little sensitivity to the mesh distortion. This aspect is particularly desirable for a finite element procedure to increase the efficiency of algorithms in terms of computational cost and accuracy in both displacement and stress components and the robustness of the numerical process with respect to a large spectrum of data. In the attempts to improve the performance of lower-order finite elements, a key solution was reducing the overly-stiffness of the standard linear triangle T_3 model. A manner of obtaining the desired softening effects is the use of strain smoothing techniques as in the family of so-called smoothed finite element methods (S-FEMs) [1]. The essential idea in the S-FEMs was to define a smoothing domain through the mesh discretization with different patterns i.e. cells, nodes, edges or faces of the background mesh. It is worth mentioning that NS-FEM is very close or coincident to simplex FEM element proposed in [3], successfully in shakedown and limit analysis. Technical models, namely plate and beam models are often formulated by introducing rotational degrees of freedom. Their spatial assemblage and connection is easy if it is based on three translational and three rotational degrees of freedom per node as in the case of folded structures or curved shells. To this aim it is convenient to include in the formulation, even as an enrichment of the shape functions, the so called drilling rotations. It is worth mentioning that the Allman's like models [4] appear to be particularly simple and the choice of assuming a quadratic interpolation function for the side transversal displacement is effective but, as counterpart, the side interpolation becomes insensitive to the nodal rotation average and the resultant element can lead to defective rank. Moreover by using Allman's like interpolation the choice of referring to relative nodal rotations(i.e. to consider them as additional rotations superimposed on the linear deformation due to the nodal displacements) leads to some difficulties in recovering the absolute rotation field. In this respect [5] proposed a modified Allman's triangular element and a clear discussion of drilling rotations is done proving that the drilling degree's of freedom should be the rotation component of the elasticity. It shows also that the drilling parameter in the original Allman's element is not the Cauchy's continuum rotation (true rotation). It can become the true rotation by introducing a simple constraint that also eliminates the spurious zero modes. Our work needs only a slight change to the original Allman's element keeping the advantages of simplicity, accuracy and compatibilities of Allman's element. In this paper improved S-FEMs formulations with an enriched displacement field, making use of modified Allman's shape functions, is presented in its assumed stress version. This mixed interpolation strategy doesn't cause any difficulties in the linear elastic analysis and is the natural context in performing lower bound strategy for shakedown, limit analysis and elastoplastic analysis [6]. In addition, as the introduction of nonlinearities in the model leads to very complicate nonlinear contributions, an essential and working well element with few addressed requirement is of great interest. Another advantage of the simple assumption made for the stress field regards the convenience of using self-equilibrated stress interpolations

in Cartesian coordinates. With few implementation changes it is possible to derive composite FEM models [7] with the same interpolation fields. The difference from S-FEMs and composite FEM is topological. Both are based on a discretization grid but in S-FEMs some mechanical field stress or strain are assumed continuous trough the portions, and kept discontinuous within the part of the grid. In this way the mesh of the elements is not coincident with the grid of the parts. In the proposed composite elements the stress is discontinuous on the element and across their sides and the mesh of the elements is coincident with the discretization of the geometry. This stress interpolation is able to address the discontinuities in the plastic strain and, in such a way, to define in their description a finer mesh with respect the basic grid.

Structures, during their operational life, are subjected to a sequence of variable actions depicting, sometimes, a very complex loading scenario [2]

In this context shakedown analysis furnishes, in a direct way, a reliable safety factor against plastic collapse, loss in functionality due to excessive deformation (ratcheting) or collapse due to low cycle fatigue (plastic shakedown), and also provides valuable information about the internal stress redistribution due to the plastic adaptation phenomenon. .

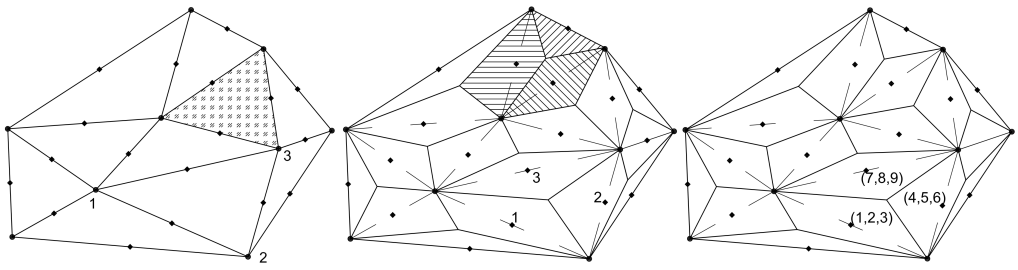
2 On an edge-based assumed stress field

In this section a new triangular mixed edge smoothed finite element (MES-FEM) is formulated. This aims to improve the displacement field by means of an Allman-like interpolation in order to account for drilling rotations.

2.1 The discretization grid and mesh

Similarly to ES-FEM, we start from a geometrical discretization of the two-dimensional domain (*grid*) (see Fig. 1(a)), by means of three node triangles (*parts*). Each part can be subdivided into three triangular *subparts* identified by each edge and the centroid of the part. On this grid the *element* is defined by the union of the subparts adjacent to each edge of the grid (see Figs. 1(b)-1(c)). The union of all the elements defines the *mesh*. Each part contributes to the elements corresponding to its sides, so in this paper the mesh (of the elements) is distinguished from the grid (of the parts).

In Figure 1(a) a part (triangle) of the grid is highlighted and an example of the numbering of the nodes of the mesh is given while in Fig. 1(b) three elements corresponding to the sides of a triangle have been indicated and a numbering example of the elements is shown. Finally in Fig. 1(c) an example of the stress parameter numbering is given.



(a) Grid of the parts (b) Mesh of the elements (c) Stress numbering

Figure 1: Grid (geometrical) and mesh (logical) discretization and stress numbering.

A suitable description of the relevant quantities of the discrete model can be obtained by using a triangular area co-ordinate system (r, s, t), related to the Cartesian ones (x, y) by the

co-ordinate transformation

$$\begin{bmatrix} r \\ s \\ t \end{bmatrix} = \frac{1}{2A} \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \end{bmatrix} \tag{1}$$

where A is the part area and the coefficients of the transformation matrix are defined as

$$a_i = x_j y_k - x_k y_j, \quad b_i = y_j - y_k, \quad c_i = x_k - x_j \tag{2}$$

following the permutation rule $i = \overrightarrow{123}$, $j = \overrightarrow{231}$, $k = \overrightarrow{312}$.

2.2 The displacement interpolation

Within each part the displacement field is described by a continuous interpolation expressed in terms of parameters $\{u_n v_n, \theta_n\}, n = 1..3$ located on its three corners (see Fig.1(a)). Making

$$\begin{cases} u_0 = u_1 r + u_2 s + u_3 t \\ v_0 = v_1 r + v_2 s + v_3 t \end{cases} \tag{3}$$

the usual shape functions for the linear triangle T_3, in the proposed model the displacement $\boldsymbol{u}[r,s,t] \equiv \{u,v\}^T$ is interpolated using the shapes proposed in [5]

$$\begin{cases} u = u_0 + \dfrac{2}{3}(\theta_1 - \theta_2) b_3 rs + \dfrac{2}{3}(\theta_2 - \theta_3) b_1 st + \dfrac{2}{3}(\theta_3 - \theta_1) b_2 tr \\ v = v_0 + \dfrac{2}{3}(\theta_1 - \theta_2) c_3 rs + \dfrac{2}{3}(\theta_2 - \theta_3) c_1 st + \dfrac{2}{3}(\theta_3 - \theta_1) c_2 tr \end{cases} \tag{4}$$

that is slightly different from the classical Allman Triangle (AT) due to the choice of the coefficient $\frac{2}{3}$ instead of $\frac{1}{2}$. By imposing the constraint

$$\frac{1}{2}\left(\frac{\partial v_0}{\partial x} - \frac{\partial u_0}{\partial y}\right) - \frac{\theta_1 + \theta_2 + \theta_3}{3} = 0 \tag{5}$$

the parameters θ_i become the nodal (drilling) rotations of the 2D Cauchy continuum $2\omega_i = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)$ and the rank defectiveness of the Allman-like triangles is sanitized. It is also possible to show that the solution is not affected by the choice of $\frac{1}{2}$ or $\frac{2}{3}$ as coefficient for drilling rotations when the constraint (5) is not imposed. We refer readers to [5] for further details.

With this choice the drilling rotations, which are useful for the connection with flexural elements, have the physical meaning of the continuum Cauchy problem and allow a coherent link with the displacement field. The displacement on each part can be expressed in compact matrix notation as

$$\boldsymbol{u}[r,s,t] = \Phi[r,s,t]\boldsymbol{u}_e \tag{6}$$

where $\boldsymbol{u}_e = [u_1, v_1, \theta_1, u_2, v_2, \theta_2, u_3, v_3, \theta_3]^T$ is the vector collecting the kinematical part parameters and Φ is the matrix collecting the quadratic shape functions (4).

Constraint equation (5), constant on the part, becomes, in matrix form

$$\mathbf{d}_e^T \boldsymbol{u}_e = 0 \tag{7}$$

where

$$\mathbf{d}_e^T = [\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3] \quad \text{with} \quad \mathbf{d}_k = \left[\frac{c_k}{4A}, \quad -\frac{b_k}{4A}, \quad \frac{1}{3}\right], \; k = 1..3.$$

2.3 The stress interpolation

The usual formulation of the ES-FEM assumes a constant strain on the element [8]. Here the finite element model is constructed by assuming independent interpolations to approximate the displacement and the stress field.

In particular the stress interpolation is assumed constant over the element and then it is piecewise constant over the parts (see Figs. 1(b)-1(c)) exactly as in the strain in the ES-FEM. For the plane stress case, the vector of the stress parameters of the part

$$\boldsymbol{\beta}_e = \left[\sigma^1_{xx}, \sigma^2_{xx}, \sigma^3_{xx}, \sigma^1_{yy}, \sigma^2_{yy}, \sigma^3_{yy}\sigma^1_{xy}, \sigma^2_{xy}, \sigma^3_{xy}\right]^T \tag{8}$$

is introduced to ordinate the stress components of the vector $\boldsymbol{\sigma}^i = [\sigma^i_{xx}, \sigma^i_{yy}, \sigma^i_{xy}]^T$ associated to each subpart $i = 1..3$. We express the selection of $\boldsymbol{\sigma}^i$ from $\boldsymbol{\beta}$ as

$$\boldsymbol{\sigma}^i = \boldsymbol{T}^i_\sigma \boldsymbol{\beta}_e \tag{9}$$

It is worth noting that the piecewise constant approximation of the stress components adopted has some significant advantages. A solution equilibrated for zero volume loads is described simply, whereas higher order interpolations are typically used in conjunction with equilibrium conditions which reduce the number of independent parameters. In addition, this choice avoids any co-ordinate transformation which could introduce noise related to the mesh distortion [9]. The mixed format makes an effective extension of the model to nonlinear analyses simpler. For application in plastic analysis [10], the method is expected to behave well as the discontinuous interpolation for the stress field addresses discontinuities in the plastic deformation field and the plastic admissibility is also imposed in a simple manner.

2.4 The part internal work

The equilibrium/compatibility operator of the part is obtained by the exact integration of the internal work. In particular, by using the Gauss theorem, the integral can be transferred to the boundary of each sub-part and, because the stress is self-equilibrated, the domain contributions are deleted, so obtaining

$$\int_A \boldsymbol{\sigma}^T \boldsymbol{\varepsilon} dA = \sum_{i=1}^{3} \int_{A_i} (\boldsymbol{\sigma}^i)^T \boldsymbol{\varepsilon} dA = \sum_{i=1}^{3} \left(\int_{\Gamma_i} \boldsymbol{n}^T \boldsymbol{\sigma}^T \boldsymbol{u} \, d\Gamma \right) = \boldsymbol{\beta}_e^T \boldsymbol{D}_e \boldsymbol{u}_e \tag{10}$$

The discrete equilibrium operator of the part is expressed in the compact form as

$$\boldsymbol{D}_e^T = \begin{bmatrix} \boldsymbol{B} & \mathbf{0} & \boldsymbol{C} \\ \mathbf{0} & \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}_y & \boldsymbol{C}_x & \mathbf{G} \end{bmatrix}, \quad \boldsymbol{B} = \frac{1}{6} \begin{bmatrix} b_1 & b_1 & b_1 & b_1 \\ b_2 & b_2 & b_2 & b_2 \\ b_3 & b_3 & b_2 & b_3 \end{bmatrix} \tag{11}$$

where matrices $\mathbf{G}$ and $\boldsymbol{B}_y$ are defined as

$$\mathbf{G} = \frac{1}{81} \begin{bmatrix} 2(4y_2x_2 - y_3x_3) - d_1 & -8(x_1b_1 + x_2b_3 + x_3b_2) \\ 2(y_3x_3 - 4x_1y_1) + d_2 & 2(4y_3x_3 - x_1y_1) - d_2 \\ -8(x_1b_2 + x_2b_1 + x_3b_3) & 2(x_1y_1 - 4y_2x_2) + d_3 \end{bmatrix}, \quad d_i = 3(x_jy_k + x_ky_j) - 5(a_j + a_k)$$

$$\boldsymbol{B}_y = \frac{1}{81} \begin{bmatrix} b_1(4b_3 - b_2) & 4b_1(b_3 - b_2) & b_1(b_3 - 4b_2) & 4b_1(b_3 - b_2) \\ b_2(b_1 - 4b_3) & b_2(4b_1 - b_3) & 4b_2(b_1 - b_3) & b_2(4b_1 - b_3) \\ 4b_3(b_2 - b_1) & b_3(b_2 - 4b_1) & b_3(4b_2 - b_1) & b_3(b_2 - 4b_1) \end{bmatrix}$$

The matrix $\boldsymbol{C}$ has the same structure as $\boldsymbol{B}$, but contains the coefficients c_i in spite of b_i (see eq. (2)). Similarly the matrix $\boldsymbol{C}_x$ is obtained by replacing, in the matrix $\boldsymbol{B}_y$, the coefficients b_i by the coefficients $-c_i$.

2.5 Implementation details

To make the preprocessing and assembly phases simple, quick and governed by standard operations, the discrete operators are defined over the part. In particular it is convenient to use a standard six node triangular grid of T_6 elements (see Fig. 1(a)) for the variable numbering. In this way the vertex nodes (indicated by bullets) of the T_6 element (see Fig. 1(a)) are used for the kinematical variable numbering while the mid-side nodes (indicated by squares) are used for the stress parameter numbering, as shown in Figs. 1(b)-1(c). With this approach the preprocessing phase is really fast, avoiding the search procedure of selecting the parts sharing the common edge.

The preprocessing phase of smoothed elements is complex, full details are given in [11], in contrast the strategy proposed here is provided freely by a mesh generator like `GMSH` [12]. In particular a similar strategy can be used in constructing the so-called cell based smoothed element (CS-FEM) that can take advantage of preprocessing based on quadratic quadrangular grids.

2.6 Comparison with other FE models

As shown in [8, 11], ES-FEM behaves well in many contexts of analysis. The reason for this good behaviour is that the discontinuity of the approximate strain field is reduced significantly, due to the smoothing effect. In this respect, ES-FEM and Imbricate FEM(EI-FEM) are close formulations, while the smoothing domain is the same. This can be easily verified by showing the effect of a single non-zero displacement component at a node for ES-FEM . It can be seen that the domain associated with the non-zero strains, is the same as that presented in Fig. 9 of [11] for the edge imbricate FEM and so similar smoothing effects are expected. Also for the MES-FEM presented in this work, the active stress components triggered by the solution, propagate the effect to the neighbourhood of the elements sharing the active kinematical variable, so increasing the smoothness of the displacement field. In this case, however, in contrast to ES-FEM, the smoothing concept is applied/extended to Allman-like kinematics.

The proposed element also has similarities with the composite $MT_{6/3}$ element [7, 13, 22], the main difference being topological. In contrast to MES-FEM, in $MT_{6/3}$ the grid and the mesh are coincident and displacement interpolation is that of the quadratic triangle T_6. The stress assumptions look similar for both models as each part is partitioned into three triangular sub-parts in which the stress is assumed constant. However for the MES-FEM the stress components are the same for the two sub-parts on a common edge while for the $MT_{6/3}$ they are different.

Note that the Allman-like interpolation can be seen as a reduction or simplification of the full quadratic displacement interpolation used for $MT_{6/3}$. In particular it can be seen as being derived from the T_6 FE by deleting the mid-side node and adding some constraints to maintain a quadratic interpolation for the displacement component transversal on the edge only, and linear in the side direction. Similarly to ES-FEM, which improves the T_3 interpolation of the part with the smoothing effect, MES-FEM also enriches the accuracy of the Allman-like interpolation. In this respect the mid-side node of the quadratic displacement interpolation used in the $MT_{6/3}$ element plays an inter-element connection role similar to the conflict domain of the ES-FEM and EI-FEM elements.

To show the importance of this smoothing effect on the accuracy of the solution, a composite FE, denoted AM-CM, derived from the $MT_{6/3}$ by replacing the original displacement shape functions of the quadratic triangle T_6 with those in Eqs.(3)-(4), is derived. In this way the difference between AM-CM and MES-FEM consists in the discontinuity of the stress fields through the sides of the triangles (coincidence of the grid and the mesh), i.e AM-CM does not have any smoothing effects. Notwithstanding the greater number of dofs of the stress interpolation of AM-CM, the smoothing effect presented in MES-FEM makes an important improvement in the accuracy as will be shown by the numerical experiments.

3 Plastic collapse analysis

We refer to the analysis of a body subjected to volume forces **b** and tractions **t**, both increasing with the same load multiplier λ. The material is assumed to be elastic-perfectly plastic, therefore the stress field is constrained to satisfy plastic admissibility inequalities which are independent of the plastic strain. The limit analysis aims to identify the plastic collapse state, characterized by the collapse multiplier λ_c and the associated plastic mechanism. A way to solve this problem is represented by the evolutive analysis which furnishes the entire structural response through the solution of a sequence of incremental elasto-plastic problems [6]. The collapse multiplier λ_c is evaluated as the limit value for the equilibrium path.

The limit analysis theorems offer an alternative way which is directly addressed to compute the lower and upper approximations of the collapse limit multiplier. The approximation furnished by the statical characterization of the collapse load, giving a safe estimate, is more attractive from an engineering point of view. In this case the collapse multiplier is individuated as a solution of the nonlinear mathematical programming problem

$$\begin{aligned} &\text{maximize} && \lambda \\ &\text{subject to} && \boldsymbol{Q}^T\tilde{\boldsymbol{\beta}} - \lambda\boldsymbol{p} = \boldsymbol{0} \\ & && \boldsymbol{f}[\boldsymbol{\beta}] \leq \boldsymbol{0} \end{aligned} \tag{12}$$

where the equality constraints are represented by the equilibrium equations, described through the global equilibrium operator $\boldsymbol{Q}^T$ and the load vector $\boldsymbol{p}$ collecting the body forces and tractions (see eqs. 17 and 19). The plastic admissibility inequalities are expressed through the vector $\boldsymbol{f}$, which collects the local restrictions imposed by the assumed yield condition over the stress state $\boldsymbol{\sigma}^i$ of the N_r regions of the domain

$$f_i[\boldsymbol{\sigma}^i] \leq 0, \quad i = 1..N_r \tag{13}$$

Within the numerical methods for the analysis of nonlinear mathematical programming problems, the Interior Point strategy represents an established approach [14], [15]. Specialized implementations of these algorithms in plastic problems can be found in [16] [17]. This method also furnishes the solution of the dual problem, namely the Lagrange multipliers of the functional

$$\mathcal{L} = \lambda + \dot{\boldsymbol{q}}^T(\boldsymbol{Q}^T\tilde{\boldsymbol{\beta}} - \lambda\boldsymbol{p}) - \dot{\boldsymbol{\mu}}^T\boldsymbol{f} \tag{14}$$

which are represented by the mechanism vector $\dot{\boldsymbol{q}}$ and the vector of the plastic multipliers $\dot{\boldsymbol{\mu}}$.

The dual problem can be written in the form

$$\begin{aligned} \text{maximize} \quad & \dot{\boldsymbol{\mu}}^T \boldsymbol{\sigma}_y \\ \text{subject to} \quad & \boldsymbol{Q}\,\dot{\boldsymbol{q}} - \dot{\boldsymbol{\mu}}^T \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\beta}} = \mathbf{0} \\ & 1 - \dot{\boldsymbol{q}}^T \boldsymbol{p} = \mathbf{0} \\ & \dot{\boldsymbol{\mu}} \geq \mathbf{0} \end{aligned} \tag{15}$$

which is derived by imposing the stationarity conditions of the Lagrange functional (14) with respect $\boldsymbol{\beta}$ and λ, variables of the primal problem (12), and by using the Euler theorem

$$f_i[\boldsymbol{\sigma}^i] = (\boldsymbol{\sigma}^i)^T \frac{\partial f_i}{\partial \boldsymbol{\sigma}^i} - {\sigma_y}^i, \quad i = 1..N_r \tag{16}$$

to express the yield conditions on each of the stress regions where the yield stress ${\sigma_y}^i$ is assigned.

Since our objective is to test the performances of the proposed element, a standard numerical tool has been chosen to solve the programming problem. In this way the comparison results are unaffected by our implementation details and tolerances, so the proposed element is tested in a standard and well consolidated environment without special modifications. In order to obtain a high level of accuracy, the MOSEK [18] toolbox for MATLAB [19] has been selected. This algorithm has proved to be highly robust and efficient in independent benchmark tests and, without any tuning of its default settings, to be able to solve very large problems with remarkable speed [20].

3.1 The discrete equilibrium equations

We consider a plane body occupying the domain Ω having a boundary Γ subjected to the body forces **b** and the traction **t**. The proncipal virtual work equation becomes

$$W[\boldsymbol{\sigma}, \boldsymbol{u}] = \int_\Omega \boldsymbol{\sigma}^T \boldsymbol{D}\boldsymbol{u}\, d\Omega - \int_\Omega \mathbf{b}^T \boldsymbol{u}\, d\Omega - \int_\Gamma \mathbf{t}^T \boldsymbol{u}\, d\Gamma \tag{17}$$

where $\boldsymbol{D}$ is the compatibility operator. When a constraint like (5) needs to be imposed, an extra Lagrange multiplier could be introduced or the functional could be penalized by adding to Eq.(17) the following term

$$W_c = \int_\Omega \lambda_L \left(\frac{1}{2} \left(\frac{\partial v_0}{\partial x} - \frac{\partial u_0}{\partial y} \right) - \frac{\theta_1 + \theta_2 + \theta_3}{3} \right) d\Omega \tag{18}$$

The modified functional W introducing the interpolation then becomes

$$W[\boldsymbol{\beta}, \boldsymbol{q}] \equiv W[\boldsymbol{\sigma}, \boldsymbol{u}] + W_c = \sum_e W_e[\boldsymbol{\beta}_e, \boldsymbol{u}_e]$$

where the eth part contribution to the functional (18) is defined as

$$W_e[\boldsymbol{\beta}_e, \boldsymbol{u}_e] = \boldsymbol{\beta}_e^T \boldsymbol{D}_e \boldsymbol{u}_e + \lambda_L \boldsymbol{D}_{ed} \boldsymbol{u}_e - \boldsymbol{u}_e^T \boldsymbol{p}_e \tag{19}$$

and $\boldsymbol{D}_e$ is the compatibility operator defined in Eq.(11), $\boldsymbol{p}_e$ is the load vector furnished by the integration of the external load components weighted with the shape functions of the displacement interpolations and the constraint contribution gives the contribution to the compatibility operator $\boldsymbol{D}_{ed} = \mathbf{d}_e A_e$.

By assembling the element contributions in the global matrices through the element incidence operators for displacements $\boldsymbol{u}_e = \boldsymbol{T}_u \boldsymbol{q}$ and stresses $\boldsymbol{\beta}_e = \boldsymbol{T}_\beta \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the global stress vector and $\boldsymbol{q}$ is the global displacement vector, the sum of the contributions (19) furnishes the global matrices

$$\boldsymbol{Q} = \sum_e \begin{bmatrix} \boldsymbol{T}_\beta^T \boldsymbol{D}_e \boldsymbol{T}_u \\ \boldsymbol{D}_{ed} \boldsymbol{T}_u \end{bmatrix}, \quad \tilde{\boldsymbol{\beta}}^T = \left[\boldsymbol{\beta}^T, \boldsymbol{\lambda}_L^T\right] \tag{20}$$

It is worth noting that matrix, the single step of the elastoplastic analysis [6, 21, 10, 22] can be seen as the solution of the mathematical problem (12) by adding the compliance matrix $\mathbf{F}$ that is block diagonal at the element (stress) level due to the constant interpolation adopted and its inverse $\mathbf{F}^{-1}$ can be directly assembled. In this way an inexpensive construction of the tangent stiffness matrix [6, 10] is obtained.

The proposed assembly rule of the finite element, once the matrices used in defining the discrete operators in Eqs. (11) are provided, is then very light and follows the standard FE format. Furthermore the resulting discrete operator has a compact form and could easily be evaluated.

4 Numerical results

The performances of the proposed mixed finite element models in evaluating the plastic collapse states have been tested by the numerical experiments reported in the following subsections. The plastic collapse states have been analyzed by using Mises yielding criterion and taking a unitary yield stress.

Cook membrane

The following results refer to the elastic analysis of the well investigated Cook's membrane, reported in Figure 2. The convergence of the numerical solution has been tested by using four meshes obtained by successive refinements of the coarse mesh (*mesh 1*), having two elements for each side.

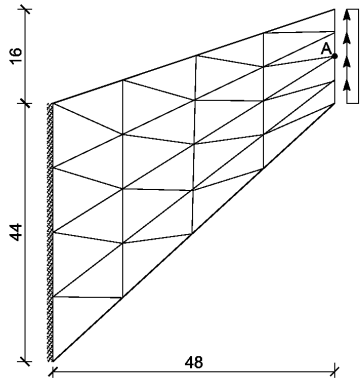


Figure 2: Cook membrane. Regular and irregular coarse meshes.

Table 1 reports a comparison of the computed values of the plastic collapse multiplier and the number of iterations spent on each analysis. The reference result [21] was obtained using a mesh having 1024 elements and 2178 dofs.

Table 1: Cook membrane. Plane stress limit analysis.

	mesh 1	*mesh 2*	*mesh 3*	*mesh 4*
	λ_c	λ_c	λ_c	λ_c
$MES-FEM$	0.428	0.405	0.399	0.396
$AM-CM$	0.451	0.426	0.405	0.399
$E-S$	0.505	0.415	0.401	0.397
ref. [21]			0.3956	

Square plate with circular hole

The square plate with a central circular hole, subjected to a constant traction, is often used as stress concentration test. The meshes are generated by splitting the right side and the top side into n parts and the remaining sides into $2n$ parts. The grids $mesh\ 1$, $mesh\ 2$ and $mesh\ 3$ refer to $n = 3, 6, 12$, respectively. Figure 3 shows the coarse mesh, in the regular and irregular versions.

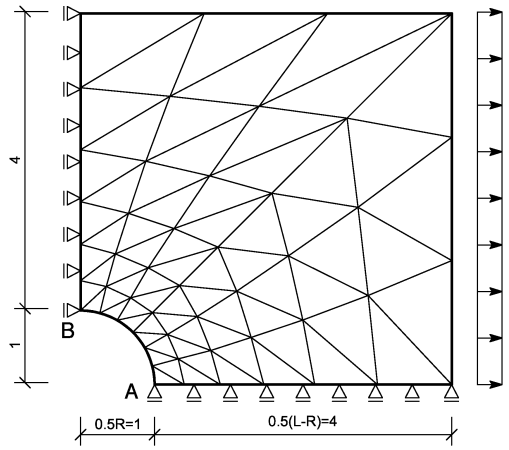


Figure 3: Square plate with circular hole. Coarse regular and irregular meshes.

The numerical values of the collapse multipliers, computed in plane stress condition and by using the Mises criterion, are compared in Table 2 for the three meshes. It is worth noting that the accurate reference results [21] were computed with 4802 dofs, 2304 elements and 9216 admissibility constraints. The exact collapse value has been computed with the analytical formula

$$\lambda_c\, t = (1 - \frac{R}{L})\sigma_y \tag{21}$$

5 Concluding remarks

This paper proposed a mixed edge-based smoothed finite element formulation enriched with drilling degrees of freedom for the analysis of structural plane problems and a composite FEM model. The main features of the model are its simplicity and ease of implementation within

Table 2: Square plate with circular hole. Plane stress limit analysis.

	mesh 1	*mesh 2*	*mesh 3*
	λ_c	λ_c	λ_c
$MES-FEM$	0.811	0.804	0.802
$AM-CM$	0.818	0.808	0.804
$E-S$	0.812	0.805	0.802
ref. [10]			0.8006
exact			0.8000

existing computational tools. More importantly, the element proves to be accurate and robust for the plastic analysis of plane problems. The efficiency of the present model comes from the properties of ES-FEM and from the choice made for the interpolation functions for the in-plane rotations that seem to match the smoothing effect of the stress field.

The numerical experimentation of the model, which was carried out on a wide range of problems and data, shows its excellent performance.

The experiments also demonstrate its accuracy in representing the rotational field, considerably improving the results of other models with the same displacement description. In particular it is shown that the MES-FEM performs better than the composite model.

REFERENCES

[1] Liu GR, Nguyen-Thoi T. Smoothed finite element methods. New York: CRR Press, Taylor and Francis Group(2010).

[2] Leonetti L, Casciaro R, Garcea G. Effective treatment of complex statical and dynamical load combinations within shakedown analysis of 3D frames. Computers and Structures 158(2015), 124-139

[3] Garcea G, Armentano G, Petrolo S,Casciaro R. Finite element shakedown analysis of two-dimensional structures. International Journal for Numerical Methods in Engineering. 63(2005)1174-1202.

[4] Allman DJ, A compatible triangular element including vertex rotations for plane elasticity analysis, Comput. Struct. 19(1984)1-8.

[5] Huang M, Zhao Z, Shen C. An effective planar triangular element with drilling rotation. Finite Elements in Analysis and Design 46(2010)1031–1036.

[6] Garcea G, Leonetti L. A unified mathematical programming formulation of strain driven and interior point algorithms for shakedown and limit analysis. International Journal for Numerical Methods in Engineering. 88(2011)1085–1111.

[7] Leonetti L.,Aristodemo M. A composite mixed finite element model for plane structural problems. Finite Elements in Analysis and Design 2015, **94**:33–46.

[8] Liu GR, T. Nguyen-Thoi T, Lam KY. An edge-based smoothed finite element method (ES-FEM) for static, free and forced vibration analyses of solyds. Journal of Sound and Vibration. 320(2009)1100–1130.

[9] Aminpour MA. Direct formulation of a hybrid finite element with drilling degrees of freedom. Int J Numer Methods Eng. 35(1992)997–1013.

[10] Bilotta A, Leonetti L, Garcea G. An algorithm for incremental elastoplastic analysis using equality constrained sequential quadratic programming. Computers and Structures. 102–103(2012)97–107.

[11] Cazes F, Meschke G. An Edge-based Imbricate Finite Element Method (EI-FEM) with full and reduced integration. Computers and Structures. 106-107(2012)154–175.

[12] Geuzaine C. Remacle JF. Gmsh: a three-dimensional finite element mesh generator with built-in pre-and post-processing facilities.

[13] Leonetti L., Le CV. Plastic collapse analysis of Mindlin-Reissner plates using a composite mixed finite element. Int J Numer Methods Eng . DOI: 10.1002/nme.4997.

[14] Karmarkar N. A New Polynomial Time Algorithm for Linear Programming. Combinatorica. 4(1984)373–395.

[15] Nemirovski A Todd M (2008) Interior-point methods for optimization. Acta Numerica. 191–234.

[16] Krabbenhoft K., Damkilde L. A general non-linear optimization algorithm for lower bound limit analysis. International Journal for Numerical Methods in Engineering. 56(2003)165-184.

[17] Pastor F, Loute E. Solving limit analysis problems: An interior-point method. Communications in Numerical Methods in Engineering. 21(2005)631-642.

[18] A.S. The MOSEK Optimization Tools Version 3.2 (Revision 8). User's Manual and Reference Available from,(2005) January http://www.mosek.com.

[19] MATLAB and Mosek Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

[20] Mittelmann HD. An independent benchmarking of SDP and SOCP solvers. Mathematical Programming. 95(2003)407–430.

[21] Bilotta A, Leonetti L, Garcea G. Three field finite elements for the elastoplastic analysis of 2D continua. Finite Elements in Analysis and Design. 47(2011)1119–1130.

[22] Bilotta A, Garcea G, Leonetti L. A composite mixed finite element model for the elastoplastic analysis of 3D structural problems. Finite Elements in Analysis and Design. 113(2016) 43–53. doi:10.1016/j.finel.2016.01.002

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON SHAKEDOWN OF ELASTIC-PLASTIC BODIES WITH TEMPERATURE DEPENDENT PROPERTIES

Michaël Peigney

Université Paris-Est, Laboratoire Navier (UMR 8205),
CNRS, Ecole des Ponts ParisTech, IFSTTAR, F-77455 Marne la Vallée, France
e-mail: michael.peigney@enpc.fr

Keywords: Plasticity, Shakedown, Temperature-dependent properties, Melan's theorem

Abstract. *For elastic-perfectly plastic structures under prescribed loading histories, the well-known Melan's theorem gives a sufficient condition for shakedown to occur, i.e. for the evolution to become elastic in the large-time limit. The original Melan's theorem rests on the assumption that the material properties remain constant in time, independently on the applied loading. This communication addresses the long standing issue of extending Melan's theorem to temperature-dependent (or time-dependent) elastic moduli. The main motivation is to extend the range of applications of Melan's theorem to thermomechanical loading histories with large temperature fluctuations: In such case, the variation of the elastic properties with the temperature cannot be neglected. Some recent results obtained in perfect plasticity and viscoplasticity are presented and discussed.*

1 INTRODUCTION

For elastic-perfectly plastic structures under prescribed loading histories, the well-known Melan's theorem [1, 2, 3] gives a sufficient condition for shakedown to occur, i.e. for the evolution to become elastic in the large-time limit. Intuitively, shakedown means that the plastic strain tends to a limit as time tends to infinity. The Melan's theorem has the distinctive property of being path-independent, i.e. independent on any residual stress that may exist initially in the structure. The shakedown theory has been the object of numerous developments, regarding both extensions of the original theorem to various nonlinear behaviors [4, 5, 6, 7, 8, 9] and numerical methods for assessing the shakedown limits in the case of parametrized loading histories [10, 11, 12, 13, 14, 15, 16].

This communication addresses the issue of extending Melan's theorem to temperature-dependent (or time-dependent) elastic moduli. The main motivation is to extend the range of applications of Melan's theorem to thermomechanical loading histories with large temperature fluctuations: In such case, the variation of the material properties with the temperature cannot be neglected. Whereas the case of temperature-dependent yield limits is well understood [17], the case of temperature-dependent elastic moduli remains a long standing issue and has been the object of several conjectures [18, 19, 20, 21]. The main difficulty is that the proof used in the original Melan's theorem – as well as in most of its knows extensions – crucially relies on some contraction properties that are lost when the elastic moduli are allowed to vary in time. A shakedown theorem has recently been proposed for elastic-perfectly plastic materials with time-periodic elastic properties [22]. The statement and proof of that theorem differ significantly from the case of constant material properties. A salient result is that time fluctuations of the elastic moduli need to remain small (in a certain sense) for shakedown to occur in a path-independent fashion.

In this communication, the theorem of [22] is presented and improved upon. The extension to elastic-viscoplastic materials is discussed.

2 EVOLUTION EQUATION FOR THE STRESS FIELD

Consider an elastic-plastic material occupying a domain Ω in the reference configuration. Under the small strains assumption, the strain ε, stress σ and plastic strain ε^p at position x and time t are related by the constitutive equations

$$\varepsilon(x,t) = L(x,t) : \sigma(x,t) + \varepsilon^\theta(x,t) + \varepsilon^p(x,t), \tag{1}$$

$$\dot{\varepsilon}^p(x,t) = \phi'(\sigma(x,t),x,t), \tag{2}$$

where ϕ is the dissipation potential, assumed to be positive, convex, and null at the origin. In (1), L is the (symmetric positive definite) elastic moduli tensor and ε^θ is the thermal strain tensor. Note the time dependence of L, ε^θ and ϕ, which may result from imposed variations of the temperature. The elastic domain of the material is denoted by $\mathscr{C}(x,t)$, i.e. $\mathscr{C}(x,t) = \{\sigma : \phi(\sigma,x,t) = 0\}$.

Assuming quasi-static evolutions, the stress field σ satisfies the equilibrium equations

$$\operatorname{div}\sigma + f = 0 \text{ in } \Omega,\ \sigma\cdot n = T \text{ on } \partial\Omega_T, \tag{3}$$

where $f(x,t)$ are body forces and $T(x,t)$ are tractions prescribed on a part $\partial\Omega_T$ of the boundary

$\partial\Omega$. The strain field ε and displacement field u satisfy

$$\varepsilon = \frac{1}{2}(\nabla u + \nabla^T u) \text{ in } \Omega,\ u = v \text{ on } \partial\Omega_u = \partial\Omega - \partial\Omega_T \tag{4}$$

where $v(x,t)$ is a given function.

As usual in shakedown analysis, it is convenient to introduce the *fictitious elastic response* (u^E, σ^E), i.e. the response of the system if it were purely elastic, defined by

$$\begin{aligned} \varepsilon^E &= L : \sigma^E + \varepsilon^\theta, \\ \operatorname{div}\sigma^E + f = 0 \text{ in } \Omega, &\ \sigma^E \cdot n = T \text{ on } \partial\Omega_T, \\ \varepsilon^E = \frac{1}{2}(\nabla u^E + \nabla^T u^E) \text{ in } \Omega,&\ u^E = v \text{ on } \partial\Omega_u. \end{aligned} \tag{5}$$

The boundary value problem defined by Eqs. (1-4) governs the evolution of the fields (u, ε, σ). An evolution equation involving only the stress field can be formulated from Eqs. (1-4), as is now detailed. Let $\mathcal{E}$ denote the space of stress fields, chosen as a subspace of symmetric second-order tensor fields with square-integrable components. The space $\mathcal{E}$ is equipped with the scalar product $\langle \sigma, \sigma' \rangle = \int_\Omega \sigma(x) : \sigma'(x) d\omega$. The associated norm is denoted by $\| \ \|$, i.e. $\|\sigma\| = \sqrt{\langle \sigma, \sigma \rangle}$. For simplicity we assume in the following that $\mathcal{E}$ is of finite dimension.

Any stress field σ satisfying (3) can be written as $\sigma = \sigma^E + \rho$ where ρ belongs to the vectorial space $\mathcal{H} \subset \mathcal{E}$ of self-equilibrated stress fields, defined by

$$\mathcal{H} = \{\rho \in \mathcal{E} : \operatorname{div}\rho = 0 \text{ in } \Omega,\ \rho \cdot n = 0 \text{ on } \partial\Omega_T\}. \tag{6}$$

For an arbitrary $\rho' \in \mathcal{H}$, it follows from (1) that

$$\int_\Omega (\rho' - \rho) : \frac{d(\varepsilon - \varepsilon^E)}{dt} d\omega = \int_\Omega (\rho' - \rho) : \frac{d}{dt}(L : \rho) d\omega + \int_\Omega (\rho' - \rho) : \phi'(\sigma) d\omega. \tag{7}$$

The principle of virtual power shows that the left-hand side of (7) is equal to 0, hence

$$-\int_\Omega (\rho' - \rho) : \frac{d}{dt}(L(x,t) : \rho) d\omega = \int_\Omega (\rho' - \rho) : \phi'(\sigma, x, t) d\omega. \tag{8}$$

Let $\Phi(\rho, t) : \mathcal{H} \mapsto \mathbb{R}$ be the convex, positive function defined by

$$\Phi(\rho, t) = \int_\Omega \phi(\sigma^E(x,t) + \rho(x), x, t) d\omega.$$

Note that $\Phi(\cdot, t)$ vanishes on the set $\mathcal{K}(t) = \{\rho \in \mathcal{H} : \rho(x) + \sigma^E(x,t) \in \mathscr{C}(x,t)\ \forall x \in \Omega\}$. For later reference, note that $\mathcal{K}(t) = \mathcal{K}_0(t) \cap \mathcal{H}$ where

$$\mathcal{K}_0(t) = \{\rho \in \mathcal{E} : \rho(x) + \sigma^E(x,t) \in \mathscr{C}(x,t)\ \forall x \in \Omega\}.$$

We have (at least formally)

$$\int_\Omega (\rho' - \rho) : \phi'(\sigma) d\omega = \langle \rho' - \rho, \Phi'(\rho, t) \rangle.$$

Further introducing the linear symmetric operator $\mathcal{L}(t) : \mathcal{H} \mapsto \mathcal{H}$ defined by

$$\langle \rho, \mathcal{L}(t)\rho' \rangle = \int_\Omega \rho(x) : L(x,t) : \rho'(x,t) d\omega \ \ \forall \rho, \rho' \in \mathcal{H},$$

Eq. (8) can be rewritten as

$$-\frac{d}{dt}(\mathcal{L}(t)\rho) = \Phi'(\rho,t). \tag{9}$$

Since the elastic moduli tensor $L(x,t)$ is symmetric positive definite, it can easily be verified that $\mathcal{L}(t)$ is self-adjoint and positive definite. Its inverse is denoted by $\mathcal{M}(t)$. Starting from a given initial state $\rho(t=0)$, the evolution of the stress field in $\mathcal{H}$ is governed by the ordinary differential equation (9). The uniqueness of the stress rate $\dot{\rho}$ has been proved in [23].

In the case of time-independent elastic moduli, a change of scalar product allows Eq. (9) to be put in the form

$$-\frac{d\rho}{dt} = \Phi'(\rho,t). \tag{10}$$

Properties of Eq. (10) are well understood. In particular, the distance between two solutions (as measured by the energy norm) is decreasing with time: For two solutions $\rho(t)$ and $\rho'(t)$ of (10), we have indeed

$$\frac{1}{2}\frac{d}{dt}\|\rho(t)-\rho'(t)\|^2 = \langle \dot{\rho}-\dot{\rho}', \rho-\rho'\rangle = -\langle \Phi'(\rho,t)-\Phi'(\rho',t), \rho-\rho'\rangle \leq 0 \tag{11}$$

where the convexity of $\Phi(\cdot,t)$ has been used. The contraction property (11) plays a crucial role when studying the asymptotic behaviour of solutions as $t \to +\infty$. If $\mathcal{L}(t)$ is a radial function of time, i.e. of the form

$$\mathcal{L}(t) = f(t)\mathcal{L}^0 \tag{12}$$

then a change of variables allows Eq. (9) to be cast in the form (10) [22]. In the general case when $\mathcal{L}(t)$ is not radial, then the properties of Eq. (9) differ significantly from those of Eq. (10) – In particular, the contraction property is lost.

In the following it will be useful to measure 'how much' $\mathcal{L}(t)$ (or $\mathcal{M}(t)$) differs from a radial function. To that purpose, set

$$\mu(t) = 1/\operatorname{tr}\mathcal{M}(t)\ , \mathcal{M}_0(t) = \mu(t)\mathcal{M}(t),$$

and define

$$\gamma(a,b) = \int_a^b |||\dot{\mathcal{M}}_0(t)|||\, dt \tag{13}$$

where $|||\cdot|||$ is the norm operator in $\mathcal{H}$, i.e. $|||\mathcal{M}(t)||| = \sup_{\rho' \in \mathcal{H}, \|\rho'\|=1} \|\mathcal{M}(t)\rho'\|$. Observe that $\gamma(a,v)$ vanishes if $\mathcal{L}(t)$ is of the form (12) on the time interval $[a,b]$, and is non negative otherwise.

Remark: For simplicity, we have assumed so far that the dissipation potential ϕ is differentiable, which is the case for viscoplasticity. For perfect plasticity, the dissipation potential is only subdifferentiable: Eqs. (9) and (10) have to be replaced with

$$-\frac{d}{dt}(\mathcal{L}(t)\rho) = \partial I_{\mathcal{K}(t)}(\rho,t) \tag{14}$$

and

$$-\frac{d\rho}{dt} = \partial I_{\mathcal{K}(t)}(\rho,t), \tag{15}$$

respectively. In Eqs (14-15), $I_{\mathcal{K}(t)}$ is the indicator function of the set $\mathcal{K}(t)$ and ∂ denotes the subdifferential operator. Eq. (15) corresponds to the so-called 'sweeping process' [24].

3 SHAKEDOWN THEOREM

In the following we study the asymptotic behavior of solutions to (9) as $t \to \infty$. We consider the case where $\mathcal{L}(t)$, $\mathcal{K}(t)$ are periodic in time (with the same period T) and assume that $\mathcal{K}_0(t)$ and $\mathcal{L}(t)$ are bounded by some constants denoted respectively by M and C, i.e.

$$\|\rho'\| \leq M \ , |||\mathcal{L}(t)||| \leq C \tag{16}$$

for all t and $\rho' \in \mathcal{K}_0(t)$. As mentioned in the introduction, shakedown corresponds to the situation where the evolution becomes elastic in the large time limit. Elastic evolutions are characterized by the fact that $\phi(\sigma(x,t),x,t) = 0$. Concerning the evolution equation (9) for the stress field, it follows that the distinctive properties of an elastic solution $\rho^*(t)$ are

$$\frac{d}{dt}(\mathcal{L}(t)\rho^*(t)) = 0, \ \rho^*(t) \in \mathcal{K}(t) \ \forall t \in [0,T]. \tag{17}$$

Shakedown corresponds to the situation where the residual stress converges towards an elastic solution as $t \to \infty$. Our main objective is to examine conditions under which *every* solution of (9) converges towards an elastic solution in the large time limit. More precisely, the main result we discuss is the following

Shakedown theorem: If
(i) There exists an elastic solution ρ^* to Eq. (9) and some $r > 0$ such that

$$\forall \rho' \in \mathcal{H}, \forall t \in [0,T], \ \|\rho' - \rho^*(t)\| \leq r \Longrightarrow \rho' \in \mathcal{K}(t);$$

(ii) the elastic moduli are such that

$$\frac{\gamma(0,T)}{\inf \mu} < \frac{r}{2CM};$$

then shakedown occurs, whatever the initial state is.

That theorem holds for elastic-perfectly plastic materials as well as for Duvaut-Lions elasto-viscoplastic materials (see Sect. 3.3). Condition (i) means that there exists an elastic solution that remains in the interior of the elasticity domain $\mathcal{K}(t)$: It is a reformulation of the Melan's condition found in the classical Melan's theorem. Condition (ii) sets restriction on the time variations of the elastic moduli. Setting such a restriction is necessary to obtain global convergence results: One can indeed find counterexamples in which condition (i) is fulfilled and some solutions of (9) do not become elastic in the large time limit [22].

3.1 Variation of the elastic energy

Let $\rho^*(t)$ be an elastic solution satisfying (i) and let $\rho(t)$ denotes an arbitrary solution of (9). We consider the positive function f defined as

$$f(t) = \frac{1}{2}\langle \tau(t), \mathcal{L}_0(t)\tau(t) \rangle. \tag{18}$$

where $\tau(t) = \mu(t)(\rho(t) - \rho^*(t))$. The function f is referred to as the elastic energy. The shakedown theorem stated above rests critically on the property

$$f(a+T) - f(a) \leq \big(2N\gamma(0,T) - r\inf\mu\big) \int_a^{a+T} \|\dot{\eta}(t)\| dt, \tag{19}$$

where

$$N = \sup_{a \le t \le a+T} \{\|\mathcal{L}(t)\rho(t)\|, \|\mathcal{L}(t)\rho^*(t)\|\}.$$

Eq. (19) bounds the variation of the elastic energy over a cycle. That bound plays the role of the contraction property that holds only in the case of time-independent elastic moduli. Observe in particular from (19) that $f(a+T)-f(a) \le 0$ if $\gamma(0,T)$ is sufficiently small.

The rest of this Section is devoted to a proof of Eq. (19). The inequality (19) is obtained by performing minor modifications to a similar result obtained in [22] for perfect plasticity, hence we only sketch the main steps of the proof and refer to [22] for details. Recalling that $d(\mathcal{L}(t)\rho^*(t))/dt = 0$ and setting $H(t) = -\frac{1}{2}\langle \tau(t), \dot{\mathcal{L}}_0(t)\tau(t)\rangle$, some straightforward manipulation shows that

$$\dot{f}(t) = \langle \tau(t), \dot{\eta}(t)\rangle + H(t)$$

where $\eta(t) = \mathcal{L}(t)\tau(t)$. Integrating on the time interval $[a, a+T]$ gives

$$f(a+T) - f(a) = \int_a^{a+T} \langle \tau(t), \dot{\eta}(t)\rangle \, dt + \int_a^{a+T} H(t)\, dt. \tag{20}$$

In the following, the two integrals in the right-hand side of Eq. (20) are bounded separately. Let us first consider the integral $\int_a^{a+T} \langle \tau(t), \dot{\eta}(t)\rangle \, dt$. Noting from (9) that $\dot{\eta}(t) = -\Phi'(\rho(t),t)$ and recalling that Φ is convex, positive and vanishes in $\mathcal{K}(t)$, we have

$$0 \le \Phi(\rho(t),t) \le \langle \dot{\eta}(t), \rho'(t) - \rho(t)\rangle \text{ for all } \rho' \in \mathcal{K}(t). \tag{21}$$

Condition (i) implies that $\rho^*(t) - r\dot{\eta}/\|\dot{\eta}\|$ is in $\mathcal{K}(t)$. Hence (21) gives $0 \le \langle \dot{\eta}(t), \rho^*(t) - r\frac{\dot{\eta}}{\|\dot{\eta}\|} - \rho(t)\rangle$ i.e. $\langle \dot{\eta}(t), \rho(t) - \rho^*(t)\rangle \le -r\|\dot{\eta}(t)\|$. Therefore we obtain

$$\int_a^{a+T} \langle \tau(t), \dot{\eta}(t)\rangle \, dt \le -r(\inf \mu) \int_a^{a+T} \|\dot{\eta}\| \, dt. \tag{22}$$

Now consider the integral $\int_a^{a+T} H(t)\, dt$. We have $\tau(t) = \mathcal{L}_0^{-1}(t)(\eta(t) - \eta(a) + s)$ where $s = \mathcal{L}(a)(\rho(a) - \rho^*(a))$ is independent on t. Some manipulations (see [22]) show that

$$\int_a^{a+T} H(t) dt \le \frac{1}{2} \int_a^{a+T} \|\eta(t) - \eta(a)\| \cdot |||\dot{\mathcal{M}}_0(t)||| \cdot \|\eta(t) - \eta(a) + 2s\| dt. \tag{23}$$

Since $\eta(t) - \eta(a) + 2s = \mathcal{L}(t)\tau(t) + \mathcal{L}(a)\tau(a) - 2\mathcal{L}(a)\rho^*(a)$, we have

$$\|\eta(t) - \eta(a) + 2s\| \le 4N. \tag{24}$$

We also have, for $t \in [a, a+T]$,

$$\|\eta(t) - \eta(a)\| = \|\int_a^t \dot{\eta}(t')dt'\| \le \int_a^t \|\dot{\eta}(t')\| dt' \le \int_a^{a+T} \|\dot{\eta}(t')\| dt'. \tag{25}$$

Substituting (24-25) in (23) and using the definition (13) of $\gamma(a,b)$, we obtain

$$\int_a^{a+T} H(t) dt \le 2N\gamma(a, a+T) \int_a^{a+T} \|\dot{\eta}(t)\| dt. \tag{26}$$

Replacing (22-26) in (19) and observing by periodicity that $\gamma(a, a+T) = \gamma(0,T)$, Eq. (19) is obtained.

3.2 Case of elastic-perfectly plastic materials

In perfect plasticity, the stress is constrained to remain in the elasticity domain $\mathscr{C}(x,t)$, which by (16) implies that $\|\rho(t)\| \leq M$ and $\|\rho^*(t)\| \leq M$ for all t. By (16) we thus have $\|\mathcal{L}(t)\rho(t)\| \leq CM$ and $\|\mathcal{L}(t)\rho^*(t)\| \leq CM$. In (19), the constant N thus satisfies $N \leq CM$. Applying Eq. (19) on the time interval $[iT,(i+1)T]$ gives

$$f((i+1)T) - f(iT) \leq -m \int_{iT}^{(i+1)T} \|\dot{\eta}(t)\|\, dt$$

where $m = r(\inf \mu) - 2CM\gamma(0,T)$ is non-negative by (ii). Summing over $i = 0, \cdots, I$ gives

$$\int_{0T}^{IT} \|\dot{\eta}(t)\|\, dt \leq \frac{1}{m} f(0).$$

Taking the limit $I \to \infty$ shows that the integral $\int_0^S \|\dot{\eta}(t)\| dt$ converges as $S \longrightarrow +\infty$. It follows that $\eta(t)$ also converges towards a limit η_∞ as $t \longrightarrow +\infty$. Setting $\rho_\infty(t) = \mathcal{M}(t)\eta_\infty$, it can be checked that $\rho(t) - \rho_\infty(t) \longrightarrow 0$ as $t \to \infty$, and that $\rho_\infty(t)$ is an elastic solution.

Remark: For perfect plasticity, the theorem stated above slightly improves on the result obtained in [22]. In [22], a condition analog to (ii) was formulated, with the factor 2 in (ii) replaced by a factor 3.

3.3 Case of elastic-viscoplastic materials

Compared to perfect plasticity, an additional difficulty is that the stress is not restricted to remain in the elasticity domain: There is no direct bound on $\rho(t)$, hence no direct on the constant N in (19) . That difficulty, however, can be overcome in the case of a Duvaut-Lions viscoplastic behavior, i.e. for a dissipation potential of the form

$$\phi(\sigma,x,t) = \frac{\alpha}{2} |\sigma - P_{\mathscr{C}(x,t)}\sigma|^2 \tag{27}$$

where $\alpha > 0$ is a viscosity parameter (assumed to be independent on (x,t) for simplicity) and $P_{\mathscr{C}(x,t)}$ denotes the projection on the elasticity domain $\mathscr{C}(x,t)$. Full details will be published elsewhere [25], but the main point is that $\eta(t) = \mathcal{L}(t)\rho(t)$ is bounded by a constant N' for large time. Moreover, that constant N' can be chosen as arbitrary close to CM. Using (19) with a value of N' such that $CM < N' < (r \inf \mu)/2\gamma(0,T)$, one can show that $\eta(t)$ converges as $t \to \infty$ and consequently that $\rho(t)$ converges towards an elastic solution (the reasoning is similar to that presented in Sect. 3.2).

4 EXAMPLE

As an illustrative example, consider the 3-bar truss represented in Fig. 1. We denote by σ_i, ε_i, ε_i^p the uniaxial stress, total strain and plastic strain in bar i. The constitutive relation in each bar reads as

$$\varepsilon_i = \frac{\sigma_i}{E_i(t)} + \varepsilon_i^p + \varepsilon_i^\theta, \dot{\varepsilon}_i^p = \phi'(\sigma_i).$$

The truss is stress-free, i.e. $\sigma_1 + \sigma_2 + \sigma_3 = 0$. The loading consists of a thermal strain $\varepsilon_2^\theta(t) = a\sin(2\pi t/T)$ in bar 2. We assume that $E_1 = E_3$ remains independent on time and that $E_2(t)$ vary as $E_2(t) = E/(1+\delta\sin^2(\pi t/T))$. The elasticity domain is the same for all bars and taken as $[-\sigma_y, \sigma_y]$ where the yield limit σ_y is independent on temperature.

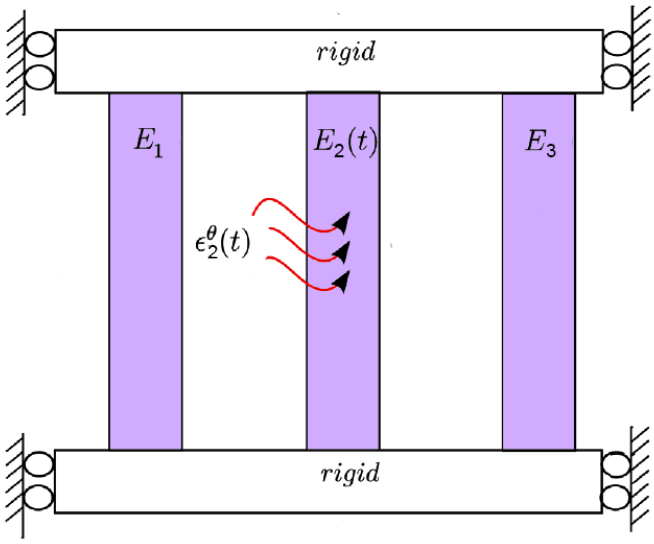


Figure 1: Three-bar truss with large thermal loading.

For the problem at hand, the vectorial space $\mathcal{E}$ of stress fields identifies with $\mathbb{R}^3$ and is equipped with the scalar product $\langle \sigma, \sigma' \rangle = \sum_{i=1}^3 \sigma_i \sigma_i'$. Writing the stress state of the system as $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, the fictitious elastic response of the system is

$$\sigma^E(t) = E \frac{\varepsilon_2^\theta(t)}{3 + 2\delta \sin^2(\pi t/T)} (1, -2, 1).$$

It follows that $\|\sigma^E(t)\| \leq \sqrt{6}\sigma^\theta/3$ where $\sigma^\theta = E \sup_t |\varepsilon_2^\theta(t)|$. Any σ in $\mathcal{K}^0$ verifies $\|\sigma + \sigma^E\| \leq \sqrt{3}\sigma_y$. Hence the constant M in (16) can be taken as

$$M = \sqrt{3}\sigma_y + \sqrt{6}\sigma^\theta/3.$$

In the present case, the space $\mathcal{H}$ of self-equilibrated stress fields is of dimension 2. Calculating the operator $\mathcal{L}(t)$ shows that

$$C = \frac{3 + 2\delta}{3E}, \gamma(0, T) = \frac{\delta}{3 + \delta}, \inf \mu = \frac{1}{2E}.$$

Note that $(0,0,0)$ is an elastic solution. As long as $\sigma^\theta/\sigma_y \leq 3/2$, it can be verified that $(0,0,0)$ satisfies condition (i) with a value or r given by

$$r = (\frac{\sigma_y}{2} - \frac{\sigma^\theta}{3})\sqrt{6}.$$

In Fig. 2 are represented the values (σ^θ, δ) satisfying conditions (i)-(ii): Shakedown occurs for any (σ^θ, δ) below the solid line shown in Fig. 2. For $\delta = 0$, we obtain that shakedown occurs provided that $\sigma^\theta \leq 3/2\sigma_y$, in agreement with the classical Melan's theorem for time-independent moduli. As can be observed on Fig. 2, the time fluctuations of the elastic moduli result in a decrease of the shakedown limit on σ^θ. Those results apply both to elastic-perfectly plastic materials and to elasto-viscoplastic material of the Duvaut-Lions type.

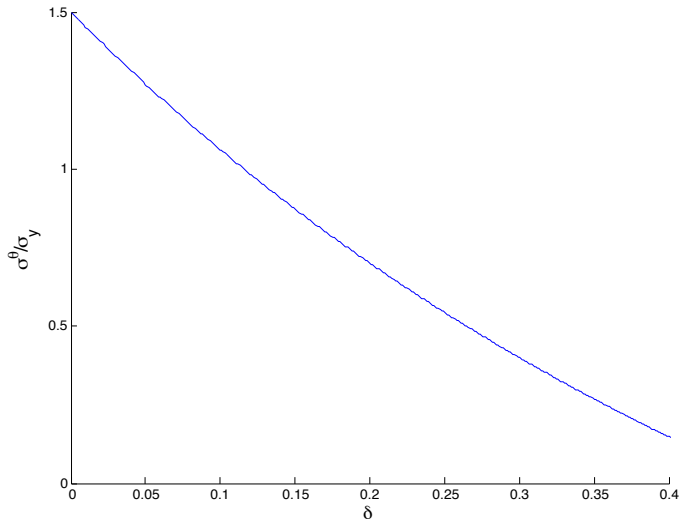


Figure 2: Shakedown limit for the three-bar truss.

In Fig. 3 are plotted the incremental evolution of the plastic strain and residual stress in bar 2, starting from an initial residual stress $\rho(0) = (0.46, -0.53, 0.07)E$. The plots on the left correspond to a elastic-perfectly plastic behavior, while the plots on the right correspond to a elastic-viscoplastic behavior. We set $\delta = 0.1$ and $\varepsilon^\theta(t) = 0.1(\sigma_y/E)\sin(2\pi t/T)$: Such values are below the shakedown limit shown in Fig. 2. As expected from the shakedown theorem, we observe the convergence of the plastic strain towards a stabilized value. In the case of perfect plasticity, that convergence is reached after the first loading cycle. In contrast,the convergence is slow in the case of viscoplasticity. Observe that the residual stress does not converge towards a stabilized value: This is a direct consequence of the time dependence of the elastic moduli.

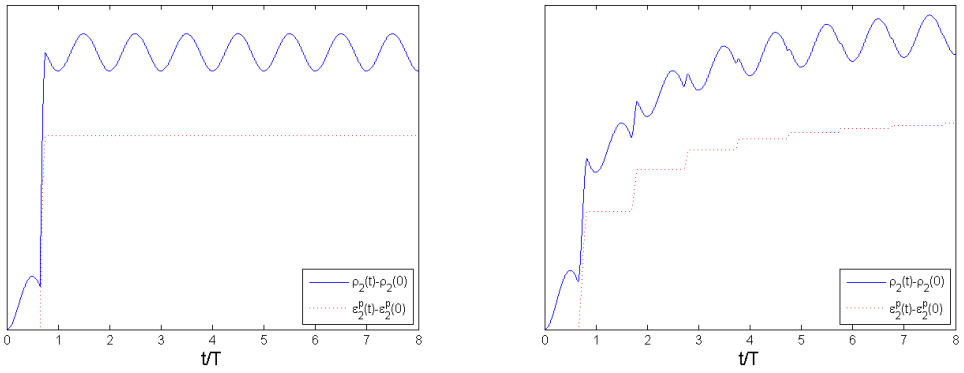


Figure 3: Residual stress and plastic strain in bar 2: (left) elastic-perfectly plastic material; (right) elastic-viscoplastic material

REFERENCES

[1] E. Melan, Theorie statisch unbestimmter systeme aus ideal-plastischen baustoff, Sitz.Berl.Ak.Wiss. 145 (1936) 195–218.

[2] P. S. Symonds, Shakedown in continuous media, J. Appl. Mech. 18 (1951) 85–89.

[3] W. T. Koiter, General problems for elastic solids, Progress in solid mechanics (1960).

[4] D. C. Pham, On shakedown theory for elastic-plastic materials and extensions, Journal of the Mechanics and Physics of Solids 56 (2008) 1905–1915.

[5] Q. S. Nguyen, On shakedown analysis in hardening plasticity, Journal of the Mechanics and Physics of Solids 51 (2003) 101–125.

[6] Y. J. Ahn, E. Bertocchi, J. Barber, Shakedown of coupled two-dimensional discrete frictional systems, Journal of the Mechanics and Physics of Solids 56 (12) (2008) 3433–3440.

[7] M. Peigney, Shakedown theorems and asymptotic behaviour of solids in non-smooth mechanics, Eur. J. Mech. A 29 (5) (2010) 784–793.

[8] M. Peigney, On shakedown of shape memory alloys structures, Annals of Solid and Structural Mechanics 6 (2014) 17–28.

[9] D. Weichert, A. Ponter, A historical view on shakedown theory, in: E. Stein (Ed.), The History of Theoretical, Material and Computational Mechanics-Mathematics Meets Mechanics and Engineering, Springer, 2014, pp. 169–193.

[10] J. Zarka, J. Frelat, G. Inglebert, A new approach to inelastic analysis of structures, Martinus Nijhoff Publishers, Dordrecht (1988).

[11] H. Maitournam, B. Pommier, J. Thomas, Détermination de la réponse asymptotique d'une structure anélastique sous chargement cyclique, C.R. Mecanique 330 (2002) 703–708.

[12] V. Carvelli, Z. Cen, Y. Liu, G. Maier, Shakedown analysis of defective pressure vessels by a kinematic approach, Archive of Applied Mechanics 69 (9-10) (1999) 751–764.

[13] M. Peigney, C. Stolz, Approche par contrôle optimal des structures élastoviscoplastiques sous chargement cyclique, C.R.Acad.Sci. Paris, II 329 (2001) 643–648.

[14] M. Peigney, C. Stolz, An optimal control approach to the analysis of inelastic structures under cyclic loading, Journal of the Mechanics and Physics of Solids 51 (2003) 575–605.

[15] J.-W. Simon, D. Weichert, Shakedown analysis of engineering structures with limited kinematical hardening, International Journal of Solids and Structures 49 (15) (2012) 2177–2186.

[16] K. V. Spiliopoulos, K. D. Panagiotou, A direct method to predict cyclic steady states of elastoplastic structures, Computer Methods in Applied Mechanics and Engineering 223 (2012) 186–198.

[17] G. Borino, Consistent shakedown theorems for materials with temperature dependent yield functions, International Journal of Solids and Structures 37 (22) (2000) 3121–3147.

[18] J. König, A shakedown theorem for temperature dependent elastic moduli, Bull. Ac. Pol. Sci., Ser. Sci. Techn. 17 (3) (1969) 161–165.

[19] B. Halphen, S. di Domizio, Evolution des structures élastoplastiques dont les coefficients d'élasticité dépendent de la température, 17ème congrès français de mécanique, Troyes (2005).

[20] S. Hasbroucq, A. Oueslati, G. De Saxcé, Inelastic responses of a two-bar system with temperature-dependent elastic modulus under cyclic thermomechanical loadings, International Journal of Solids and Structures 47 (14) (2010) 1924–1932.

[21] S. Hasbroucq, A. Oueslati, G. De Saxcé, Analytical study of the asymptotic behavior of a thin plate with temperature-dependent elastic modulus under cyclic thermomechanical loadings, Int. J. Mech. Sci. 54 (1) (2012) 95–104.

[22] M. Peigney, Shakedown of elastic-perfectly plastic materials with temperature-dependent elastic moduli, Journal of the Mechanics and Physics of Solids 71 (2014) 112–131.

[23] B. Halphen, Elastic perfectly plastic structures with temperature dependent elastic coefficients, C.R. Mécanique 333 (8) (2005) 617–621.

[24] J. J. Moreau, Evolution problem associated with a moving convex set in a hilbert space, J. Diff. Eq. 26 (3) (1977) 347–374.

[25] M. Peigney, On Melan's theorem in temperature-dependent viscoplasticity, to appear (2016).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A NUMERICAL STUDY ON THE ENDURANCE LIMIT OF PARTICULATE REINFORCED METAL MATRIX COMPOSITES (PRMMCS) USING THE DIRECT METHOD AND THE STATISTICAL LEARNING

Geng Chen[1], Bezold Alexander[1], Christoph Broeckmann[1], and Dieter Weichert[2]

[1] Institute for Materials Applications in Mechanical Engineering, RWTH Aachen University
Augustinerbach 4, Aachen, Germany
e-mail: {g.chen, a.bezold, c.broeckmann}@iwm.rwth-aachen.de

[2] Institute of General Mechanics, RWTH Aachen University
Templergraben 64, Aachen, Germany
e-mail: weichert@iam.rwth-aachen.de

Keywords: Direct Method, RVE, Metal Matrix Composite (MMC), Statistics, Classification

Abstract. *Particulate reinforced metal matrix composites (PRMMCs) are characterized by their stochastic and irregular microstructure. The strength of the material has been noticed to have a large scatter and great dependence on the underlying composite structure. In order to understand how size and distribution of the reinforcement particles contribute to the effective material behaviors, we elaborated in present study a numerical approach which incorporates homogenization technique, shakedown analysis, and statistical learning. To demonstrate this approach, a typical PRMMC material, WC - 20 wt.% Co, was taken as example and numerous representative volume element (RVE) samples of this material were built based on both real and artificial microscope images. Multiple effective behaviors, including the ultimate strength and endurance limit, were evaluated on each RVE sample using the direct method. In order to understand how multiple structural factors jointly affect the overall composite strength, the predicted strength and other selected features extracted from the samples are submitted to few statistical models such as logistic regression and artificial neural network (ANN) to establish predictive models. On the basis of these models we discussed how few structural factors, which have been identified to have nontrivial impact, contribute to the endurance limit of the selected PRMMC material.*

1 INTRODUCTION

Particulate reinforced metal matrix composites (PRMMCs) are becoming increasingly popular for being used as structural materials due to their preferable mechanical properties such as high strength and good wear resistance. Compared to fiber reinforced composites, mechanical behaviors of PRMMCs are more isotropic, and the materials are in general less expensive to be fabricated. As composed of metal matrix and discrete ceramic particles, the microstructure of the PRMMCs are irregular and highly randomized. Therefore, the mechanical behaviors of the material, including the strength, have been noticed to have a large scatter and great dependence on the underlying composite structure. In order to enhance the economic use of the material and accelerate the development period of new PRMMCs compositions, it is essential to have an engineering tool for predicting the strength of materials from their microstructural morphologies and identifying the reasons contributing to the material strength; and this is the objective of the present paper.

One common practice for determining the macroscopic behaviors of composite materials is through numerical homogenization. According to this approach, finite element (FE) models of the composite structure, the so called representative volume element (RVE), should be built; then by applying loads on it and deriving the effective terms of its responses, the global behaviors of the composite can be obtained. Beyond determining the mechanical behaviors such as elasticity, numerical homogenization has been used also for determining the strength of the composite materials. To do this, one viable way is to apply the direct method (DM) to RVE models and converting the obtained results to the macroscopic strengths. Previously, Weichert et al. [1] have elucidated in their studies the mathematical formulation of this approach based on the static theorem, while Chen H. and Pontor [2] proposed a formulation serving the same aim but based on the kinematic theorem. By applying these methods to composites reinforced by periodically aligned fibers, the in-plane strength of the material have been successfully calculated. The availability of this technique has been extended in our previous studies to random heterogeneous materials [3]. For random heterogeneous materials such as PRMMCs, one intrinsic difficulty is to represent the material by RVE models. To have a model which smears out the influence raised by the random size and the distribution of the reinforcement particles, it requires the size of the model to be exceptionally large and this will cause sever computational difficulty. One feasible remedy to this problem is, instead of using one RVE, employing many RVEs which are considered to be statistically equivalent. In the light of this idea, an in-house code was developed for simplify the modeling of RVE samples by automatically converting SEM images to adaptively meshed RVE models. Taking advantage of this code, we performed DM on many samples converted from SEM image of a typical PRMMC material, WC - 20 wt.% Co, and obtained the statistics of their macroscopic ultimate strengths and endurance limits.

As has been addressed, the interests of studying the strength of PRMMCs is not only confined to determining the strength; indeed, it is also important to have an insight of factors contributing to the strength. To expose the relationship between the composite structure and the material strength, one advisable solution is through statistical learning. Statistical learning is a fast evolving discipline in recent years. Many algorithms arisen from this field have been successfully implemented for solving challenging medical and engineering problems. The problems of statistical learning can be coarsely categorized to supervised and unsupervised. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations among a set of input measures. Depending on if the outputs are qualitative or

quantitative, problems of the supervised learning can be further subdivided to regression and classification (or pattern recognition). Although problems belonging to unsupervised learning prevail in engineering practices, most of the studies dedicated to applying statistical learning to engineering problems focus on the supervised learning. Some achievements made in this respect can be seen from monograph [4] and review articles [5, 6, 7]. In the scope of composite structures and their mechanical behaviors, recently the studies where statistical learning is employed as the main approach increase in both number and the diversity [8, 9]. Among these studies, the application of learning models to either experimental or simulated data gradually becomes a new paradigm for results interpretation and knowledge discovery.
In the present study, beside determining the strength of a PRMMC material, WC - 20 wt.% Co, one additional objective is to understand how multiple structural factors jointly affect the overall composite endurance limit. To answer this question, we constructed predictive models based on both logistic regression and back propagation artificial neural network (ANN). Based on analyzing results of direct method using these models, the relationship between endurance limit and selected micromechanical measurements will be discussed.

2 SHAKEDOWN OF RVE MODELS

To study the effective behaviors of heterogeneous materials, the material is reflected in two scales: the microscopic scale is small enough for the heterogeneities to be identified. The macroscopic scale is large enough for the heterogeneities to be expelled. The two scales are well-separated and they are described by two coordinate systems: the global coordinate system $\boldsymbol{x}$ and local coordinate system $\boldsymbol{y}$. The following relationship holds

$$\boldsymbol{y} = \frac{\boldsymbol{x}}{\epsilon} \,. \tag{1}$$

ϵ is a small scale parameter which determines the size of the RVE.
For a heterogeneous meterial, once submitted to an external loading, its microscopic stress field $\boldsymbol{\sigma}$ in $\boldsymbol{y}$ and its macroscopic counterpart $\boldsymbol{\Sigma}$ satisfy the relationship

$$\boldsymbol{\Sigma} = \frac{1}{\Omega} \int_{\Omega} \boldsymbol{\sigma}(y) dV = \langle \boldsymbol{\sigma}(\boldsymbol{y}) \rangle \,. \tag{2}$$

Here $\langle \cdot \rangle$ stands for the averaging operator, and Ω indicates the RVE domain. Similarly, the relationship between strain measures satisfies

$$\boldsymbol{E} = \frac{1}{\Omega} \int_{\Omega} \boldsymbol{\varepsilon}(y) dV = \langle \boldsymbol{\varepsilon}(\boldsymbol{y}) \rangle \,. \tag{3}$$

The local strain $\boldsymbol{\varepsilon}$ can be decomposed into two parts: The average value $\boldsymbol{E}$ and a fluctuating part $\boldsymbol{\varepsilon}^*$

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = \boldsymbol{E} + \boldsymbol{\varepsilon}^* \,. \tag{4}$$

When all constituents of a RVE are elastic, the overall behavior of the RVE is elastic as well. In this circumstance, $\boldsymbol{\Sigma}$ and $\boldsymbol{E}$ are correlated by an effective elastic tensor $\overline{\mathbb{C}}$

$$\boldsymbol{\Sigma} = \overline{\mathbb{C}} : \boldsymbol{E}. \tag{5}$$

In case that the heterogeneous material to be considered behaves isotropically in the macro scale, same to the single phase material, $\overline{\mathbb{C}}$ can be uniquely determined from two elastic constants, such as effective Young's modulus $\bar{E}$ and effective Poission's ratio $\bar{\nu}$.

When the composite material is composed of elasto-plastic constituents, its macroscopic ultimate strength Σ_U and endurance limit Σ_∞, which correspond to plastic and shakedown limit in the micro scale, can be studied by incorporating homogenization with direct method. As formulated by Magoariec et al. [10], when the shakedown state is attained in the micro scale, stress field pertained to the reference elastic body $\mathcal{B}^E$, $\boldsymbol{\sigma}^e$, and the time invariant residual stress field $\bar{\boldsymbol{\rho}}$ are required to satisfy following conditions

$$\boldsymbol{\sigma}^e : \begin{cases} \boldsymbol{\nabla} \cdot \boldsymbol{\sigma}^e = 0 & \text{in } \Omega \,, \\ \boldsymbol{\sigma}^e = \mathbb{C} : (\boldsymbol{E} + \boldsymbol{\varepsilon}^*) & \text{in } \Omega \,, \\ \boldsymbol{\sigma}^e \cdot \boldsymbol{n} & \text{anti-periodic on } \partial\Omega \,, \\ \boldsymbol{u}^* & \text{periodic on } \partial\Omega \,, \\ \langle \boldsymbol{\varepsilon} \rangle = \boldsymbol{E} \,. \end{cases} \tag{6}$$

$$\bar{\boldsymbol{\rho}} : \begin{cases} \boldsymbol{\nabla} \cdot \bar{\boldsymbol{\rho}} = 0 & \text{in } \Omega \,, \\ \bar{\boldsymbol{\rho}} \cdot \boldsymbol{n} & \text{anti-periodic on } \partial\Omega \,. \end{cases} \tag{7}$$

Here, Ω indicates the RVE domain, $\partial\Omega$ the surface, $\boldsymbol{n}$ the outer normal, and $\boldsymbol{u}^*$ the fluctuation part of the displacement corresponds to $\boldsymbol{\varepsilon}^*$.
Although shakedown problem in the RVE scale can be studied by either strain or stress approach [10], in present study we consider exclusively the stress approach. For stress approach the load prescribed on RVE is the macroscopic stress $\boldsymbol{\Sigma}$. Because the material to be studied is non-periodic, a small specification is made on conditions (6) and (7), where, instead of enforcing the node-wise anti-periodicity of the residual stresses and periodicity of the fluctuating displacement, we apply the statically uniform boundary conditions (SUBC) on the purely elastic reference RVE. As a consequence, the shakedown problem yields $\bar{\boldsymbol{\rho}} \cdot \boldsymbol{n} = \boldsymbol{0}$ on $\partial\Omega$ and one can easily prove that, in the absence of the body force $\langle \bar{\boldsymbol{\rho}} \rangle = \boldsymbol{0}$, so $\bar{\boldsymbol{\rho}}$ does not contribute to the macroscopic stress.
By discretizing the physical fields in (6) and (7) by means of the FE formulations, the application of the static theorem to RVEs composed of elastic perfectly plastic materials leads to following optimization problem:

$$\begin{aligned} \underset{\bar{\boldsymbol{\rho}},\alpha}{\text{minimize}} \quad & -\alpha \\ \text{subject to} \quad & \sum_{i=1}^{NG} \boldsymbol{C}_i \, \bar{\boldsymbol{\rho}}_i = \boldsymbol{0} \,, \\ & F(\alpha \boldsymbol{\sigma}^e_{ik} + \bar{\boldsymbol{\rho}}_i) - \sigma^2_{Y\,i} \leq 0 \\ & \forall i \in [1, NG] \,; k \in [1, NV] \,. \end{aligned} \tag{8}$$

Here, α is referred to as the load factor, $\boldsymbol{C}$ the equilibrium matrix, $\bar{\boldsymbol{\rho}}_i$ the stress tensor associated with the ith Gaussian point, $\boldsymbol{\sigma}^e_{ik}$ the abbreviation of $\boldsymbol{\sigma}^e_i(\boldsymbol{P}_k)$ which means the $\boldsymbol{\sigma}^e$ at Gaussian point i and load vertex k, σ_Y the yield strength, F the yield function, NG the number of Gaussian points, and NV the number of vertices. Solving (8) yields the load capacity of the RVE, and depending on if $k = 1$ or $k > 1$ the calculated strength corresponds to either plastic limit or endurance limit. In the present study, the load scenario considered is restricted to non-reversed uniaxial stress, in this case $NV = 2$ and $\boldsymbol{\sigma}^e_{ik} = 0$ for all $k = 2$.

3 CLASSIFICATION MODELS

After solving (8) pertained to each RVE and obtaining $\Sigma_\infty = \alpha \cdot \Sigma$, the results are fed to statistical models for gaining a better interpretation. The statistical models considered in the present study are classification models. There are many advantages for having classification models: First, classification models are robust; our observations convinced that these models are insensitive to RVE size and sample size. Next, classification models have great practical values. Because material designers often face the problem of comparing between different composite structures, it is beneficial to have simple and inexpensive models for evaluating strength in an relative sense. However, it is worthy to note, that the purpose for introducing the predictive statistical model is not to replace the DM calculation. In fact, the study addresses the use of predictive model as a mean to understand the data; in other words, the interpretability of the model is accentuated compared to its predictivity.
We consider two classification models: the logistic regression and the ANN, both of them determine from a large set of observations $(X_1, Y_1), (X_2, Y_2), \ldots (X_m, Y_m)$ by which Y_m are categorical variables, to which class Y_i a new observation X_i would belong to at its highest probability. Although both logistic regression and ANN serve the same goal, in comparison the former one is convenient to be explained but lacks sufficient flexibility in prediction, while the latter one renders excellent prediction but is hard to be interpreted. The algorithms of both models, as well as the strategies adopted to avoid overfitting, are far beyond the scope of the present paper. More detailed explanations on these issues can be found in [11].

4 FEATURE SELECTION FOR THE STATISTICAL MODELS

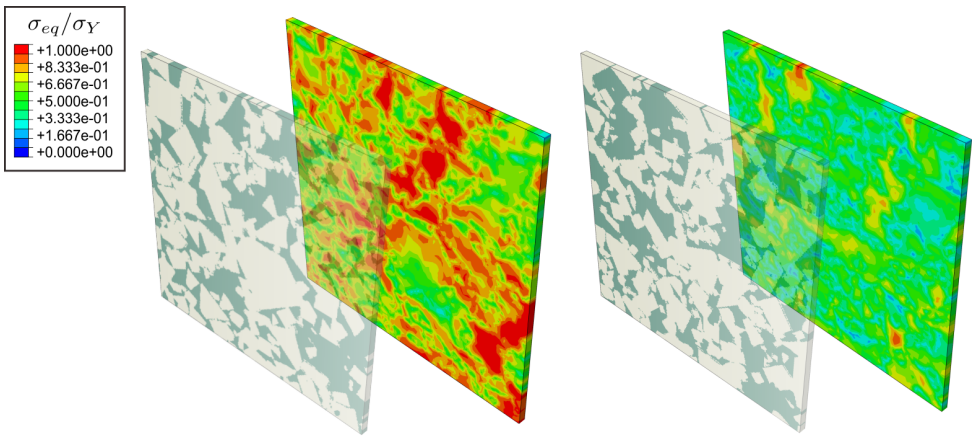
A key factor for having a well performed predictive model is that features X are properly selected. In the present case, the simplest way to select X is to use RVE images directly as features. Although this approach prevails in the computer science community, it is not adopted because such feature is physically less interpretable. On selecting features for the statistical model, the first set of features adopted are Vol.% WC, $\bar{E}$, $\bar{\nu}$, as well as peak valued equivalent elastic stress in both phases $\sigma^e_{WC,max}$ and $\sigma^e_{Co,max}$. Additionally, as shown in Figure 1, we noticed from few samples that Σ_∞ is related to the distribution of the stress and thus features reflecting the distribution characteristics of the stress field should be incorporated into the model. Because total stress $\boldsymbol{\sigma} = \alpha\boldsymbol{\sigma}^e + \bar{\boldsymbol{\rho}}$ is unknown before solving the shakedown problem, features were extracted from the elastic stresses field instead.
The features chosen for representing the distribution characteristics of the stress field is highly stressed volume fraction (V_{eff}) defined as

$$V_{eff}^{n\%\sigma} = \frac{\sum_{i=1}^{NG} V_i \cdot \mathcal{H}(\sigma_i, n\%\sigma^e_{\max})}{\sum_{i=1}^{NG} V_i} \,. \tag{9}$$

Here V_i is the volume and σ_i the equivalent stress at Gaussian point i; $\mathcal{H}$ the Heaviside function defined as

$$\mathcal{H}(\sigma_i, n\%\sigma^e_{\max}) = \begin{cases} 1, & \sigma_i \geq n\%\sigma^e_{\max} \,, \\ 0, & \text{Oherwise} \,. \end{cases} \tag{10}$$

One can see from (9), that V^n_{eff} sums over volumes whose stress exceeds the threshold value $n\%\sigma^e_{\max}$, therefore by adjusting the magnitude of n the stressed volume at each level can be easily accessed. In the present study, $V_{eff}^{n\%\sigma}$ was evaluated for each phase separately, where



(a) $Z_{\Sigma_\infty} = 1.50$ (b) $Z_{\Sigma_\infty} = -1.49$

Figure 1: Endurance limit and highly stressed volume

$\sigma^e_{\max}$ in (10) refers to $\sigma^e_{WC,max}$ or $\sigma^e_{Co,max}$, and calculation renders $V^{n\%\sigma}_{Co,eff}$ or $V^{n\%\sigma}_{WC,eff}$. The calculation fixed n to multiple values, where $n \in \{10, 20, \ldots, 90\}$ and results were collected into a vector.
In summary, the features selected for the predictive models can be divided into two categories:

- Category 1 (Homogenized material parameter): WC Vol.%, $\bar{E}$, $\bar{\nu}$
- Category 2 (Results of elastic simulation): $\sigma^e_{\max}$, $V^{n\%\sigma}_{eff}$

5 RESULTS AND DISCUSSIONS

The numerical study is based on 500 40 μm- 40 μm- 1μm RVE samples. Each sample is modeled from an individual artificial morphology generated by a simple random sequential adsorption (RSA) algorithm as shown in Figure 2. The algorithm is developed in Matlab [12] on the matrix basis. According to this algorithm, the RVE domain is initialized as a zero matrix and the program continuously projects prism shaped geometry into this matrix. After each projection, zero elements in the matrix are set to one if they belong to the prism domain and remain zero otherwise. The value of elements will not be reset if they have already been picked in previous iterations. Parameters controlling the projection, such as prism size, rotation angle, and center of the projection, are all random numbers. In order to be consistent with real WC-Co microstructures, the algorithm adopts a configuration that d_{WC} obeys a normal distribution with mean value $3\mu m$ and standard deviation $0.8\mu m$. The position where each particle locates is independent from the others and therefore there is no predefined clustering.
The models were built in commercial FE solver ABAQUS [13] and meshed by a uniform mesh configuration: the element type is fixed to linear wedge elements (C3D6); elements covering non-critical regions were assigned with a global size of 0.8m; while elements near the phase boundaries are of a finer density with an edge size of 0.2m. Under this configuration, the number of elements for each RVE sample varies between 15,000-20,000. In all models, both constituents are considered as elastic-perfect plastic materials with parameters given in Table 5. Meanwhile, the study is restricted to small deformation and assumes plastic failure as the only failure mechanism.
RVEs are prescribed by a global uni-axial stress $\Sigma = 10$ MPa for calculating $\boldsymbol{\sigma}^e$. After $\boldsymbol{\sigma}^e$ has been calculated, results were output to MATLAB together with geometrical set up to construct

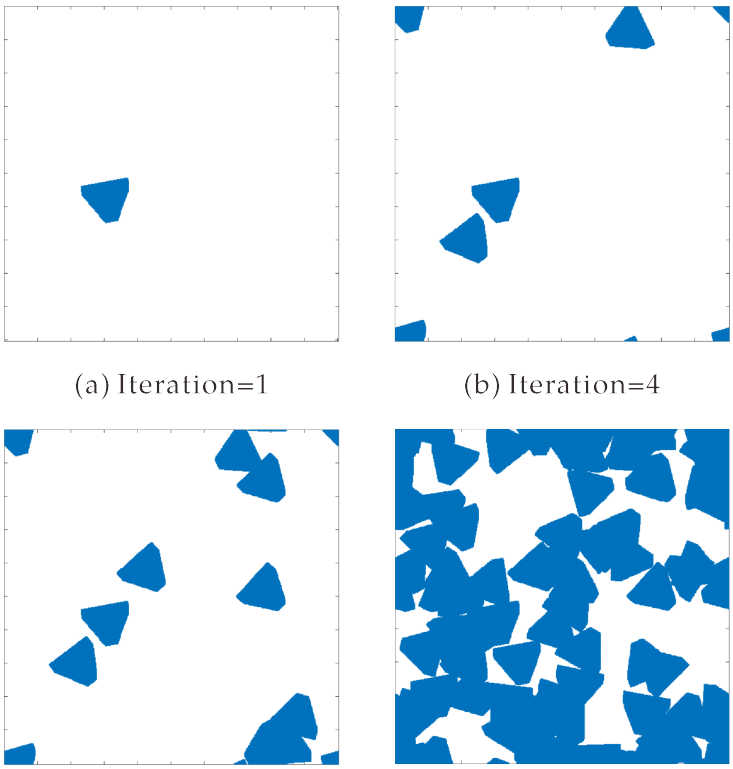


Figure 2: Inclusion process with fixed grain size

	E [GPa]	ν [−]	σ_Y [MPa]
WC	700	0.24	2000
Co	210	0.30	683

Table 1: Material properties of both phases

optimization problem. Once these optimization problems were built and solved by an interior point method based solver Gurobi [14], the solutions were further used for building classification models. Two classification models used, namely the logistic regression and the ANN, were realized in the statistics and machine learning toolbox of the MATLAB. In the logistic regression model, input data are linear, squared, and interaction terms of aforementioned features. In the ANN classifier, model parameters are determined based on evidences gained from both parametric study and comparison of the cross validation errors between different model settings. In the end, a single hidden layer model consisted of 20 neurons in the hidden layer has been chosen as the ANN structure.
The criterion adopted in classification models are Σ_∞ where the Z score is defined as

$$Z = \frac{x - \bar{x}}{s} \, . \tag{11}$$

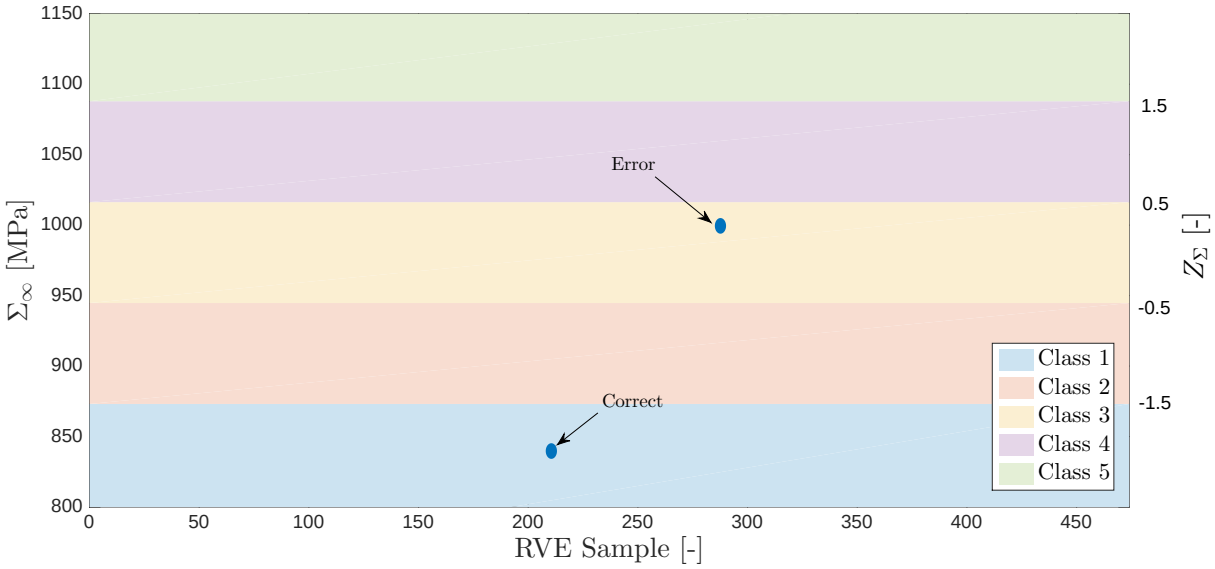


Figure 3: Domain of classes for the classification problems

Here $\bar{x}$ and s indicate the first and second moment of the sample set, respectively. According to Z score, the entire span of Σ_∞ is sub-divided into 4 or 5 classes as follows

$$\text{4 Classes} \begin{cases} \text{Class 1}: \ Z_{\Sigma_\infty} < -1.0 \\ \text{Class 2}: \ -1.0 < Z_{\Sigma_\infty} \leq 0 \\ \text{Class 3}: \ 0 < Z_{\Sigma_\infty} \leq 1.0 \\ \text{Class 4}: \ Z_{\Sigma_\infty} > 1.0 \end{cases} \tag{12}$$

$$\text{5 Classes} \begin{cases} \text{Class 1}: \ Z_{\Sigma_\infty} < -1.5 \\ \text{Class 2}: \ -1.5 < Z_{\Sigma_\infty} \leq -0.5 \\ \text{Class 3}: \ -0.5 < Z_{\Sigma_\infty} \leq 0.5 \\ \text{Class 4}: \ 0.5 < Z_{\Sigma_\infty} \leq 1.5 \\ \text{Class 5}: \ Z_{\Sigma_\infty} > 1.5 \end{cases} \tag{13}$$

Herein, the five-classes problem, as schematically presented in Figure 3, is only considered upon the success of the four-classes problem.
In all predictive models the feature x is normalized as follows before submitted to a learning algorithm

$$x_i' = 2 \cdot \frac{x_i - \min(x)}{\max(x) - \min(x)} - 1 \ . \tag{14}$$

This way, the normalized feature $x_i' \in [-1, 1]$, and the model will not face the numerical difficulty caused by the poor scaling.
To demonstrate the result of classification, we plot the sample point in the position where it actually locates with the color corresponding to the predicted class. When decision of a classification model is correct, then the color of a sample overlaps with its background color (see Figure 3). The wrong classification means the strength of a sample is either over- or underestimated. The overestimation of the strength is refereed to as false positive (FP), while underestimation as false negative (FN). Comparatively, FP is riskier than FN. According to the definition (12-13), the classes are aligned continuously and thus a certain amount of the classification tolerance should be allowed. In other words, if a sample lies closely enough to the classification boundary, then the wrong classification of the sample to the adjacent class should not be counted in

the error metrics. To exclude this particular typed error from the error metrics, we set tolerance to $1/10 \cdot s$ and employed modified error rate as the measurement of the classification error.
In the course of building statistical models for classifying Σ_∞, we noticed that the highly stressed volume fraction $V_{eff}^{n\%\sigma}$ plays a critical role. When the features in Category 2 were omitted, the results of classification is similar to Figure 4 where the overall error rate is 40.5%. Also, there is hardly any improvement in performance when model parameters were tuned or an alternative learning algorithm was used. The result in Figure 4 shows that the model without features in Category 2 fails to classify a sample into a correct strength class.

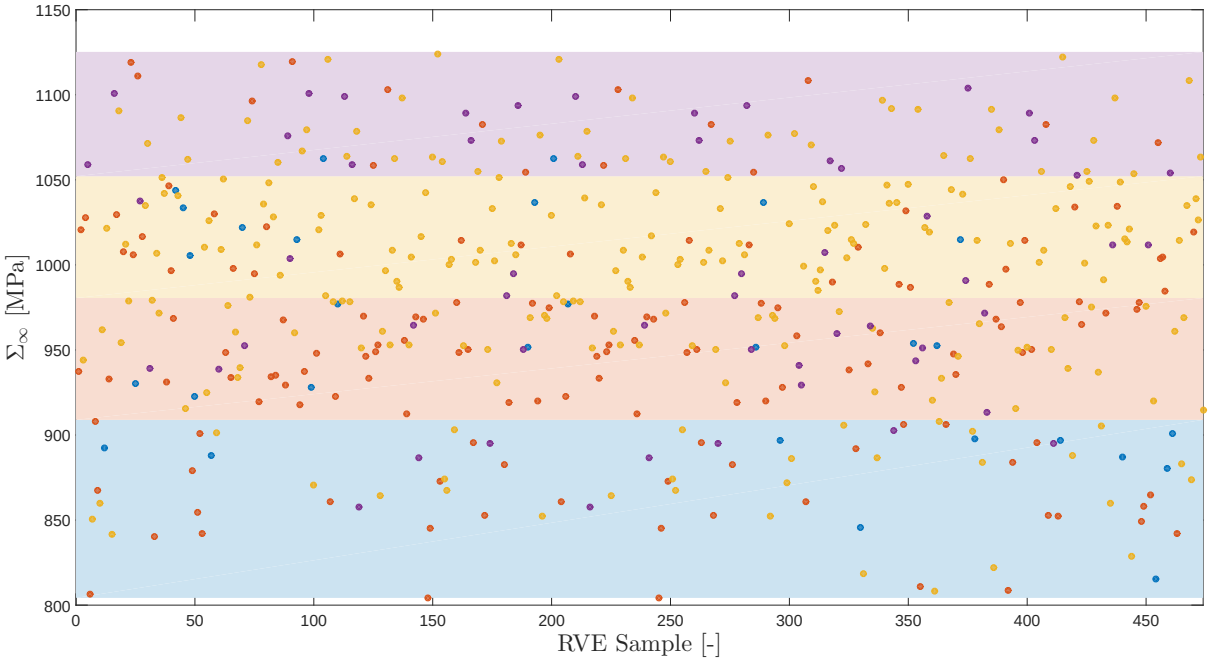


Figure 4: Results of an ANN classification model on Σ_∞ without features in Category 2

After the features in Category 2 were introduced back to the model, we observed that even without $\sigma^e_{\max}$, the accuracy of an ANN classifier was significantly boosted to more than 90% in the sense of modified error rate. Furthermore, by complete the features with $\sigma^e_{\max}$, the accuracy was further increased for around 4% percent and the final result is given in Figure 5. For a better demonstration of the result, we plot samples that have been mistakenly classified with a deviation greater than the tolerance in Figure 6. One can intuitively see from this figure that there are only few errors and all of them are quantitative instead of qualitative. In other words, the confusion occurs only between two neighboring classes. The pattern of classification errors can be seen clearer from the confusion matrix in Figure 7 where all non-zero elements are inside the diagonally bordered band. Based on the confusion matrix, it is justified to conclude that the classifier demonstrates an excellent performance and the pattern of the data have been exploited.
The generality of the ANN classifier has been confirmed through applying it to the test data obtained in another load direction. According to the confusion matrix given in Figure 8, the degree of over-fitting is insignificant. Because the classifier demonstrates sound performance in handling the present classification problem, we established a new ANN in an analogous manner for solving the five-classes classification problem. As expected the performance of the new classifier is satisfactory (see Figure 9), its overall correct rate is 95.8%, and correct rate of the

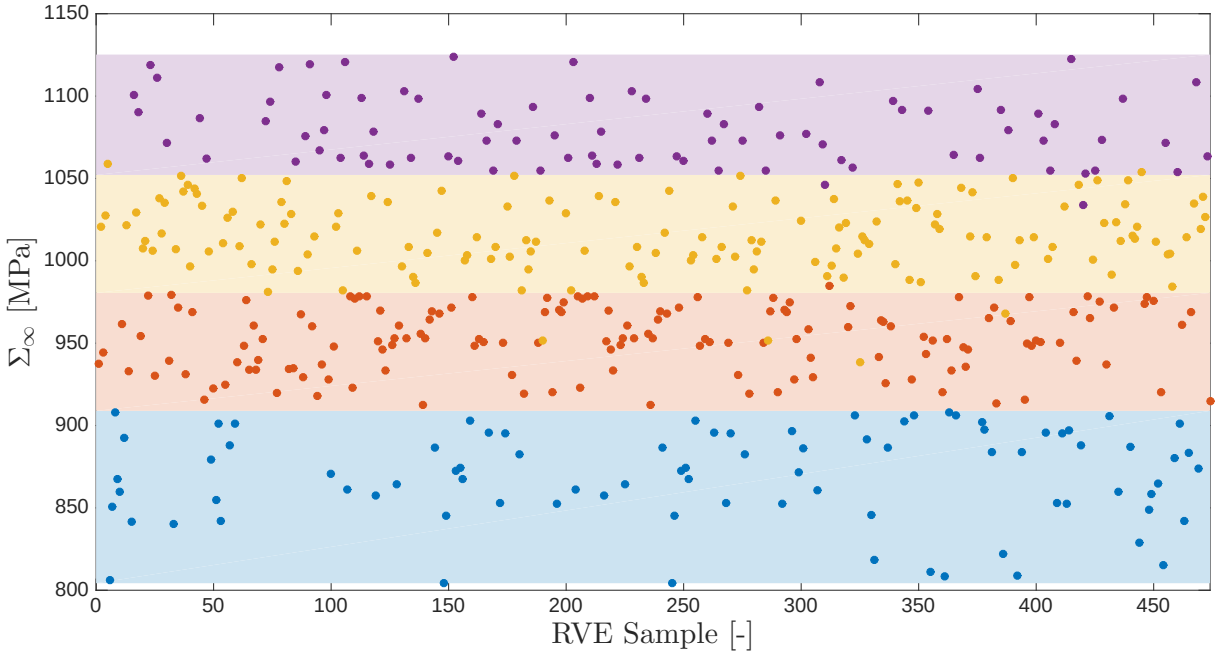


Figure 5: Results of an ANN classification model on Σ_∞ with features in Category 2

test set is 93.7%. If we set the classification tolerance to 1/5 s, then these two numbers will increase to 97.0% and 95.9%, respectively.
Although results in Figure 7 were retrieved from an ANN classifier, the accuracy of the classifier built by logistic regression is in the same order where the overall correct rate is about 94.1%, and more than 80 percent of the errors are confusions between two neighboring classes. In addition to that, the over-fitting problem of the logistic regression classifier is insignificant as well. The success of the logistic regression classifier confirms that characteristics of the endurance limit can be well prescribed by linear, interaction, and the squared terms of all considered features.
In summary, the strength class of Σ_∞ can be accurately predicted from RVE samples by features considered. Among them, the features in Category 2 play a particularly significant role and can fundamentally improve the performance of the classifier.

6 CONCLUSIONS

This paper elaborates how to apply shakedown analysis to predict the strength of random PRMMC materials and using statistical models to interpret the results. The study reveals that the endurance limit of the PRMMC material, WC- 20 Wt.% Co, is strongly dependent on the evenness of the effective stress in the binder, and an increased highly stressed volume fraction will lead to a greater endurance limit. The study confirms the benefit of incorporating direct method and statistical learning for studying the strength of random heterogeneous materials. Lastly, one should notice that the failure mechanism considered in the present study is restricted to the plastic failure, more complex failure scenarios, such as plastic damage of the binder phase of the brittle fracture of the reinforcement particles, are not considered and should be investigated in the future.

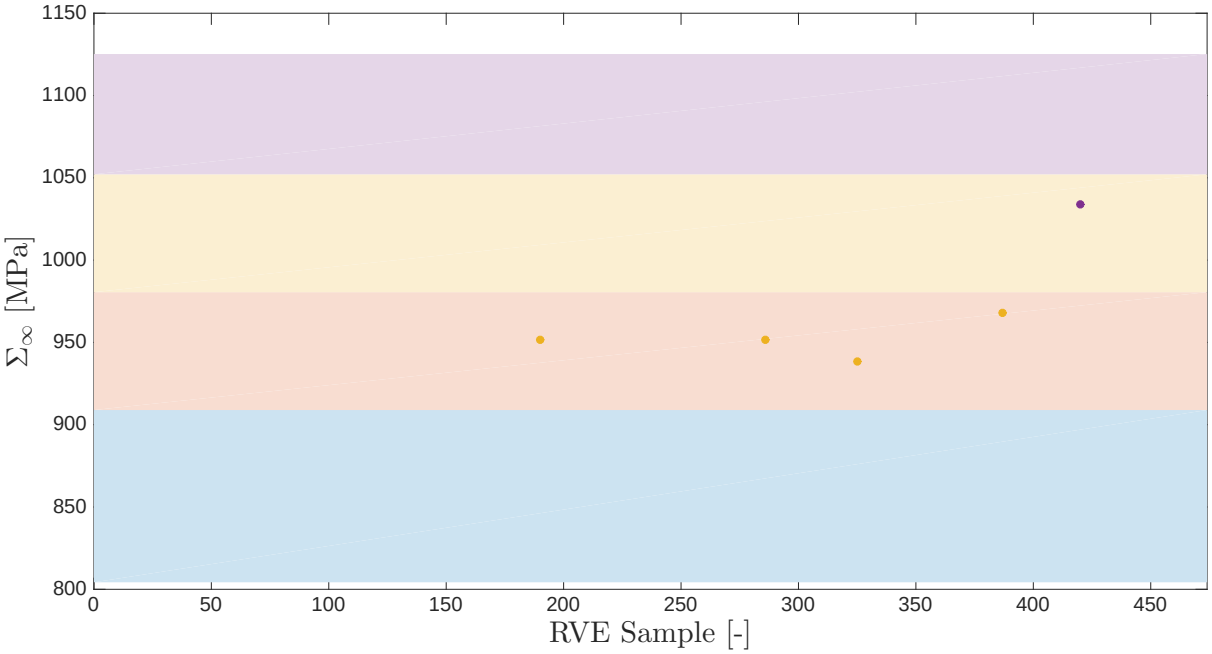


Figure 6: Errors of an ANN classification model on Σ_∞ with features in Category 2

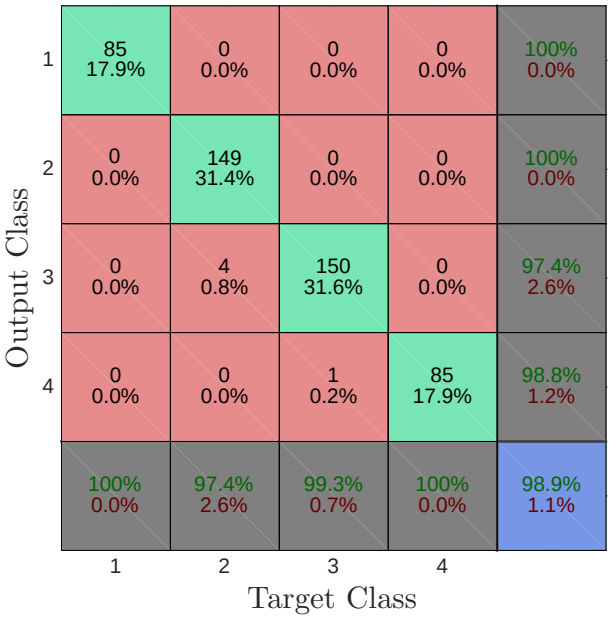


Figure 7: Confusion matrix of an ANN classier on Σ_∞ with features in Category 2

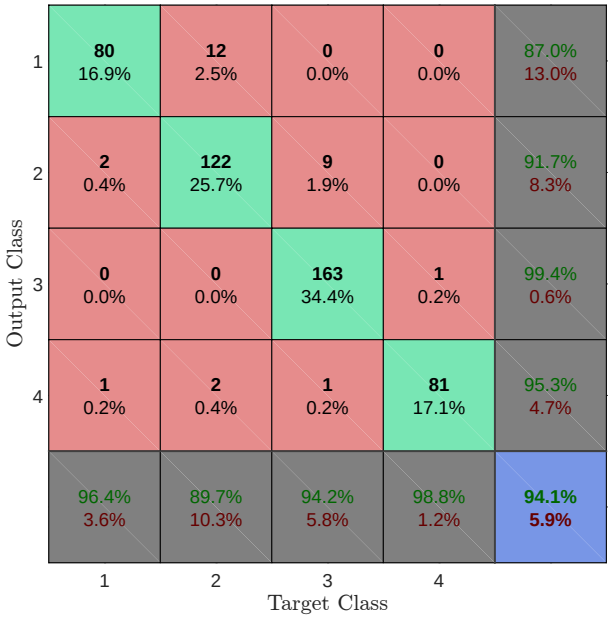


Figure 8: Confusion matrix of an ANN classier on Σ_∞ applied to test set

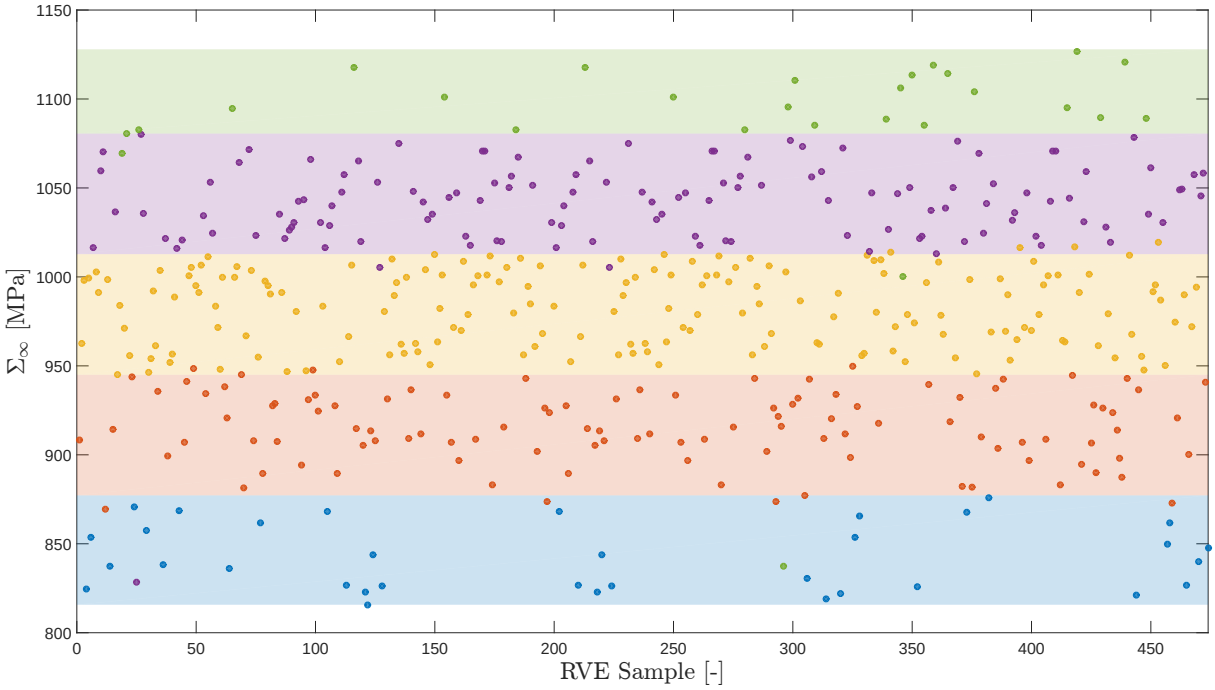


Figure 9: Results of an ANN classification model on Σ_∞ with features in Category 2 (5 classes)

REFERENCES

[1] D. Weichert, A. Hachemi, and F. Schwabe. Application of shakedown analysis to the plastic design of composites. *Archive of Applied Mechanics*, 69(9-10):623–633, 1999.

[2] H.F. Chen and A.R.S. Ponter. Shakedown and limit analyses for 3-D structures using the linear matching method. *International Journal of Pressure Vessels and Piping*, 78(6):443–451, 2001.

[3] G. Chen, U. A. Ozden, A. Bezold, C. Broeckmann, and D. Weichert. On the statistical determination of yield strength, ultimate strength, and endurance limit of a particle reinforced metal matrix composite (PRMMC). In D. Weichert P. Fuschi, A. A. Pisano, editor, *Direct Methods for Limit and Shakedown Analysis of Structures: Advanced Computational Algorithms and Material Modelling*, pages 105–122. Springer, 2015.

[4] S.M. Sapuan and I.M. Mujtaba. *Composite materials technology: neural network applications*. CRC Press, 2009.

[5] Z. Zhang and K. Friedrich. Artificial neural networks applied to polymer composites: a review. *Composites Science and Technology*, 63(14):2029 – 2044, 2003. Polymer Composites: Design, Materials, Manufacturing, Dedicated to Professor M. Neitzel.

[6] H.E. Kadi. Modeling the mechanical behavior of fiber-reinforced polymeric composite materials using artificial neural networksła review. *Composite Structures*, 73(1):1 – 23, 2006.

[7] A.O. Addin, S.M. Sapuan, E. Mahdi, and M. Othman. Prediction and detection of failures in laminated composite materials using neural networks–a review. *Polymers and Polymer Composites*, 14(4):433–442, 2006.

[8] M. Shabani and A. Mazahery. Application of finite element model and artificial neural network in characterization of Al matrix nanocomposites using various training algorithms. *Metallurgical and Materials Transactions A*, 43(6):2158–2165, 2012.

[9] M.A. Nik, K. Fayazbakhsh, D. Pasini, and L. Lessard. A comparative study of metamodeling methods for the design optimization of variable stiffness composites. *Composite Structures*, 107:494 – 501, 2014.

[10] H. Magoariec, S. Bourgeois, and O. Débordes. Elastic plastic shakedown of 3D periodic heterogeneous media: a direct numerical approach. *International Journal of Plasticity*, 20(8):1655–1675, 2004.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2001.

[12] MATLAB. *version 8.4.0 (R2014b)*. The MathWorks Inc., Natick, Massachusetts, 2014.

[13] ABAQUS. *ABAQUS/CAE user's manual : version 6.13*. Simulia, Dassault Systémes, 2013.

[14] Gurobi Optimization Inc. Gurobi optimizer reference manual(2014), 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SHAKEDOWN ANALYSIS OF PLATE BENDING UNDER STOCHASTIC UNCERTAINITY BY CHANCE CONSTRAINT PROGRAMMING

Ngọc Trình Trần[1], **Thanh Ngọc Trần**[2], **H.G. Matthies**[3], **G.E. Stavroulakis**[4] and **M. Staat**[1]

[1] Aachen University of Applied Sciences, Jülich Campus, Institute for Bioengineering, Biomechanics Lab., Heinrich-Mußmann-Str. 1, 52428 Jülich, Germany, trinhdhkt@gmail.com, m.staat@fh-aachen.de, http://www.fh-aachen.de/forschung/institut-fuer-bioengineering/
[2] University of Duisburg-Essen, Chair of Mechanics and Robotics, Lotharstr. 1, 47057 Duisburg, Germany, thanh.tran@uni-due.de
[3] TU Braunschweig, Institute of Scientific Computing, Hans-Sommer-Str. 65, 38092 Braunschweig, Germany, wire@tu-bs.de
[4] Technical University of Crete, 73100 Chania, Greece, gestavr@dpem.tuc.gr

Keywords: Limit analysis, Shakedown analysis, stochastic programming, chance constrained programming, non-linear programming.

Abstract. *In this paper we propose a stochastic programming to analyze limit and shakedown of plate bending under uncertainty conditions of strength. The Kirchhoff plate theory is used to formulate chance constrained problems. Based on the duality theory, the shakedown load multiplier formulated by the kinematic theorem is proved actually to be the dual form of the shakedown load multiplier formulated by static theorem. In this investigation a dual chance constrained programming algorithm is developed to calculate simultaneously both the upper and lower bounds of the plastic collapse limit and the shakedown limit.*

1 INTRODUCTION

An essential concern in mechanical and civil engineering design is to determine the ultimate load bearing capacity of structures beyond the elastic limit. Structures are considered safe if a state is reached after initial plastic deformation, such that the system does neither fail due to instantaneous or incremental collapse nor due to alternating plasticity. Using the classical step-by-step methods for such calculations can be cumbersome and computationally expensive in many cases. Furthermore, the disadvantage of these methods is that the exact knowledge of the loading history is necessary, which is not realistic in many technical applications. Limit and shakedown analysis are appropriate direct methods to avoid these problems. Reliability analysis of the structures takes the uncertainty of the actual load-carrying capacity of the structure into consideration since all resistance and loading variables are random in nature [1, 9, 10, 20].

Chance constrained programing is an approach of stochastic programming to limit and shakedown analysis under uncertainty [1, 2]. Under uncertainty the shakedown problem can be stated with random objective function or with random constraints. A probability is set with which the constraint has to be satisfied. Deterministic limit analysis of plates subjected to bending has been studied analytically and numerically. Both the Kirchhoff plate and the Mindlin plate models are considered based on the kinematic theorem of limit analysis and the finite element method. Based on the duality theory, Tran [3] developed a dual algorithm to calculate simultaneously both the upper and lower bounds of the plastic collapse limit and shakedown load of plates. In this study, we present a new primal-dual numerical algorithm of the shakedown problem under uncertainty in which the material strength is considered as a normally distributed random variable. Using chance constrained programming, we obtain deterministic equivalent formulations based on upper bound and lower bound theorems and then prove that formulations are actually dual to each other. The four-node discrete Kirchhoff quadrilateral (DKQ) element is used to analyze the problem. The proposed algorithm shows good performance in numerical test examples.

2 SHAKEDOWN ANALYSIS OF PLATE BENDING

2.1 Basic relations

Similar to the elastic theory, the inelastic behavior of thin plates may be analyzed under Kirchhoff's assumption that the normal to the middle plane of the plate remains straight and normal to the deformed middle plane. This assumption yields $u = -z\dfrac{\partial w}{\partial x}$, $v = -z\dfrac{\partial w}{\partial y}$, in which x, y are the coordinates in the middle plane of the plate, z is the distance from the middle plane, w is the deflection, u and v are the displacements in the x and y directions respectively. By differentiation, the strains are obtained as

$$\begin{Bmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \gamma_{xy} \end{Bmatrix} = \begin{Bmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial v}{\partial y} \\ \dfrac{\partial v}{\partial x} + \dfrac{\partial u}{\partial y} \end{Bmatrix} = \begin{Bmatrix} -z\dfrac{\partial^2 w}{\partial x^2} \\ -z\dfrac{\partial^2 w}{\partial y^2} \\ -2z\dfrac{\partial^2 w}{\partial x \partial y} \end{Bmatrix} = z \begin{Bmatrix} \chi_x \\ \chi_y \\ \chi_{xy} \end{Bmatrix} \tag{1}$$

The vector $\boldsymbol{\chi} = \left[\chi_x \ \chi_y \ \chi_{xy} \right]^T$ is called the vector of curvatures.

The kinematic relations can be written as follows:

$$\dot{\boldsymbol{\chi}} = -\nabla^2 \dot{w} \tag{2}$$

where $\dot{\boldsymbol{\chi}}$ is the curvature rate vector and $\dot{w}$ is the transversal velocity.

In this paper, the von Mises yield criterion is considered. Expressing in terms of moments, the criterion takes the form $f(m_x, m_y, m_{xy}) = \left(m_x^2 - m_x m_y + m_y^2 + 3m_{xy}^2 \right)^{1/2} - m_0 = 0$.

In matrix form the von Mises yield criterion can be written as follows:

$$f(\mathbf{m}) = \sqrt{\mathbf{m}^{\mathrm{T}} \mathbf{P} \mathbf{m}} - m_0 = 0 \tag{3}$$

where $\mathbf{m} = \left[m_x, m_y, m_{xy} \right]^{\mathrm{T}}$ is the vector of bending and twisting moments, $m_0 = \sigma_y t^2 / 4$ is the plastic limit moment per unit length of a plate section, t is the thickness of the plate, σ_y is the yield stress of the material, and

$$\mathbf{P} = \frac{1}{2} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 6 \end{bmatrix}. \tag{4}$$

2.2 Static formulation

Consider a convex polyhedral load domain $\mathcal{L}$ and a special loading path consisting of all load vertices $\hat{P}_k (k = 1, ..., m) \in \mathcal{L}$. Let a point on the problem domain A (the midplane) be identified by a vector variable $\mathbf{x}$ and the fictitious elastic moment vector be $\mathbf{m}^{\mathrm{E}}$. The static shakedown theorem states that shakedown occurs if, and only if, there exists a time-independent self-equilibrium residual moment field $\bar{\boldsymbol{\rho}}$ which is statically admisible so that the actual moment field $\mathbf{m} = \alpha \mathbf{m}^{\mathrm{E}} + \bar{\boldsymbol{\rho}}$ does not anywhere violate the yield condition at any point of the structure and for all posible load combinations. Based on this theorem, we can find a statically admisible residual moment field in order to obtain a maximum load domain $\alpha \mathcal{L}$ that guarantees (3). Therefore, the shakedown load factor α^- can be obtained by solving the following optimization problem:

$$\begin{aligned} \alpha^- &= \max \alpha \\ s.t.: &\begin{cases} \nabla^2 \bar{\boldsymbol{\rho}}(\mathbf{x}) = 0 & \text{in } A \\ f(\alpha \mathbf{m}^{\mathrm{E}} + \bar{\boldsymbol{\rho}}) \le m_0 \end{cases} \end{aligned} \tag{5}$$

By discretizing the entire problem domain into finite elements and applying the Gauss-Legendre integration technique, Eqs (5) can be rewritten in the following form:

$$\alpha^{-} = \max \alpha$$
$$\text{s.t.:} \begin{cases} \sum_{i=1}^{NG} w_i \mathbf{B}_i^T \bar{\boldsymbol{\rho}}_i = 0 \qquad \text{in } A \\ f(\alpha \mathbf{m}_{ik}^E + \bar{\boldsymbol{\rho}}_i) \le m_0 \qquad \forall i = \overline{1, NG}, \quad \forall k = \overline{1, m} \end{cases} \tag{6}$$

in which $\mathbf{B}_i$ is the deformation matrix, w_i is integration weight at Gauss point i; NG denotes the total number of Gauss points of the structure, and m is the number of load vetices..

2.3 Kinematic formulation

An upper bound to the shakedown limit of plate can be obtained using the kinematic shakedown theorem which has the following two statements:

(a) *Shakedown will occur for a structure subject to repeated or cyclic loads, if the plastic dissipation power exceeds the work rate of external forces for any admissible plastic strain-rate cycles and all loading paths*
(b) *Shakedown cannot occur, if the plastic dissipation power is less than the work rate of external forces for any one admissible plastic strain-rate cycle or any one loading path.*

We introduce here an admissible cycle of plastic curvature field $\Delta\boldsymbol{\chi}^p$. At each load vertex, the plastic curvature rate may not necessarily be compatible at each instant during the time cycle, but the plastic curvature accumulation over the cycle is required to be kinematically compatible such that

$$\Delta\boldsymbol{\chi}^p = \sum_{k=1}^{m} \dot{\boldsymbol{\chi}}_k^p = \nabla^2 \dot{w}. \tag{7}$$

Based on the above statements and the mathematical programming theory, an upper bound of the shakedown load factor can be found by solving the following convex nonlinear programming (the superscript p is neglected for simplicity):

$$\alpha^{+} = \min \sum_{k=1}^{m} \int_A D_{\text{int}}(\dot{\boldsymbol{\chi}}_k)\, dA$$
$$\text{s.t.:} \begin{cases} \Delta\boldsymbol{\chi} = \sum_{k=1}^{m} \dot{\boldsymbol{\chi}}_k = \nabla^2 \dot{w} \qquad \text{in } A \\ \dot{w} = 0 \qquad on \quad \partial A \\ \sum_{k=1}^{m} \int_A \mathbf{m}^E(x, \hat{P}_k)^T \dot{\boldsymbol{\chi}}_k dA = 1 \end{cases} \tag{8}$$

where $D_{\text{int}}(\boldsymbol{\chi})$ is the plastic dissipation power per unit area. According to the von Mises yield condition, the plastic dissipation can be expressed in term of the curvature rates of the middle surface: $D_{\text{int}}(\dot{\boldsymbol{\chi}}) = m_0 \sqrt{\dot{\boldsymbol{\chi}}_k{}^T \mathbf{Q} \dot{\boldsymbol{\chi}}_k}$
with

$$\mathbf{Q} = \mathbf{P}^{-1} = \frac{1}{3}\begin{bmatrix} 4 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

We denote the nodal variables of the finite element by $\mathbf{q} = \left[w \;\; \partial w/\partial x \;\; \partial w/\partial y \right]^{\mathrm{T}}$, the discretized formulation by FEM is as follows:

$$\begin{aligned} \alpha^{+} &= \min \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i m_0 \sqrt{\dot{\boldsymbol{\chi}}_{ik}{}^{T} \mathbf{Q} \dot{\boldsymbol{\chi}}_{ik}} \\ \text{s.t.:} &\begin{cases} \sum_{k=1}^{m} \dot{\boldsymbol{\chi}}_{ik} = \mathbf{B}_i \dot{\mathbf{q}} & \forall i = \overline{1, NG} \\ \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i \dot{\boldsymbol{\chi}}_{ik}^{T} \mathbf{m}_{ik}^{E} = 1 \end{cases} \end{aligned} \tag{10}$$

3 STOCHASTIC SHAKEDOWN ANALYSIS OF PLATE BENDING BY CHANCE CONSTRAINED PROGRAMMING

3.1 Static Approach to chance constrained programming

Consider the situation that the strength of material is not given, fixed quantity, but must be modelled $R = R(\omega)$ through random variables on a certain probability space. Hence, the shakedown problem (6),(10), respectively, under stochastic uncertainty must be reformulated by appropriate deterministic problems which are provided by adopting stochastic optimization techniques. A technique used effectively called chance constrained programming is presented here.

For stochastic plastic moment $\mathbf{m}_i(\omega)$ depending on some random ω, a stochastic formulation can be obtained by assuming that the inequality constraints of (6) are satisfied at least by a chance $\psi = \left[0.5 \,;\, 1\right]$ (or reliability level) at Gauss point i.

$$\begin{aligned} \alpha^{-} &= \max \alpha \\ \text{s.t.:} &\begin{cases} \sum_{i=1}^{NG} w_i \mathbf{B}_i^{T} \bar{\boldsymbol{\rho}}_i = 0 \quad \text{in } A \\ \mathrm{Prob}\left[f(\alpha \mathbf{m}_{ik}^{E} + \bar{\boldsymbol{\rho}}_i) - m_{0i}(\omega) \le 0 \right] \ge \psi \quad \forall i = \overline{1, NG}, \quad \forall k = \overline{1, m} \end{cases} \end{aligned} \tag{11}$$

Based on the methodology of chance constrained programming, the stochastic program (11) is to be relaxed into an equivalent deterministic optimization problem:

$$\begin{aligned} \alpha^{-} &= \max \alpha \\ \text{s.t.:} &\begin{cases} \sum_{i=1}^{NG} w_i \mathbf{B}_i^{T} \bar{\boldsymbol{\rho}}_i = 0 \quad \text{in } A \\ f(\alpha \mathbf{m}_{ik}^{E} + \bar{\boldsymbol{\rho}}_i) \le \mu_i - \kappa_i \sigma_i \quad \forall i = \overline{1, NG}, \quad \forall k = \overline{1, m} \end{cases} \end{aligned} \tag{12}$$

where $\mu_i, \sigma_i, \kappa_i = \Phi^{-1}(\psi_i)$ is the mean value, standard deviation and the inverse of the cumulative distribution function of plastic moment at Gauss point i, respectively.

3.2 Kinematic approach to chance constrained programming

As mentioned above the deterministic formulation to calculate an upper bound shakedown load factor:

$$\alpha^+ = \min \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i m_0 \sqrt{\dot{\boldsymbol{\chi}}_{ik}{}^T \mathbf{Q} \dot{\boldsymbol{\chi}}_{ik}}$$

$$\text{s.t.:} \begin{cases} \sum_{k=1}^{m} \dot{\boldsymbol{\chi}}_{ik} = \mathbf{B}_i \dot{\mathbf{q}} \qquad \forall i = \overline{1, NG} \\ \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i \dot{\boldsymbol{\chi}}_{ik}^T \mathbf{m}_{ik}^E = 1 \end{cases} \tag{13}$$

If the yield stress of material is random, then the plastic moment is uncertainty quantity and the objective function of (13) is a stochastic variable. Firstly, we must properly define the minimum of a random function. This can be done in such a way that one looks for a minimum lower bound η objective function under the constraint that the probability of violation of that bound is prescribed [4].

$$\min \eta$$

$$\text{s.t.:} \begin{cases} \text{Prob}\left(\sum_{k=1}^{m} \sum_{i=1}^{NG} w_i m_0(\omega) \sqrt{\dot{\boldsymbol{\chi}}_{ik}^{\mathrm{T}} \mathbf{Q} \dot{\boldsymbol{\chi}}_{ik}} \geq \eta \right) = \psi \\ \sum_{k=1}^{m} \dot{\boldsymbol{\chi}}_{ik} = \mathbf{B}_i \dot{\mathbf{q}} \qquad \forall i = \overline{1, NG} \\ \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i \dot{\boldsymbol{\chi}}_{ik}^{\mathrm{T}} \mathbf{m}_{ik}^E = 1 \end{cases} \tag{14}$$

The chance constrained program technique also can convert the stochastic program (14) into an equivalent deterministic program as follows:

$$\alpha^+ = \min \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i \left(\mu_i - \kappa_i \sigma_i \right) \sqrt{\dot{\boldsymbol{\chi}}_{ik}^{\mathrm{T}} \mathbf{Q} \dot{\boldsymbol{\chi}}_{ik}}$$

$$\text{s.t.:} \begin{cases} \sum_{k=1}^{m} \dot{\boldsymbol{\chi}}_{ik} = \mathbf{B}_i \dot{\mathbf{q}} \qquad \forall i = \overline{1, NG} \\ \sum_{k=1}^{m} \sum_{i=1}^{NG} w_i \dot{\boldsymbol{\chi}}_{ik}^{\mathrm{T}} \mathbf{m}_{ik}^E = 1 \end{cases} \tag{15}$$

4 DUALITY APPROACH TO CHANCE CONSTRAINED PROGRAMMING

4.1 Dual theorem for chance constrained programming

Some new notations is intruduced for the sake of simplicity:

$$\dot{\mathbf{k}}_{ik} = w_i \mathbf{Q}^{1/2} \dot{\boldsymbol{\chi}}_{ik}, \qquad \mathbf{t}_{ik} = \left(\mathbf{Q}^{-1/2}\right)^T \mathbf{m}_{ik}^E, \qquad \hat{\mathbf{B}}_i = w_i \mathbf{Q}^{1/2} \mathbf{B}_i \tag{16}$$

where

$$\mathbf{Q}^{1/2}\mathbf{Q}^{-1/2} = \mathbf{I}, \;\; \mathbf{Q} = \left(\mathbf{Q}^{1/2}\right)^T \mathbf{Q}^{1/2} \tag{17}$$

By substituting (16) into (15) one obtains a simplified version for upper bound of shakedown limit load (primal problem)

$$\alpha^+ = \min \sum_{k=1}^{m} \sum_{i=1}^{NG} (\mu_i - \kappa_i \sigma_i) \sqrt{\dot{\mathbf{k}}_{ik}^T \dot{\mathbf{k}}_{ik}}$$

$$\text{s.t.:} \quad \begin{cases} \sum_{k=1}^{m} \dot{\mathbf{k}}_{ik} - \hat{\mathbf{B}}_i \dot{\mathbf{q}} = \mathbf{0} \qquad \forall i = \overline{1, NG} & \text{(a)} \\ \sum_{i=1}^{NG} \sum_{k=1}^{m} \dot{\mathbf{k}}_{ik}^T \mathbf{t}_{ik} - 1 = 0 & \text{(b)} \end{cases} \tag{18}$$

It is seen that in the case of limit analysis there exists a dual form for (18), see, e.g. Heitzer and Staat [16] , Andersen *et al.* [17]. Vu *et al*. [18], [19] generalized this theory for the case of shakedown analysis. An extension of their theory to shakedown analysis of plate is investigated in this work through the following theorem:

> **Theorem**. *If there exists a finite solution α^+ for the upper bound shakedown limit load multiplier (20) then the static formulation (12) is exactly the dual problem of the kinematic one (20) such that*

$$\alpha^+ = \min_{\mathbf{h}\left(\dot{\mathbf{k}}_{ik}, \dot{\mathbf{q}}\right) = \mathbf{0}} \sum_{k=1}^{m} \sum_{i=1}^{NG} (m_i - \kappa_i \sigma_i) \sqrt{\dot{\mathbf{k}}_{ik}^T \dot{\mathbf{k}}_{ik}} = \max_{\substack{\sum_{i=1}^{NG} w_i \mathbf{B}_i^T \bar{\boldsymbol{\rho}}_i = \mathbf{0} \\ f\left(\alpha \mathbf{m}_{ik}^E + \bar{\boldsymbol{\rho}}_i\right) \le 0}} \alpha = \alpha^- \tag{19}$$

4.2 Primal-dual shakedown algorithm

A difficult happening when dealing with the non-linear constrained optimization problem (20) is that the objective function is not everywhere differentiable. In order to allow a direct non-linear, smooth optimization problem, a 'smooth regularization method' should be used for overcoming this barrier. For this purpose, a very small positive number, ε_0^2 will be added to $D_{\text{int}}(\dot{\boldsymbol{\chi}}_k)$. An efficient technique for large-scale optimization problems, which are successfully applied in [13] is used. Using penalty method to eliminate the first constraint in (18) lead to a penalty function

$$P = \sum_{i=1}^{NG} \left\{ \sum_{k=1}^{M} \left((m_i - \kappa_i \sigma_i) \sqrt{\dot{\mathbf{k}}_{ik}^T \dot{\mathbf{k}}_{ik} + \varepsilon_0^2} \right) + \frac{C}{2} \left(\sum_{k=1}^{m} \dot{\mathbf{k}}_{ik} - \hat{\mathbf{B}}_i \dot{\mathbf{q}} \right)^T \left(\sum_{k=1}^{m} \dot{\mathbf{k}}_{ik} - \hat{\mathbf{B}}_i \dot{\mathbf{q}} \right) \right\} \tag{20}$$

where c is a penalty parameter such that $c \gg 1$. The corresponding Lagrangian of (20) is

$$L = P - \alpha \left(\sum_{i=1}^{NG} \sum_{k=1}^{m} \dot{\mathbf{k}}_{ik}^T \mathbf{t}_{ik} - 1 \right) \tag{21}$$

We denote

$$\boldsymbol{\beta}_i = -c\left(\sum_{k=1}^{m}\dot{\mathbf{k}}_{ik} - \hat{\mathbf{B}}_i\dot{\mathbf{q}}\right) \tag{22}$$

By employing Newton method to solve the KKT conditions of the Lagrangian in (39) and after some manipulations, one gets the following system:

$$\mathbf{K}d\dot{\mathbf{q}} = -\mathbf{K}\dot{\mathbf{q}} + \mathbf{f}_1 + \mathbf{f}_2(\alpha + \mathrm{d}\alpha) \tag{23}$$

in which

$$\begin{aligned}
\mathbf{K} &= \sum_{i=1}^{NG}\hat{\mathbf{B}}_i^T\mathbf{E}_i^{-1}\hat{\mathbf{B}}_i \\
\mathbf{f}_1 &= -\sum_{i=1}^{NG}\hat{\mathbf{B}}_i^T\mathbf{E}_i^{-1}\sum_{k=1}^{m}\mathbf{M}_{ik}^{-1}(\boldsymbol{\beta}_i + \alpha\mathbf{t}_{ik})\frac{\dot{\mathbf{k}}_{ik}^T\dot{\mathbf{k}}_{ik}}{\sqrt{\dot{\mathbf{k}}_{ik}^T\dot{\mathbf{k}}_{ik} + \varepsilon_0^2}} \\
\mathbf{f}_2 &= \sum_{i=1}^{NG}\hat{\mathbf{B}}_i^T\mathbf{E}_i^{-1}\sum_{k}^{m}\mathbf{M}_{ik}^{-1}\sqrt{\dot{\mathbf{k}}_{ik}^T\dot{\mathbf{k}}_{ik} + \varepsilon_0^2}\,\mathbf{t}_{ik}
\end{aligned} \tag{24}$$

Solving the system (23), we have the incremental vectors of nodal variables $\dot{\mathbf{q}}$, curvature rate $\dot{\mathbf{k}}_{ik}$ and $\boldsymbol{\beta}_i$ as follows :

$$\begin{aligned}
d\dot{\mathbf{q}} &= d\dot{\mathbf{q}}_1 + d\dot{\mathbf{q}}_2\left(\alpha + d\alpha\right) \\
d\dot{\mathbf{k}}_{ik} &= (d\dot{\mathbf{k}}_{ik})_1 + (d\dot{\mathbf{k}}_{ik})_2\left(\alpha + d\alpha\right) \\
d\boldsymbol{\beta}_i &= (d\boldsymbol{\beta}_i)_1 + (d\boldsymbol{\beta}_i)_2\left(\alpha + d\alpha\right)
\end{aligned} \tag{25}$$

and

$$\left(\alpha + d\alpha\right) = \frac{1 - \sum_{i=1}^{NG}\sum_{k=1}^{m}\mathbf{t}_{ik}^T\left[\dot{\mathbf{k}}_{ik} + \left(d\dot{\mathbf{k}}_{ik}\right)_1\right]}{\sum_{i=1}^{NG}\sum_{k=1}^{m}\mathbf{t}_{ik}^T\left(d\dot{\mathbf{k}}_{ik}\right)_2} \tag{26}$$

The vectors $\mathrm{d}\dot{\mathbf{q}}, \mathrm{d}\dot{\mathbf{k}}_{ik}, \mathrm{d}\boldsymbol{\beta}_i$ and $\mathrm{d}\alpha$ are actually Newton directions which assure that a suitable step along them will lead to a decrease of the objective function of the primal problem (18) and to an increase of the objective function of the objective function of the dual problem (12). Based on(25), (26) we can update the vectors of $\dot{\mathbf{q}}, \dot{\mathbf{k}}_{ik}$, $\boldsymbol{\beta}_i$ and α.

5 NUMERICAL EXAMPLES

In this section, the numerical solution of some problems is presented to test the performance of the dual shakedown algorithm. Plates subjected to uniform or concentrated loads are considered. The 4-node DKQ plate element is applied for structural discretization. For all examples the following was assumed: length $L = 10m$, plate thickness $t = 0.1\mathrm{m}$, the

mean value of yield stress $\mathrm{E}(\sigma_y) = 250\,\mathrm{MPa}$ and the standard deviation $\sigma = 0.1\mathrm{E}(\sigma_y)$. A reliability level $\psi = 0.9999$ is chosen.

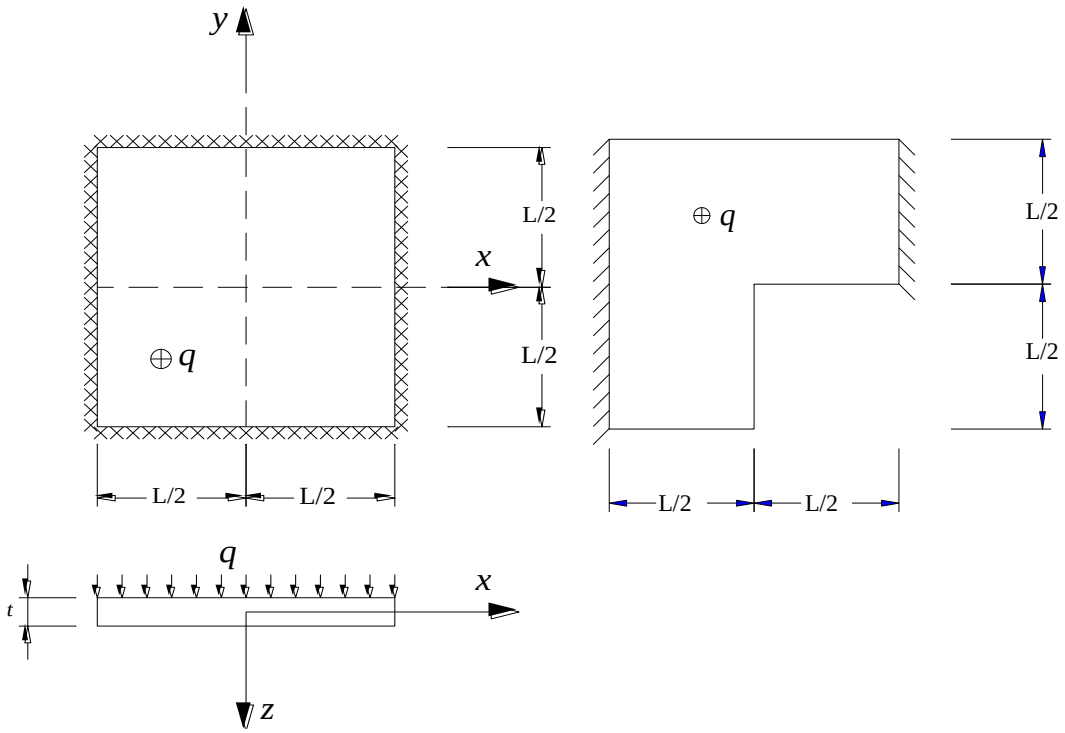


Figure 1. Square plate and L-shape plate loaded by a uniform pressure

5.1 Limit analysis of square plate subjected to uniform load

Firstly, we consider a square plate subjected to uniform pressure q as shown on Figure 1. In this analysis, the plate is modelled by 256 DKQ elements due to symmetry. Table 1 shows the comparison of the present numerical results for both cases, simply supported and clamped plates.

Authors	Upper/Lower bound (deterministic)		Upper/Lower bound (random strength)	
	Spl. supported	Clamped	Spl. supported	Clamped
Hodge *et al.* [14]	26.54/24.86	49.25/42.86	–/–	–/–
Lubliner [17]	27.71/23.81	52.01/–	–/–	–/–
Capsoni *et al.* [15]	25.02/–	45.29/–	–/–	–/–
Le *et al.* [16]	25.01/–	45.29/–	–/–	–/–
Present	25.04/25.04	45.06/45.06	15.72/15.72	28.36/28.36

Table 1. Limit load factor of square plates in comparison with other solutions $\left(\frac{m_0}{qL^2}\right)$.

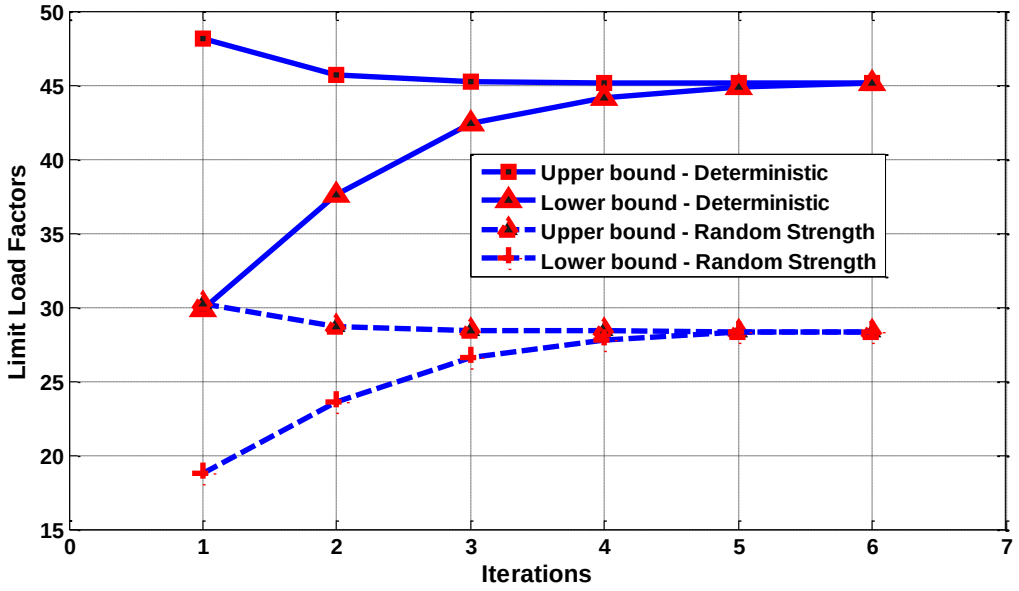


Figure 2. Clamped square plate: Convergence of limit load factors.

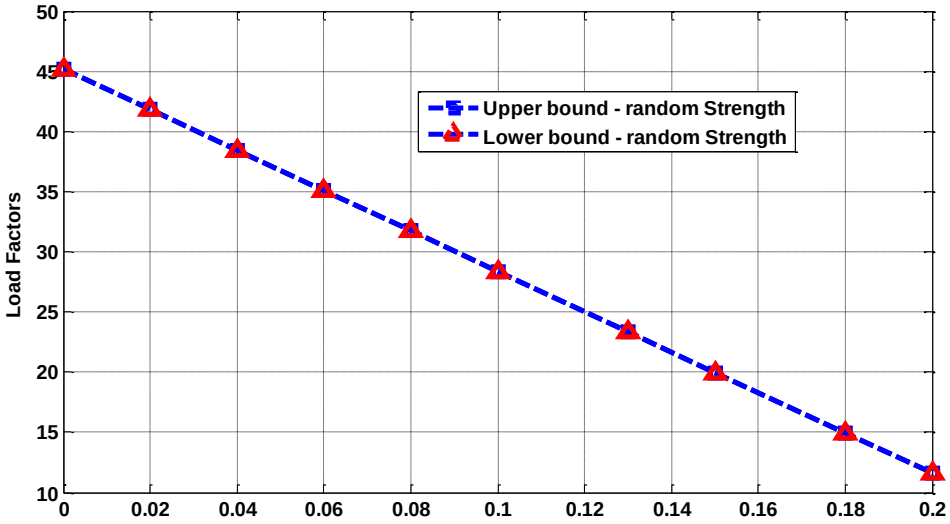


Figure 3. Dependence of load factors on the coefficient of variation $\xi\left(\xi=\frac{\sigma}{E}\right)$.

5.2 L-shaped plate subjected to uniform load.

In the second example, we investigate an L-shape plate subjected to uniform pressure q (Figure 1) which can be constant or vary within a range $q \in [0 \;\; q_{\max}]$. In this analysis, the plate is modelled by 768 DKQ elements. Figure 4 and figure 5 show the convergence of the upper bound and lower bounds for the simply supported case.

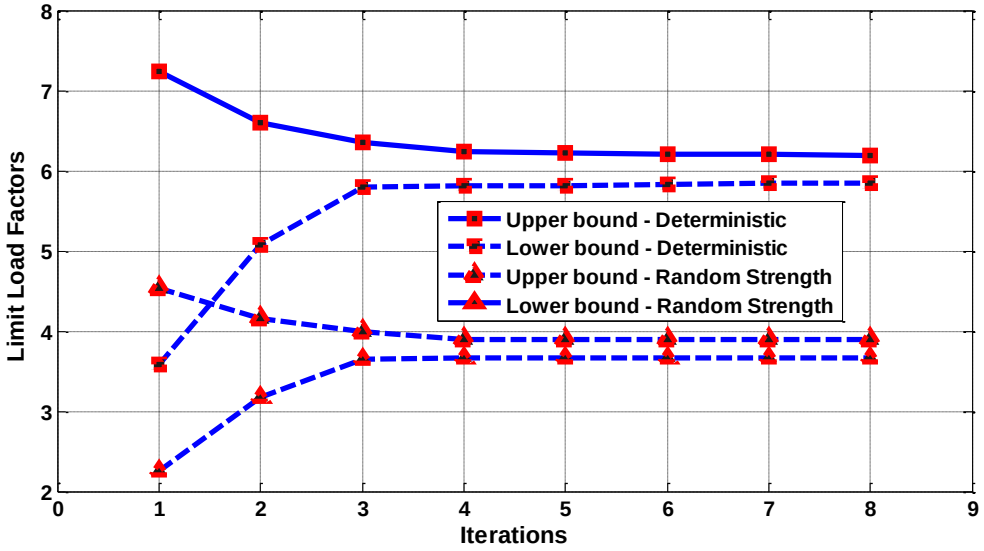


Figure 4: L-shape Plate: Convergence of limit load factors.

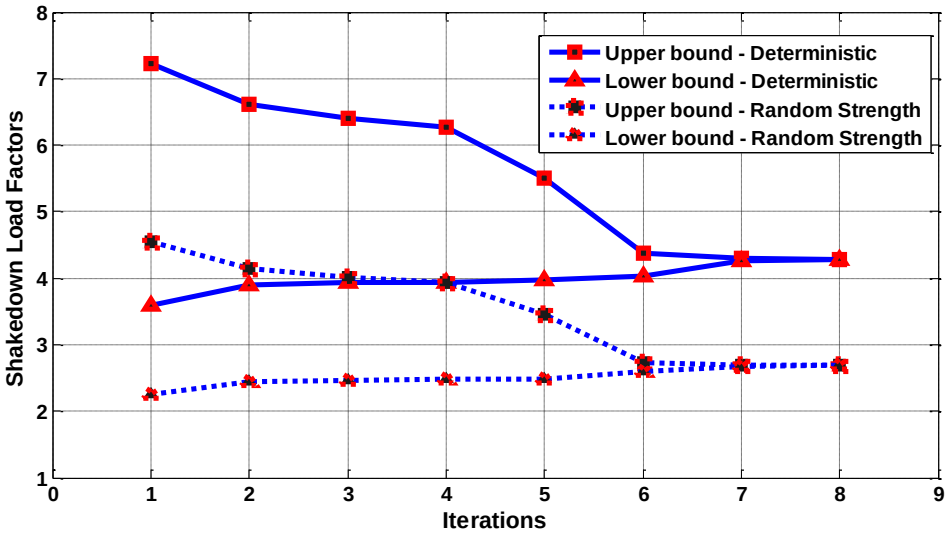


Figure 5: L-shape Plate: Convergence of shakedown load factors

Authors	Upper/Lower bound (deterministic)		Upper/Lower bound (random strength)	
	Limit	Shakedown	Limit	Shakedown
Le *et al.* [16]	6.289/–	–/–	–/–	–/–
Present	6.19/5.85	4.28/4.28	3.89/3.67	2.69/2.69

Table 2. Limit load factor of plate in comparison with other solutions $\left(\frac{m_0}{qL^2}\right)$

6 CONCLUSIONS

- Direct structural reliability design can be achieved by chance constraint programming.
- In the general case chance constraint programming is a hard problem because probabilities have to be calculated as high dimensional integrals during in the optimization algorithm.
- For normally distributed stochastic variables deterministic equivalents can be formulated (for linear programming).
- Extension to nonlinear programming is on its way.

For engineering design:

- Structural reliability is post design
- Stochastic programming is (pre) design
- The load factor decreases "quickly" with increasing coefficient of variation of the strength.
- High structural reliability can be achieved with moderate reduction of the load factor

REFERENCES

[1] M. Staat, M Heitzer, Probabilistic limit and shakedown problems. In: M. Staat, M. Heitzer (eds.) *Numerical methods for limit and shakedown analysis. Deterministic and probabilistic approach*. Part VII. NIC Series Vol. 15, Jülich (2003) 217-268. http://hdl.handle.net/2128/2926

[2] N.T. Trần, T.N. Trần, H.G. Matthies, G.E. Stavroulakis, M. Staat, FEM shakedown analysis of uncertain structures by chance constraint programming. Joint DMV and GAMM Annual Meeting, March 7-11, 2016, Braunschweig, Germany.

[3] T.N. Tran, A dual algorithm for shakedown analysis of plate bending. *International Journal for Numerical Methods in Engineering* **86,** 862-875, 2011.

[4] Shinji Kataoka, A Stochastic programming model, *Econometrica*, **31**, 181-196, 1963.

[5] A. Charnes, W.W. Cooper, Chance-constrained programming. *Management Science*, **6**(1), 73-79, 1959. http://dx.doi.org/10.1287/mnsc.6.1.73

[6] S.S. Rao, *Engineering optimization. theory and practice*. 4th ed. John Wiley, Hoboken, New Jersey (2009). http://dx.doi.org/10.1002/9780470549124

[7] K.A. Sikorski, A. Borkowski, Ultimate load analysis by stochastic programming. In: D.L. Smith (ed.) *Mathematical programming methods in structural plasticity*. CISM Courses and Lectures No. 299, Springer, Wien (1990) 403-424. http://dx.doi.org/10.1007/978-3-7091-2618-9_20

[8] F. Tin-Loi, L. Qi, Z. Wei, R.S. Womersley, Stochastic ultimate load analysis: models and solution methods. *Numer. Func. Anal. Opt.*, **17**(9-10), 1029-1043, 1996. http://dx.doi.org/10.1080/01630569608816740.

[9] T.N. Tran, R. Kreißig, M. Staat, Probabilistic limit and shakedown analysis of thin plates and shells. *Structural Safety*, **31**(1), 1-18, 2009. http://dx.doi.org/10.1016/j.strusafe.2007.10.003

[10] T.N. Tran, M. Staat, Uncertain multimode failure and shakedown analysis of shells. In: P. Fuschi, A.A. Pisano, D. Weichert (eds.) *Direct methods for limit and shakedown analysis of structures*. Chap. 14. Springer, (2015) 279-298. http://dx.doi.org/10.1007/978-3-319-12928-0_14

[11] D.K. Vu, A.M. Yan, D.H. Nguyen, A primal-dual algorithm for shakedown analysis of structure. *Computer methods in Applied Mechanics and Engineering* **193**, 4663-4674, 2004

[12] P.G. Hodge, T. Belytschko, Numerical methods for the limit analysis of plates. *Transactions of ASME, Journal of Applied Mechanics* **35**, 796-801, 1986.

[13] A. Capsoni, L. Corradi, Limit analysis of plates – a finite element formulation. *Structural Engineering and Mechanics* **8**, 325-341, 1999.

[14] V.C. Le, M. Gilbert, H. Asker, Limit analysis of plates using the EFG method and second-order cone programming. *International Journal for Numerical Methods in Engineering* **78**, 1532-1552, 2009.

[15] J. Lubliner, *Plasticity theory*. Macmillan, 1990.

[16] M. Heitzer, M. Staat: FEM-computation of load carrying capacity of highly loaded passive components by direct methods. *Nuclear Engineering and Design* **193**, 349-358, 1999. http://dx.doi.org/10.1016/S0029-5493(99)00190-9

[17] K.D. Andersen, E. Christiansen, A.R. Conn, M.L. Overton, An efficient primal-dual interior-point method for minimizing a sum of Euclidean norms. *SIAM Journal on Scientific Computing* **22**, 243-262, 2000.

[18] D.K. Vu, A.M. Yan, D.H. Nguyen, A dual form for discretized kinematic formulation in shakedown analysis. *International Journal of Solid and Structures* **41**, 267-277, 2004.

[19] D.K. Vu, M, Staat, Shakedown analysis of structures made of materials with temperature-dependent yield stress. *International Journal of Solids and Structures*, **44**(13), 4524-4540, 2007. http://dx.doi.org/10.1016/j.ijsolstr.2006.11.038

[20] M. Staat, Limit and shakedown analysis under uncertainty. *Int. J. Comput. Methods*, **11**(2), 1343008, 2014. http://dx.doi.org/10.1142/S0219876213430081

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

LIMIT ANALYSIS LOCUS OF HIGH STRENGTH STEEL PLATES UNDER NON-QUADRATIC YIELD CRITERIA

Konstantinos D. Nikolaou[1] **and Christos D. Bisbos**[2]

[1]NIKI Ltd. Digital Engineering,
Ethnikis Antistasis 205, Katsika, Ioannina, 45500 Greece
e-mail: konstantinos.nikolaou@nikitec.gr

[2]Dept. of Civil Engineering, Aristotle University of Thessaloniki,
University Campus, Thessaloniki, 54124 Greece
e-mail: cbisbos@civil.auth.gr

Keywords: anisotropy, limit analysis, high strength steel, nonlinear mathematical programming.

Abstract. *High strength steel sheets, widely used in industrial applications, exhibit a significant anisotropy of plastic response between their rolling and transverse direction. Hills yield criterion is the most popular in sheet metal forming FE simulations, because its parameters are easily determined. In the literature, several other criteria are proposed that can better describe the anisotropic behavior of specific materials. In limit analysis the material elastic properties are irrelevant. A formulation, based on the static theorem, requires only the equations of equilibrium and a respective yield criterion to describe the safety margins of a structure. Limit analysis implements standard FEM data into mathematical optimization packages to yield the respective safety factor. Depending on the yield criterion type, a large scale nonlinear mathematical programming problem must be solved. Quadratic yield criteria, like Hill, lead to Second Order Cone Programming problems. Non-quadratic yield criteria, on the other hand, lead to problems that can be treated only by general non-linear mathematical programming algorithms. In the present study, a polynomial criterion proposed by Yoshida et al. is used to construct the limit load locus of high strength steel structures. Results are compared to the von Mises and Hill criteria.*

1 INTRODUCTION

High strength steel sheets due to their crystallographic structure, chemical properties and forming process exhibit a significant anisotropy in their mechanical properties between rolling and transverse direction. Defining the safety margins of structures using such materials is crucial for their design.

Limit analysis allows the direct determination of the load bearing capacity of structures subjected to monotonically increasing loads, requiring only limited input data [1-5]. Standard FEM data is appropriately combined with optimization techniques to yield the respective safety factor. For common engineering structures usually a large scale mathematical programming problem has to be solved. Depending on the selected yield criterion a linear quadratic or general nonlinear optimization problem arises.

Within the framework of the direct methods of plasticity, anisotropic materials have been studied using the kinematic theorem [1] (upper bound approach). For example, general anisotropic structure limit analysis is presented in [6, 7], while in [8] limit analysis of orthotropic composite laminates is studied using the linear matching method. In [9] yield stresses for the different directions are incorporated in a single ellipsoidal yield surface used in 2D and 3D problems.

In the present work, limit analysis of anisotropic structures based on the static theorem is considered. More precisely, we use as basis the mathematical dual of an upper bound problem. To this goal, several yield criteria from the literature are compared. Together with the popular quadratic criteria von Mises [10] and Hill [11], a highly flexible "user-friendly" nonlinear criterion proposed by Yoshida [12] is used.

The plan of this paper is as follows. First limit analysis is presented as a nonlinear programming problem. Then, the yield criteria used are discussed. Finally, the paper closes with a numerical example and some concluding remarks.

2 LIMIT ANALYSIS AS A NONLINEAR PROGRAMMING PROBLEM

Let us consider a structure Ω made of an anisotropic elastic-plastic material. Let Ω be discretized using NE finite elements with NU free degrees of freedom and NG numerical integration points (Gauss points) for the whole structure. Limit analysis using the static theorem leads to the following non linear programming problem:

$$\begin{aligned} P_{LA} \qquad & \max a \\ s.t.: \qquad & \mathbf{H}_j\, \mathbf{s}_j = a\, \phi^{(ext)} \\ & \mathbf{s}_j \in \mathcal{F}_j, \quad for\ j = 1, \ldots, NG \end{aligned} \tag{1}$$

Unknowns are the total elastoplastic stresses $\mathbf{s}_j$ and the load multiplier (safety factor) a. $\phi^{(ext)}$ is a vector of size NU that contains all the nodal loads applied to the structure. $\mathbf{H}_j$ is the equilibrium matrix depending on the discretization and the boundary conditions. $\mathcal{F}_j$ is the local yield criterion that has to be satisfied at every stress checking point. Depending on the type of $\mathcal{F}_j$ a different optimization problem has to be solved. Quadratic yield criteria lead to second order cone programming problems, for which several very efficient optimization packages exits (i.e. MOSEK [13]). Non-linear, non-quadratic yield criteria lead to problems that can be treated only by general non-linear mathematical programming algorithms as IPOPT [14].

The safety factor depends only on the yield criterion, discretization and boundary conditions. Elastic properties (i.e. Young's modulus or elastic stresses) do not affect the solution of the limit analysis problem P_{LA}. Obviously, anisotropy is taken into account only through appropriate yield criteria.

3 ANISOTROPIC YIELD CRITERIA

Yield criteria are central to the limit analysis problem. In this section the different criteria used will be shortly described for the plane stress case. The von Mises criterion [10] reads:

$$\sigma_{11}^2 + \sigma_{22}^2 - \sigma_{11}\,\sigma_{22} + 3\,\sigma_{12}^2 \leq \sigma_y^2 \tag{2}$$

σ_y is the uniaxial tensile yield stress. Hill proposed a number of criteria for the anisotropic behavior of metals. Hill-48 [11] criterion is the generalization of von Mises and one of the most popular in elastoplastic analyses of anisotropic structures. For plane stress case:

$$A_1\,{\sigma_{11}}^2 - A_2\,\sigma_{11}\,\sigma_{22} + A_3\,{\sigma_{22}}^2 + 3\,A_4\,{\sigma_{12}}^2 = 1 \tag{3}$$

where $A_{1,\ldots,4}$ are the anisotropic parameters. They can be determined using experimental results like $r-$values r_0, r_{45} and r_{90} for the three tension axis directions from rolling direction of a sheet, as follows:

$$A_1 = 1, \quad A_2 = \frac{2r_0}{1+r_0}, \quad A_3 = \frac{r_0(1+r_{90})}{r_{90}(1+r_0)}, \quad A_4 = \frac{(r_0+r_{90})(1+2r_{45})}{3r_{90}(1+r_0)} \tag{4}$$

Hill criterion is always convex, quadratic and can be written as euclidean length constraint. Another interesting criterion is the "user-friendly" one, for high strength steels proposed by Yoshida et al [12]. This criterion is a sixth-order-polynomial with a high flexibility of describing anisotropic behavior of steel sheets using 16 coefficients. It is always convex and in good agreement with experimental results. For plane stress it reads:

$$\begin{aligned} &C_1\sigma_{xx}^6 - 3C_2\sigma_{xx}^5\sigma_{yy} + 6C_3\sigma_{xx}^4\sigma_{yy}^2 - 7C_4\sigma_{xx}^3\sigma_{yy}^3 + 6C_5\sigma_{xx}^2\sigma_{yy}^4 - 3C_6\sigma_{xx}\sigma_{yy}^5 \\ &+C_7\sigma_{yy}^6 + 9\left(C_8\sigma_{xx}^4 - 2C_9\sigma_{xx}^3\sigma_{yy} + 3C_{10}\sigma_{xx}^2\sigma_{yy}^2 - 2C_{11}\sigma_{xx}\sigma_{yy}^3 + C_1 2\sigma_{yy}^4\right)\sigma_{xy}^2 \\ &+27\left(C_{13}\sigma_{xx}^2 - C_{14}\sigma_{xx}\sigma_{yy} + C_{15}\sigma_{yy}^2\right)\sigma_{xy}^4 + 27C_{16}\sigma_{xy}^6 = \sigma_y^6 \end{aligned} \tag{5}$$

Parameters $C_{1,\ldots,16}$ are given as functions of some anisotropic coefficients described in detail in the Appendix of reference [12]. If the anisotropy parameters for Hill or Yoshida criterion are set equal to ones, the criteria become equivalent to von Mises.

4 NUMERICAL EXAMPLE

The numerical example concerns the construction of the limit analysis locus for a square plate with a central hole with $d/L = 0.20$ (Fig. 1) subjected to loads P_1 and P_2. This example has been widely studied in the literature [1-4]. 880 CST elements were used to model the structure; due to symmetry only one quarter of the plate was examined. Material is high strength steel $HSS590$ with yield stress equal to $590MPa$. X is considered as the rolling direction.

Three yield criteria are used, namely von Mises, Hill and the one proposed by Yoshida in their original nonlinear form. Geometrical nonlinearities and damage were not considered in the present work. Open source optimization package IPOPT [14] was used to solve the arising nonlinear programming problems.

To construct the locus a series of limit analysis problems have to be solved. Let us consider the two basic load cases P_1 and P_2. Parameter $\theta \in [0^o, 90^o]$ is used to variate the load pattern. For $\theta = 0^o$ only P_1 load is applied, while for $\theta = 90^o$ only P_2 is applied.

$$\mathcal{V}_\theta = \{P : P = P_1\,cos(\theta) + P_2\,sin(\theta),\ \ P_1 = P_2 = 100\,MPa\} \tag{6}$$

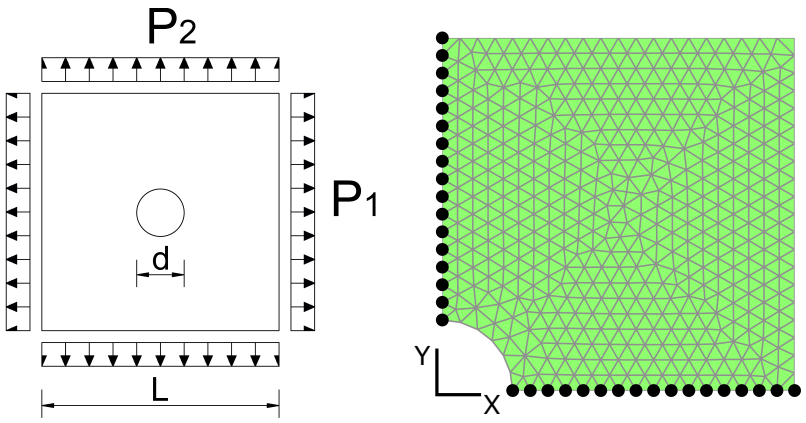


Figure 1: Square plate with central hole and FEM model.

For the von Mises yield criterion the only required parameter is the tensile yield stress. While for the Hill yield criterion anisotropy parameters $A_{1,\ldots,4}$ have to be defined.

For our case:

$$r_0 = 0.43, \quad r_{45} = 1.41, \quad r_{90} = 0.61$$

$$A_1 = 1.0000, \quad A_2 = 0.6014, \quad A_3 = 0.7936, \quad A_4 = 1.5181$$

For the Yoshida yield criterion the 16 anisotropy parameters as described in [12] for $HSS590$, are presented in Table 1.

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
1.0000	0.6014	0.4841	0.3982	0.4375	0.5752	0.7591	1.0350
C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
0.6789	0.6664	0.7540	0.9423	1.1601	1.0615	1.2470	1.7732

Table 1: Yoshida yield criterion anisotropy parameters.

Figure 2 depicts the three different loci constructed for the three yield criteria. Values on the figure are normalized using the uniaxial tensile yield stress. Significant differences of the safety factor (not always conservative) exist. Table 2 includes some of the limit analysis results for different values of parameter θ. Fourth and sixth column depict the difference compared to the results obtained using von Mises yield criterion.

Angle θ	von Mises	Hill	Diff. (%)	Yoshida	Diff. (%)
0°	0.813	0.808	-0.54	0.810	-0.33
15°	0.917	0.874	-4.67	0.881	-3.87
30°	1.062	0.982	-7.57	1.005	-5.40
45°	1.272	1.145	-9.99	1.222	-3.92
60°	1.062	1.097	+3.26	1.094	+3.03
75°	0.917	0.986	+7.57	0.942	+2.81
90°	0.813	0.906	+11.61	0.849	+4.53

Table 2: Limit analysis results for different θ.

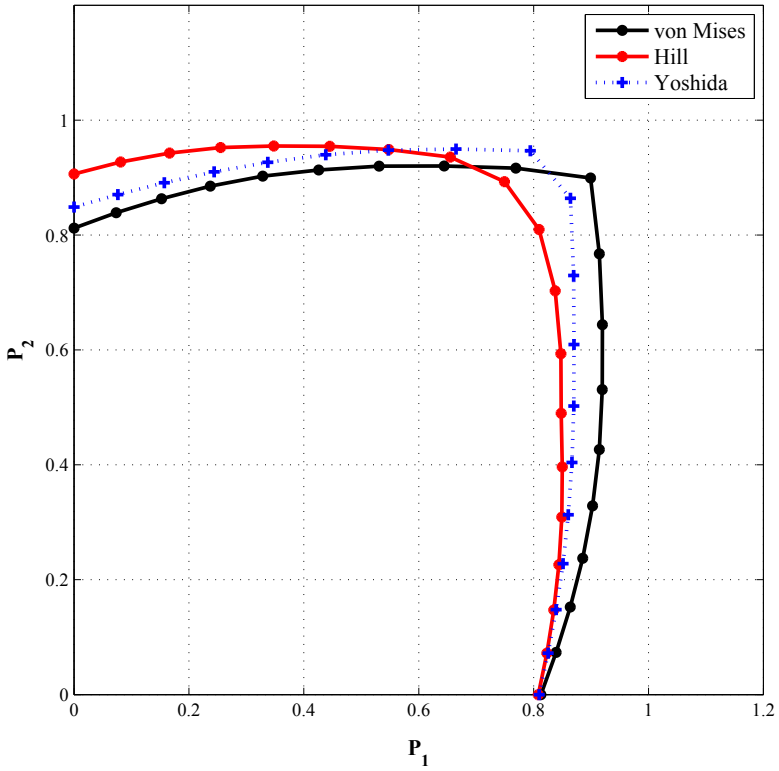


Figure 2: Limit analysis locus.

5 CONCLUSIONS

In the present work limit analysis is applied to high strength steel structures, exhibiting strong plastic anisotropy. Limit analysis requires only the equations of equilibrium and a respective yield criterion to describe the safety margins of a structure, since the material elastic properties are irrelevant to the collapse load analysis. So, in limit analysis anisotropy is described using only an adequate non-linear yield criterion like for example the popular Hill criterion, a generalization of von Mises, with parameters that are easily determined. Moreover, a polynomial yield criterion, with high flexibility of describing the anisotropic behaviour proposed by Yoshida et al was also used to illustrate the differences in the elastoplastic behaviour of high strength steels. Significant variations of the safety factor were observed for the numerical example studied, depending on the yield criterion used.

REFERENCES

[1] W.T. Koiter, *General theorems for elastic-plastic solids*, Amsterdam, 1960.

[2] M.Z. Cohn and G. Maier (Eds.), *Engineering Plasticity by Mathematical Programming: Proceedings of the NATO Advanced Study Institute, University of Waterloo*, Pergamon Press, 1979.

[3] D. Weichert and K. Spiliopoulos (Eds.), *Direct methods for limit states of structures and materials*, Springer, 2014.

[4] P. Fuschi, A.A. Pisano and D. Weichert (Eds.), *Direct methods for limit and shakedown analysis of structures*, Springer, 2015.

[5] M. Staat and M. Heitzer, *Numerical methods for limit and shakedown analysis*, Series of John von Neumann institute for computing, 2003.

[6] A. Capsoni, L. Corradi and P. Vena. Limit analysis of anisotropic structures based on the kinematic theorem. *International Journal of Plasticity*, **17**(11), 1531-1549, 2001.

[7] L. Corradi, L. Luzzi and P. Vena. Finite element limit analysis of anisotropic structures. *Computer methods in applied mechanics and engineering*, **195**(41), 5422-5436, 2006.

[8] A.A. Pisano and P. Fuschi. A numerical approach for limit analysis of orthotropic composite laminates. *International journal for numerical methods in engineering*, **70**(1), 71-93, 2007.

[9] H.X. Li and H.S. Yu. Limit analysis of composite materials based on an ellipsoid yield criterion. *International journal of plasticity*, **22**(10), 1962-1987, 2006.

[10] R. von Mises. Mechanik der festen krper im plastisch deformablen zustand.*Nachrichten von der Gesellschaft der Wissenschaften zu Gttingen, Mathematisch-Physikalische Klasse* 582-592, 1913.

[11] R. Hill. A theory of the yielding and plastic flow of anisotropic metals. *In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **193**, 281-297, 1948.

[12] F. Yoshida, H. Hamasaki and T. Uemori. A user-friendly 3d yield function to describe anisotropy of steel sheets. *International Journal of Plasticity*, **45**, 119-139, 2013.

[13] E.D. Andersen, C. Roos and T. Terlaky. On implementing a primal-dual interior point method for conic quadratic optimization. *Mathematical Programming*, **95**, 249-77, 2003.

[14] A. Waechter and L.T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large scale non linear programming. *Mathematical Programming*, **106**, 25-57, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

PARTICLE BASED SIMULATION OF FLUID FLOW IN PERIODICALLY GROOVED CHANNELS

Dorothea Kasiteropoulou, Theodoros Karakasidis, and Antonios Liakopoulos

Hydromechanics and Environmental Engineering Laboratory,

School of Engineering University of Thessaly, 38334 Pedion Areos, Volos, Greece

dkasiter@uth.gr, thkarak@uth.gr, aliakop@uth.gr

Keywords: Dissipative Particle Dynamics, coarse-graining, fluid/wall interactions, protrusion size effect, friction factor, cut-off radius.

Abstract. *The method of Dissipative Particle Dynamics is applied to investigate the effect of the parameters employed to model the presence of periodic rectangular wall roughness on planar nanochannel flow. The parameters considered here include the fluid/wall interactions, the range of interaction of fluid particles and wall particles (cut-off radius), the external applied force and the coarse-graining parameter (number of atoms per DPD particle). Protrusions of upper wall are modeled by periodically spaced rectangular protruding elements. The dependence of flow pattern on protrusion length and the amplitude of these parameters is investigated. The computed macroscopic quantities of practical interest include density, velocity, and pressure fields. Fluid particle localization near the solid wall is affected by the conservative force and the cut-off radius. Fluid velocity reduces as the protrusion length decreases for constant parameters and reduces as both the conservative force and the cutoff radius increases. The pressure is uniform across most of the channel and their pattern near and inside the cavities depend on the protrusion shape, the conservative force and cut-off radius. For the coarse graining parameter, the density and pressure remain almost constant in the core of the channel and their pattern near and inside the cavities depends on the protrusion shape.*

1 INTRODUCTION

An adequate description of fluid flow over a solid boundary with surface roughness on multiple length scales often requires resolution of fine microscopic details on the flow structure while retaining peculiarities of a macroscopic picture [1]. Starting from the molecular/particle description, which is employed to model a small part of the computational domain where the continuum models fail to capture the physics of the system, there are many reports on wall structure, wall wettabillity and stiffness, the fluid-wall interface and also the combination of these parameters. For example, Davies et al. [2] investigated the laminar flow in a microchannel with superhydrophobic walls exhibiting transverse ribs and found that reductions in the friction factor and enhancements in the fluid slip are greater as the cavity-to-rib length ratio is increased (increasing shear-free fraction) and as the channel hydraulic diameter is decreased. Niavarani and Priezjev [3], investigated the combined effect of surface roughness and shear rate on slip flow of simple fluids by molecular dynamics simulation and found that in the region where the curved boundary faces the mainstream flow, the local slip is suppressed due to the increase in pressure. Kamali and Kharazmi [4] studied the surface roughness effects on the fluid flow in a nanochannel of simple fluids in hydrophobic and hydrophilic walls by molecular dynamic simulation. The simulation results show that the roughness and protrusions of the same dimensions induce different local density pattern while the overall average might be the same. Asproulis and Drikakis [5] investigated the effects of surface roughness on the flow behavior inside micro and nanofluidic devices. They found that as the surface attraction energy or the roughness height increase, the density layering near the wall is enhanced by higher values or secondary layering phenomena. Kasiteropoulou et al. [6] discussed the relation of friction factor, f, with the flow Reynolds number in a dissipative particle dynamics study of flow in periodically grooved nanochannels and found that the computed value for friction factor increases as the length decreases for the same Reynolds number. Sun et al. [7] investigated the effects of surface wettability and topology in the multi-scale liquid flow in micro/nanochannels.

Analysis on the flow friction shows that the pressure gradient decreases in a power law with increasing channel height. The confinement on the liquid molecules will equivalently narrow the channel, where larger pressure gradient is needed to keep the flow conditions from changing. The larger the roughness is, the more obvious the influence will be. When the channel height becomes larger, both velocity slip and relative slip length will gradually converge to 0, which means that different flow boundaries will be unified to be non-slip at conventional spatial scale. Priezjev [8] reported the results obtained from Molecular Dynamic simulations of the friction at an interface between polymer melts and weakly attractive crystalline surfaces. He found that the friction coefficient at small slip velocities exhibits a distinct maximum which appears due to shear – induced alignment of semiflexible chain segments in contact with solid walls and at large slip velocities, the friction coefficient is independent of the chain stiffness. Later, Chen et al. [9] estimated the velocity slip on curved surfaces in a Couette flow studied by molecular dynamics simulation. They found that the slip length as conventionally measured at a flat wall in Couette flow is the same as that for all other cases with curved and rotating boundaries, provided the atomic interactions are the same and boundary shape is properly taken into account. Priezjev [10] investigated the effect of interfacial slip on steady-state and time-periodic flows of monatomic liquids by using non-equilibrium molecular dynamics simulations. They found that the velocity profiles in oscillatory flows are well described by the Stokes flow solution with the slip length that depends on the local shear rate. For both types of flows, the friction coefficient at the liquid–solid interface correlates well with the structure of the first fluid layer near the solid wall.

Sofos et al. [11] studied the wetability and stiffness on diffusion in grooved nanochannel and observed a significant anisotropy along the directions parallel and normal to the flow inside the grooves.

On the other hand, at macroscale description slip or wettabillity parameters are not usually performed in numerical simulations whereas the surface construction, such as the protruding elements mentioned above, seem to concern the scientific community for many years. We report here some examples of these investigations. Sahan et al. [12] investigated the two-dimensional isothermal flow in a periodically grooved channel. They presented low-order models successfully which describe the dynamical characteristics of the flow for Re close to the design conditions. They found that far from the design conditions, the reduced models predict quasi-periodic or period-doubling routes to chaos as Re is increased [13]. Greiner et al. [14] investigated the three-dimensional flow in a channel with symmetric, transverse grooves on two opposite walls by using the spectral element technique. They observed the transition from steady two- dimensional flow in the three-dimensional mixing through two and three dimensional wave structures as the Reynolds number was increased. Chung et al. [15] studied the unsteady laminar flow in grooved channel and also in a sharp 180° bend for low Reynolds number. This model is used in the cooling of high performance modern electronic systems [16]. Adachi & Hasegawa [17] investigated the flow influence of the number of grooves in which the flow repeats periodically. Above these, flows around and inside grooves at the larger dimensions can also present some other interesting dynamic flow phenomena. In particular, when the flow increases its speed can experience shock waves and shock/vortex interactions. Large eddy simulations can be used for understanding such phenomena [18, 19] and can also be combined with MD [20, 21] to shed light on scale-up effects.

In the present work we study the flow in a nanochannel of height 4.72nm and seek further insight on the effects of wall protrusions and parameters inserted on the system on the flow rate based on mesoscopic Dissipative Particle Dynamics simulations. Protrusions are introduced by periodically placing rectangular protruding elements on one of the two channel walls. Pressure, density and velocity profiles are calculated for different protrusion lengths. Protrusion size, fluid-wall interaction, cut-off radius and coarse-graining parameter plays an important role on flow properties such as density, velocity and pressure distributions.

We remind here the reader that in the nanochannel cases the coarse graining parameter is equal to 1 (an atom per DPD particle) while in the microchannel cases is $2.5x10^6$. The corresponding heights are set equal to 4.72 nm for the nanochannel and 1.1μm for the microchannel.

This paper is set up as follows. Section 2, gives a description of the system and the simulation method. In Section 3 the results are shown and analyzed, whereas section 4 contains concluding remarks.

2 SIMULATION METHOD

2.1 Channel Geometry

We study pressure-driven (or body-force driven) flow between two parallel walls. The lower wall is flat whereas the upper wall consists of protrusions modeled by rectangular elements (see Fig. 1). We considered protrusions of three length values and one height at the upper wall: $l_{r_1}=0.5l_{tot}$, $l_{r_2}=0.25l_{tot}$, $l_{r_3}=0.125l_{tot}$ *or* $l_{r_3}=0.167l_{tot}$ for the microchannel case (protrusion length) and h=0.10H *or* h=0.125 $\bar{h}$ (protrusion height), where l_{ri} and h represent the length

and the height of the rectangular grooves on the upper wall and $l_{tot,}$ H represent the length of the computational domain along the x-direction and it's height in the y-direction respectively and $\bar{h}$ represents the distance between the solid walls. Therefore case $l_{r1}=0.5l_{tot}$ corresponds to a computational domain with one rectangular groove, $l_{r2} = 0.25l_{tot}$ corresponds to two rectangular grooves, and $l_{r3} = 0.125l_{tot}$/ $l_{r3}=0.167l_{tot}$ corresponds to three/four rectangular grooves. Baseline dimensions of all computational domains in x-, y- and z-directions are 3.77nm x 4.72nm x 3.77nm for the nano- and L_x x L_y x L_z = 1.11 μm x 1.3μm x 1.1 μm for the microchannel (in DPD units the dimensions are L_x x L_y x L_z = 11.09 r_c x 13.87 r_c x 11.09 r_c and 11.14 r_c 13.00 r_c x 11.14 r_c respectively). Conversion from DPD units to physical units is discussed in detail in Kumar et al. [32].

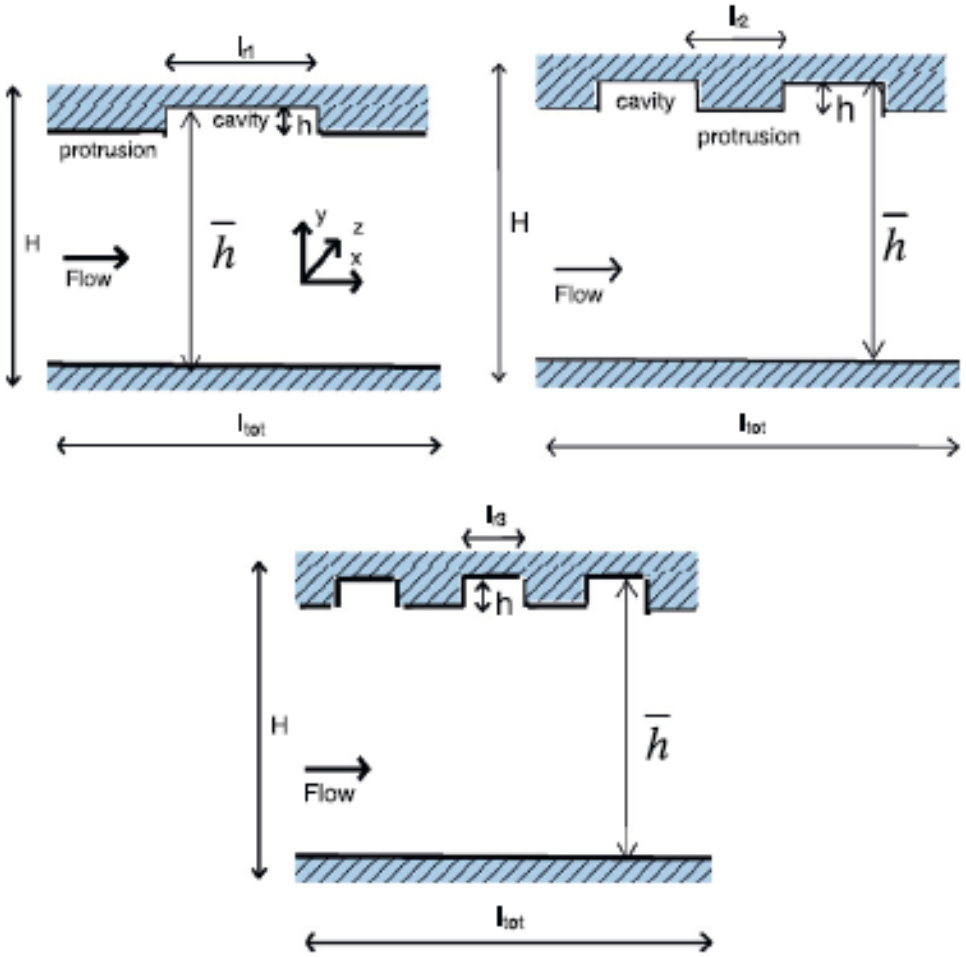


Fig. 1. Schematics of computational domains (adapted from Kasiteropoulou et al. [35]).

2.2 Mathematical Model and Computational details

2.2.1. Dissipative Particle Dynamics

The DPD system consists of N particles. For the i-th particle we denote mass m_i, position $\mathbf{r}_i$ and velocity $\boldsymbol{\upsilon}_i$, i=1,2,…N. For a single-component DPD liquid the forces exerted on a particle *i* due to particle *j* consists of three terms: 1) the conservative force $\mathbf{F}_{ij}^{C}$, 2) the dissipative force $\mathbf{F}_{ij}^{D}$ and 3) a random force $\mathbf{F}_{ij}^{R}$, given by,

$$\mathbf{F}_{ij}^{C}\left(r_{ij}\right)=\begin{cases}\alpha_{ij}\left(1-r_{ij}/r_{c}\right)\hat{\mathbf{r}}_{ij}, & for \quad r_{ij} \le r_{c}\\ 0, & for \quad r_{ij} > r_{c}\end{cases} \tag{1}$$

$$\mathbf{F}_{ij}^{D}=-\gamma\omega^{D}(r_{ij})(\mathbf{\upsilon}_{ij}\cdot\hat{\mathbf{r}}_{ij})\hat{\mathbf{r}}_{ij} \tag{2}$$

$$\mathbf{F}_{ij}^{R}=\sigma\omega^{R}(r_{ij})\xi_{ij}\hat{\mathbf{r}}_{ij} \tag{3}$$

where $\mathbf{r}_{ij}=\mathbf{r}_{i}-\mathbf{r}_{j}$, $r_{ij}=\left|\mathbf{r}_{ij}\right|$, $\hat{\mathbf{r}}_{ij}=\dfrac{\mathbf{r}_{ij}}{\left|\mathbf{r}_{ij}\right|}$, $\mathbf{\upsilon}_{ij}=\mathbf{\upsilon}_{i}-\mathbf{\upsilon}_{j}$ [22] and α_{ij} is the maximum repulsion between particles i and j [23]. The coefficients γ and σ determine the amplitude of the dissipative and random forces, respectively, while ω^{D} and ω^{R} are appropriate weight functions [24]. The weight functions ω^{D} and ω^{R} provide the range of interaction for the dissipative and random forces [25]. In Eq. (3), ξ_{ij} is a random variable with Gaussian statistics [23]. By enforcing $\xi_{ij}=\xi_{ji}$ one satisfies the principle of momentum conservation [26]. All the forces between particle i and j vanish beyond a cut-off radius r_{c} [27]. The random force coefficient, σ, the system temperature and the simulation timestep is the same for all cases studied here and their value was chosen following the methodology proposed by Groot and Warren [23].

The requirement of canonical distribution sets two conditions linking the random and dissipative forces. The first one couples the weight functions through

$$\omega^{D}\left(r_{ij}\right)=\left[\omega^{R}\left(r_{ij}\right)\right]^{2} \tag{4}$$

and the second one the strengths of the random and dissipative forces via

$$\sigma^{2}=2\gamma k_{B}T \tag{5}$$

where k_{B} is the Boltzmann constant [27]. The typical choice for the weight functions is

$$\omega^{R}\left(r_{ij}\right)=\begin{cases}\left(1-r_{ij}/r_{c}\right)^{p}, & for \quad r_{ij} \le r_{c}\\ 0, & for \quad r_{ij} > r_{c}\end{cases} \tag{6}$$

where $p=1$, for the standard DPD method [28].

The time evolution of velocities and positions of particles are described by the following equations [29]

$$d\mathbf{r}_{i}=\mathbf{\upsilon}_{i}dt \tag{7}$$

$$d\mathbf{\upsilon}_{i}=\frac{1}{m}\left(\mathbf{F}_{i}^{C}dt+\mathbf{F}_{i}^{D}dt+\mathbf{F}_{i}^{R}\sqrt{dt}\right) \tag{8}$$

where $\mathbf{F}_{i}^{C}=\sum_{i\neq j}\mathbf{F}_{ij}^{C}$, and $\mathbf{F}_{i}^{R}=\sum_{i\neq j}\mathbf{F}_{ij}^{R}$.

All simulations are conducted using the open source LAMMPS package [30]. We use different values for the parameter a_{ij} when we describe fluid–fluid interactions, wall–wall interactions and wall–fluid interactions: a_{ww} for wall–wall interactions, a_{wf} for wall–fluid interaction, a_{ff} for fluid–fluid interactions. In our simulations we have kept the fluid–fluid parameter interaction constant: a_{ff} = 25 for N_m=1 and a_{ff} = 7.5 for N_m=2.5x10^6. The value of parameter a_{ij} has been selected so that the dimensionless compressibility of the simulated DPD fluid corresponds to a typical liquid, such as water both in the nanochannel and the

microchannel case studied (for the calculation of compressibility see [23]). This value of a_{ij} is employed in fluid–fluid interactions. We remind the reader that the cross interaction term obeys the relation $\alpha_{wf}=(\alpha_{ww}\ \alpha_{ff})^{1/2}$ (see for example Pivkin and Karniadakis [29]). The type of the surface (hydrophilic or hydrophobic) is controlled via the fluid–wall interaction through the conservative force parameter a_{ij} (Equation (1)), specifically by choosing appropriately the ratio of the a_{ij} parameters for the fluid–wall interaction versus a_{ij} parameters for fluid–fluid interactions [31]. The number density of the DPD fluid, n_f, is equal to 3 for the nanochannel case and 10 for the microchannel while the random and dissipative forces, σ and γ respectively, are set equal to those described in Pivkin and Karniadakis [29]. System temperature is kept constant at T*=1 (in real units the system temperature is equal to 300K).
In this work we investigated all the parameters inserted on a DPD system along with grooved channels. In the first place we employed a_{ff} = 25 and a_{wf} = 25 with r_c = 1.0 for the cut-off radius for particle interactions (equal to 1.0σ for the nanochannel, where σ denotes the characteristic length scale of the Lennard-Jones potential (0,3402nm) and not the force parameter of Eq.5) and N_m=1 (one atom per DPD particle – nanochannel case) and we examined the effect of the magnitude of the externally applied force by assigning four values (F_{ext} = 0.01, 0.02, 0.03, 0.04 DPD units). The external driving force is applied on each particle along the x-direction to drive the flow.
In the rest of the simulations F_{ext} = 0.02 DPD units was used. For the case a_{ff} = a_{wf} = 25 and N_m=1 we explored the effect of range of interactions of DPD particles (cut-off distance) rc = 1.0σ, 1.5σ and 2.0σ. Keeping fixed the cut-off distance r_c = 1.0σ and N_m=1 we varied the values of wall–fluid interaction a_{wf} using three values of the conservative force parameter equal to a_{wf} = 25, 100, 200. Finally we changed the number of atoms per DPD particle equal to $2.5\text{x}10^6$ (microchannel case) atoms per particle, along with the particle interactions coefficient: a_{ff} = a_{wf} = a_{ww} = 7.5 and rc = 1.0σ. In the last case F_{ext} = 0.02 DPD units was also used. By changing the cut-off distance we fix the fluid type. The wall material is controlled via the fluid-wall interaction.

Periodic boundary conditions are employed along x- and z-directions. Appropriate boundary conditions need to be enforced in order to avoid that fluid particles cross the wall, since the effective forces are not sufficient to prevent wall penetration and several models have been employed and tested to impose various boundary conditions, such as specular, Maxwellian and bounce back [29-33]. In general, according to the boundary conditions employed we may observe fluctuations, near the solid walls, in the density, velocity or temperature which in cases of modeling microscopic systems may be inconsistent with continuum models. Such variations are also observed in cases of nanochannels studied with Dissipative Particle Dynamics simulations [see for example 6, 31].
In the present study we have chosen to employ the bounce back conditions, in which both components of the velocity are reversed [33]. Wall particles are bound on sites of a cubic lattice and their velocities are set equal to zero.

The simulation timestep is $\Delta t = 0.01\sqrt{mr_c^2/k_BT}$ (0.015ps for the nanochannel and $7.6\text{x}10^{-3}$μs for the microchannel in physical units). The duration of each simulation is $5\text{x}10^5$ time steps. Pressure, number density, temperature and streaming velocity bin values are averaged over the last $2.5\text{x}10^5$ time steps of the simulation.

2.3 Macroscopic property evaluations

Local macroscopic property values are calculated in parallelepiped bins. Results presented in this work are obtained by dividing the computational domain into 4x80x80, 8x80x80, 16x80x80 bins along the direction x, y, z respectively for post processing. To extract average profiles (for pressure, density, velocity), mean values over space are computed at each layer along the channel for each time step, and these values are then averaged over time.

Pressure values are obtained from the trace of the stress tensor

$$p = -\frac{1}{3} tr\mathbf{S} \tag{9}$$

where the stress tensor, $\mathbf{S}$, is calculated using the Irving– Kirkwood theory [34,35]:

$$\mathbf{S} = -\frac{1}{V}\left[m\sum_i (\boldsymbol{\upsilon}_i - \boldsymbol{\upsilon})(\boldsymbol{\upsilon}_i - \boldsymbol{\upsilon}) + \frac{1}{2}\sum_i \sum_{j\neq i} \mathbf{r}_{ij}\mathbf{F}_{ij} \right] \tag{10}$$

where V is the volume of the computational bin, $\boldsymbol{\upsilon}$ is the corresponding stream velocity and $\mathbf{F}_{ij}$ is the interparticle force on particle i due to particle j:

$$\mathbf{F}_{ij} = \mathbf{F}_{ij}^{C} + \mathbf{F}_{ij}^{D} + \mathbf{F}_{ij}^{R} \tag{11}$$

3 RESULTS AND DISCUSSION

3.1 Nano-channel results: number density, flow velocity and pressure

In this section we present results concerning the variation of fluid number density (particle localization), velocity and pressure as a function of variation of fluid/wall interaction, the range of particle interactions (variation of cut-off radius) as well as the effect of the driving force.

3.1.1. Particle localization

The average number density profiles as a function of the fluid/wall interaction are presented in Fig. 2a for a typical protrusion length. We observe that as the conservative force increases the number density inside the cavities and near the walls become higher and more distant from the wall, a behavior which is called layering (for a more detailed description see Kasiteropoulou et al. [31]). This behavior is in qualitative agreement with the results of Sofos et al. [36] who investigated the surface wettability effects on flow in rough nanochannels and found that as the wall hydrophobicity decreases there is a kind of increased fluid atom localization. Actually we observe that the wall surface behaves more like a hydrophobic one as the conservative force parameter increases. In the lower flat wall the number density is higher for all cases studied here than the number density at the upper wall and this could be explained by the presence of the wall protrusions.

Number density profiles as a function of the distance of the particle interactions (cut-off distance) are presented in Fig.2b. As we can see the number of peak density inside the cavity regions depend on the cut-off distance. In fact, for cut-off distance r_c=2.0σ there are two number density peaks which are decreased as the cut-off distance decreases. Moreover we observe that near the lower and protrusion wall the increased range of interactions results in more important layering effect and density variations extending over larger distances. This behavior is in general detected in all the protrusion length investigated in this work and this

means that the fluid is less homogeneous as the range of interactions increases.

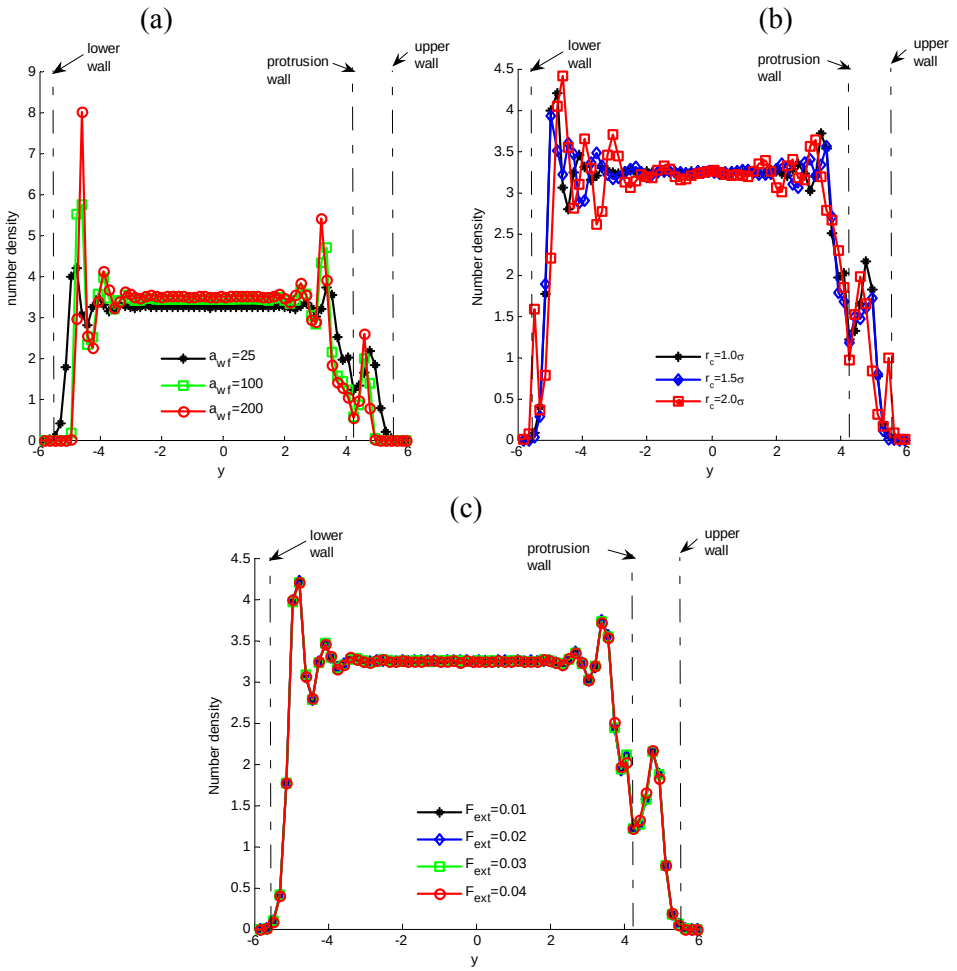


Fig. 2. Average number density profiles for $l_{r3}=0.125l_{tot}$ obtained by varying (a) wall–fluid interactions, a_{wf} (parameters kept constant: the cut-off radius $r_c = 1.0\sigma$, external driving force $F_{ext} = 0.02$, $a_{ff} = 25$), (b) the cut-off radius, r_c (parameters kept constant $a_{ff} = a_{wf} = 25$, and $F_{ext} = 0.02$), (c) the external driving force, F_{ext} (parameters kept constant $a_{ff} = a_{wf} = 25$, and $r_c = 1.0\sigma$). N_m=1. Dash-dot lines denote solid wall limits and y is in r_c units.

The external driving force seems to have no effect on the fluid particle localization (Fig. 2c) for all grooved channels, at least for the force range examined in the present work. Similar results are also reported by Young et al. [37] who reported that for a given surface wettability the fluid densities are independent of the driving force in a nanochannel flow.

The influence of the protrusion length for constant conservative force parameter is presented in Fig.3. We observe that number density inside the cavities and near the protrusion walls are reduced as the protrusion length decreases.

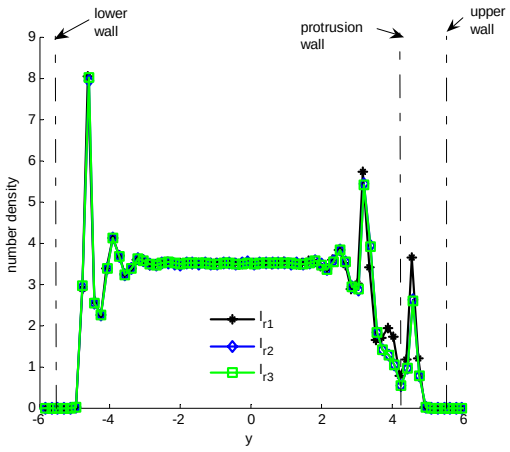
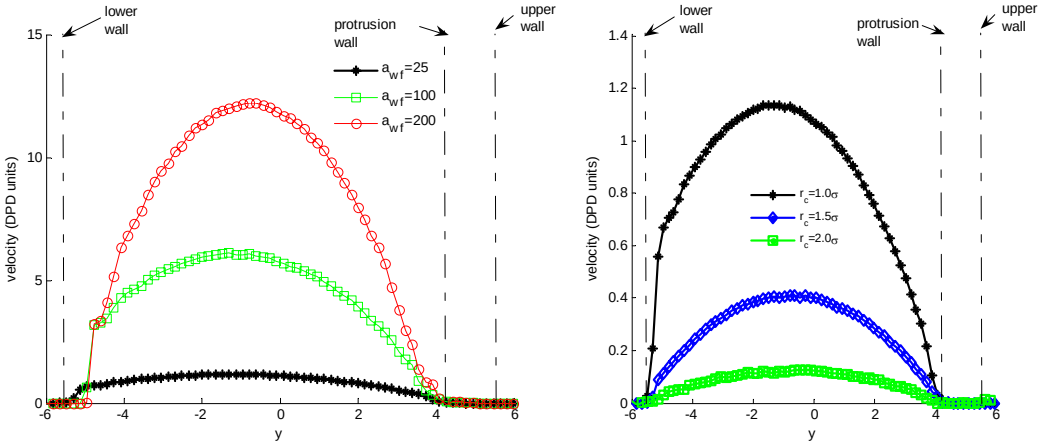


Fig. 3 Average number density profiles for a) $l_{r1}=0.5l_{tot}$, b) $l_{r2}=0.25l_{tot}$, c) $l_{r3}=0.125l_{tot}$ obtained by parameters a_{wf} = 200, a_{ff} = 25, r_c = 1.0σ and F_{ext} = 0.02. N_m=1. Dash-dot lines denote solid wall limits and y is in r_c units.

3.1.2. Flow velocity

The average velocity as a function of the conservative force parameter is presented in Fig. 4a. We observe that in the main core of the channel the flow velocity increases significantly as the fluid/wall interaction also increases and this behavior is also presented in the work of Sofos et al [36]. This behavior can be explained by the equations (1) and (8). As the conservative force parameter increases the total conservative force also increases (equation (1)) and this leads in higher velocity values (according to equation (8)).

(a) (b)



(c)

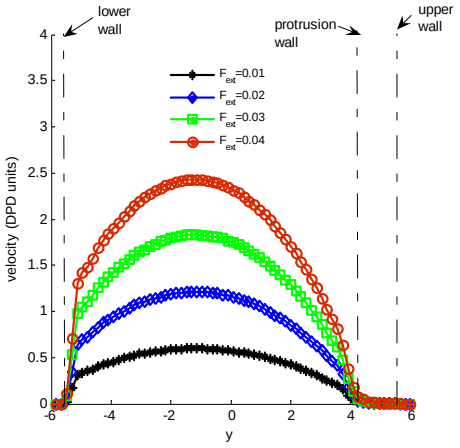


Fig. 4. Average velocity profiles for l_{r2}=0. 25l_{tot} obtained by varying (a) wall–fluid interactions, a_{wf} (parameters kept constant: the cut-off radius r_c = 1.0σ, external driving force F_{ext} = 0.02, a_{ff} = 25), (b) the cut-off radius, r_c (parameters kept constant a_{ff} = a_{wf} = 25, and F_{ext} = 0.02), (c) the external driving force, F_{ext} (parameters kept constant a_{ff} = a_{wf} = 25, and r_c = 1.0σ). N_m=1. Dash-dot lines denote solid wall limits and y is in r_c units.

As far as the range of interaction of the fluid and the wall particles (cut-off distance) concerned, we observe that the average velocity (Fig. 4b) is decreased as the cut-off radius increases.

The average velocity increases systematically as the external driving force increases (Fig. 4c). These results are in qualitative agreement with the results of Priezjev [38] where the effect of surface roughness on rate-dependent slip in simple fluids has been investigated.

3.1.3. Pressure profiles

The pressure profiles of all channel cases are presented in Fig.5 as a function of all the parameters investigated in this work. For all these parameters the pressure increases both inside the cavities and in the core of the channel as the parameters' value increases. The higher values of pressure are detected in the case of the cut-off distance increment (r_c=2.0σ) and also in the higher conservative force value (a_{wf}=200). In fact the pressure increases as the range of interaction increases among fluid as well as the fluid and wall particles. On the other hand, the conservative force parameter increases between the fluid and wall particles and this leads in higher conservative total force (according to Eq.6 and 10) and though higher pressure (according to Eq. 10). The pressure profile for various external driving force magnitudes is presented in Fig. 5c. Both in the core of the channel and inside the cavities the pressure peaks are detected in the same location for all external driving force magnitudes. The pressure increases systematically as the driving force increases but this behavior is of small interest relative to the other parameters described above.

The variation of protrusion lengths reveals some interesting characteristics. The isobars for constant conservative force parameter, equal to a_{wf}=25, are presented in Fig. 6. For case l_{r1} (Figure 6(a)), pressure reduces inside the cavities adjacent to the upper wall. However, there is a region of high pressure inside the cavities surrounded by two regions of low pressure. We believe that such high-pressure regions are observed because of the trapping of the fluid particles inside the cavities and the increased localization, as described is Section 3.1.1 (see also Figure 2(a)) [6]. This behavior results in smaller interparticle distances and thus higher interparticle forces leading to an increase of pressure. Adjacent to the protrusion wall, pressure is reduced, especially for smaller protrusion lengths, because particles are not localized there. For cases l_{r2} and l_{r3} (Figure 6(b) and 6(c)), at the rough wall, inside the

cavities, we observe that the shape of high-pressure regions depend on the protrusion length. More specifically, as the protrusion length decreases the length of the high pressure region increases.

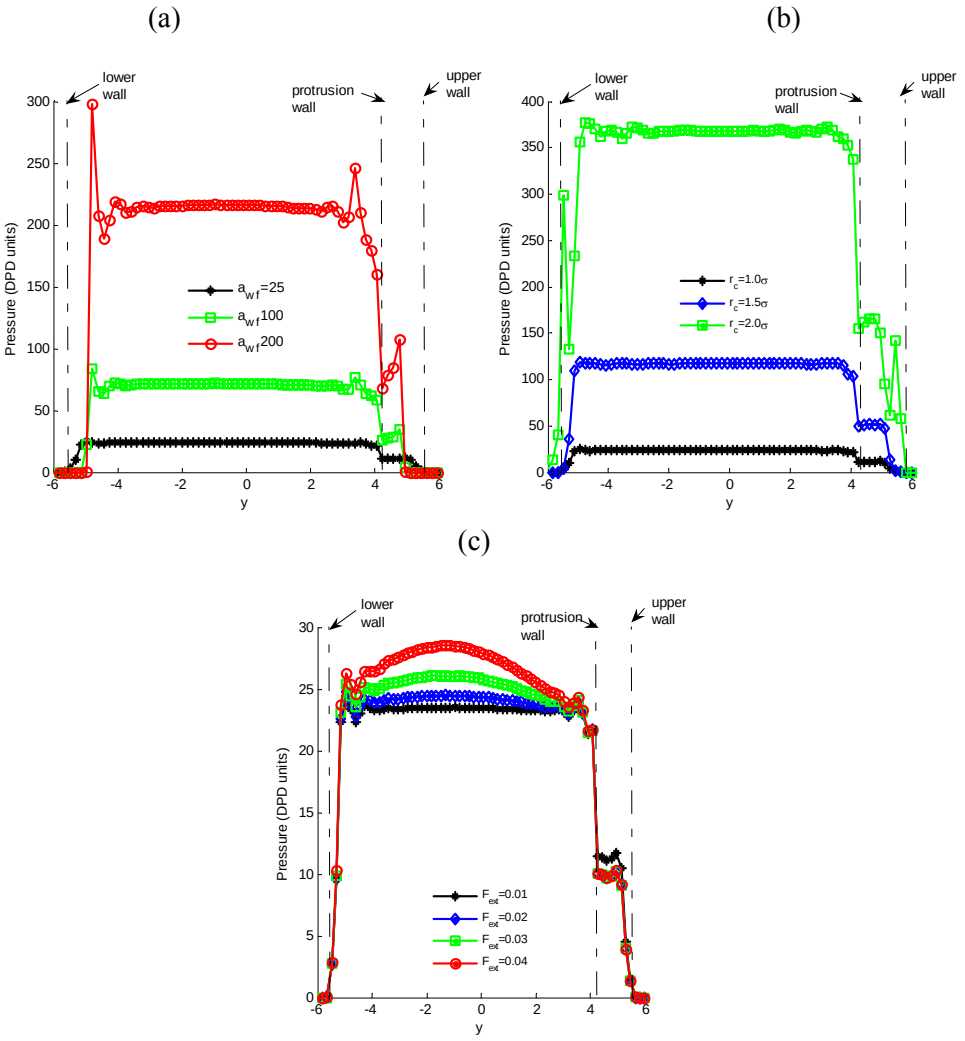
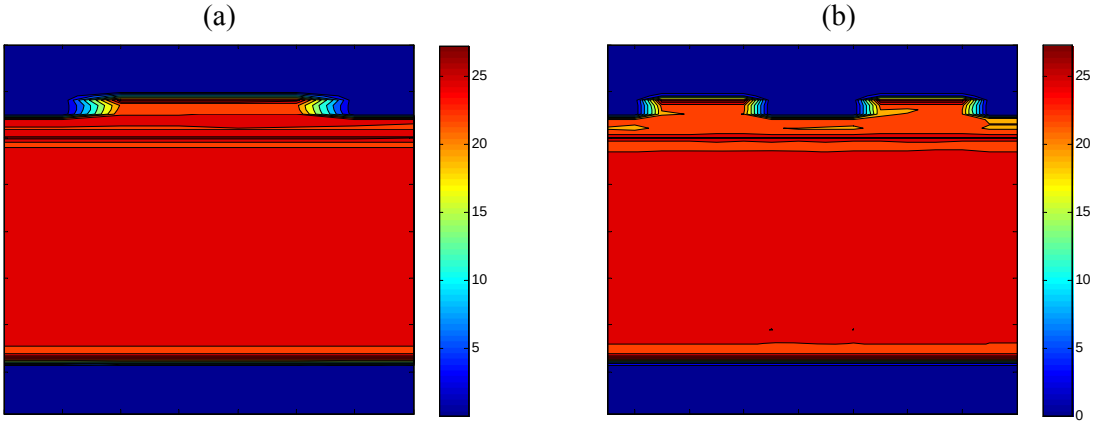


Fig. 5. Average pressure profiles for l_{r3}=0. 125l_{tot} obtained by varying (a) wall–fluid interactions, a_{wf} (parameters kept constant: the cut-off radius r_c = 1.0σ, external driving force F_{ext} = 0.02, a_{ff} = 25), (b) the cut-off radius, r_c (parameters kept constant a_{ff} = a_{wf} = 25, and F_{ext} = 0.02), (c) the external driving force, F_{ext} (parameters kept constant a_{ff} = a_{wf} = 25, and r_c = 1.0σ). N_m=1. Dash-dot lines denote solid wall limits and y is in r_c units.



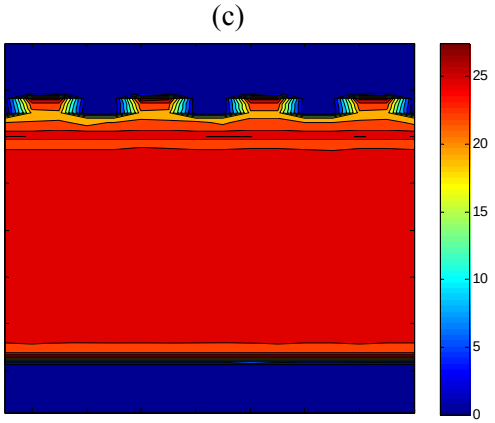


Fig. 6 Isobars for a) $l_{r1}=0.5l_{tot}$, b) $l_{r2}=0.25l_{tot}$, c) $l_{r3}=0.125l_{tot}$ obtained by parameters $a_{wf}=25$, $a_{ff}=25$, $r_c=1.0\sigma$ and $F_{ext}=0.02$. $N_m=1$. Dash-dot lines denote solid wall limits and y is in r_c units.

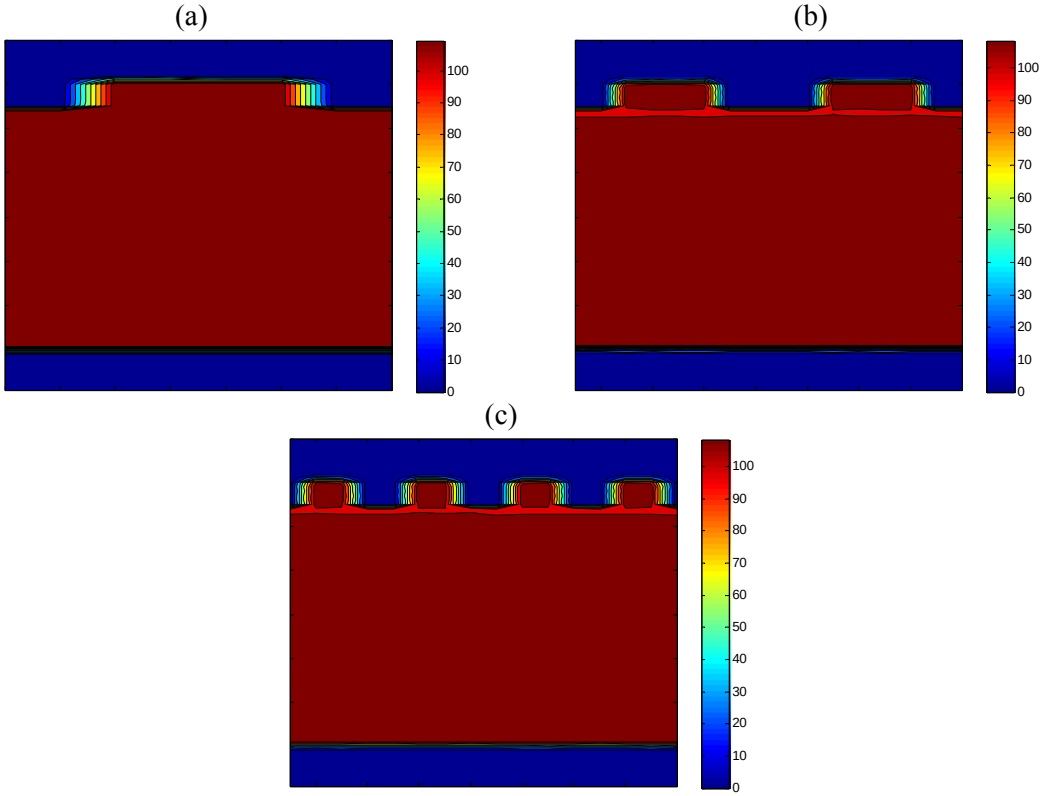


Fig. 7 Isobars for a) $l_{r1}=0.5l_{tot}$, b) $l_{r2}=0.25l_{tot}$, c) $l_{r3}=0.125l_{tot}$ obtained by parameters $a_{wf}=25$, $a_{ff}=25$, $r_c=1.5\sigma$ and $F_{ext}=0.02$. $N_m=1$. Dash-dot lines denote solid wall limits and y is in r_c units.

The isobars for constant cut-off distance equal to $r_c=1.5\sigma$ are presented in Fig. 7. Again, the pressure inside the cavities depends on the protrusion length. Although both inside and in the core of the channel the pressure is equal for the three protrusion lengths within statistical error, the distribution of the pressure inside the cavities is affected by the length. More thoroughly, for protrusion length equal to l_{r1} the high pressure region length exceeds all the protrusion height, whereas this height decreases slightly as the protrusion length decreases (see Figs. 7b and 7c for cases l_{r2} and l_{r3} respectively). Figure 7 reveal that the protrusion shape has small influence in the pressure map for constant cut-off distance.

3.2 Micro-channel results: number density, flow velocity, velocity slip, pressure and temperature

In this section we present results concerning the variation of fluid number density (particle localization), velocity and pressure as a function of the coarse-graining parameter by keeping constant the external driving force and the 'dimensionless' cut-off distance. By changing the coarse-graining parameter from value 1.0 to value $2.5x10^6$ we move through the scale of the channel and actually we "reach" the microscale dimension in our simulations.

3.2.1. Particle localization

Number density profiles at the protrusion and the cavity midplanes for the microchannel case are presented in Fig. 8. Fluid particle localization is similar for the nanochannel (Fig 2), and the microchannel case with peaks located at a distance from the walls ($0.87r_c$ and $0.44r_c$ for the nanochannel and the microchannel case respectively) for the same external driving force (F_{ext}=0.02 DPD units) and the same "dimensionless" cut-off distance (r_c=1.0). For the nanochannel case, at the protrusion midplanes density is homogeneous and slightly lower than its average value in the core of the channel while at the cavity midplanes is slightly higher (for the nanochannel case see also Fig.2 from Kasiteropoulou et al. [6]).

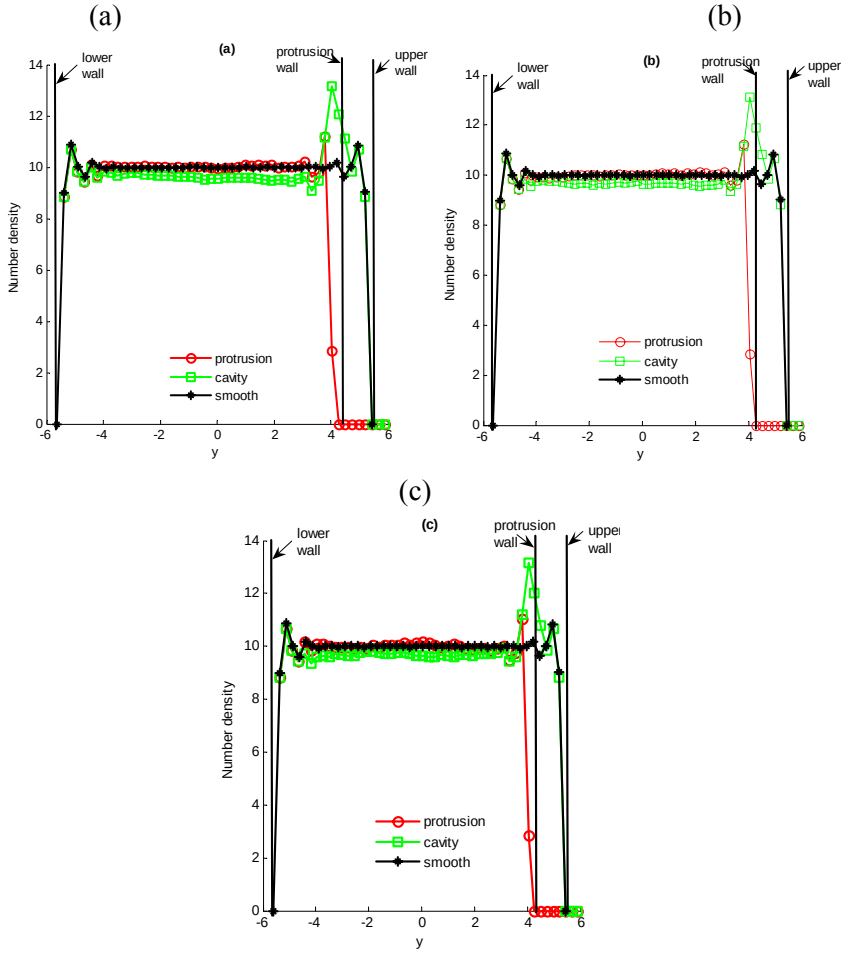


Fig. 8. Local number density profiles at the protrusion midpoints and the cavity midpoints obtained for $N_m=2.5x10^6$ and $a_{wf} = a_{ff} = a_{ww} = 7.5$ (parameters kept constant: the cut-off radius $r_c = 1.0$ and the external driving force $F_{ext} = 0.02$) for cases a) $l_{r1}=0.5l_{tot}$, b) $l_{r2}=0.25l_{tot}$, c) $l_{r3}=0.167l_{tot}$. Dash-dot lines denote solid wall limits and y is in r_c units.

In the microchannel case, at the protrusion midplanes density is almost equal to its average value in the core of the channel within statistical errors, while at the cavity midplanes density is lower. It should be also noted that in the microchannel case a high density peak is detected in the density profile very close to the protrusion surface ($y \approx 4.00 r_c$). Inside the cavities high number density regions are detected for both channel scales and suggest the possibility of particle trapping in these regions. Particle trapping is confirmed by computing the particle residence time and by the analysis of particle trajectories and is discussed in detail in Kasiteropoulou et al. [6] (nanochannel) and in Kasiteropoulou et al. [39] (microchannel).

Regarding the density fluctuations close to the solid walls, we must make some comments already discussed in [40]. For the nanoscale simulations using DPD, these oscillations are realistic. However, as the DPD particles get larger in size (the case of mesoscale simulations), these oscillations close to the wall extend over larger regions than those corresponding to physical layering zones as discussed in [41–43]. There are several methods that have been developed in order to limit such oscillations [41–43]. Thus, the observed oscillations in the mesoscale case have to be seen with some caution. Of course we must mention that if one wants to represent in detail the situation close to the wall, it is crucial to implement the above-mentioned methods. However, in our case, we are interested mostly on the qualitative flow behavior as a function of the protrusion shape and the parameters inserted on a DPD system and not to a detailed quantitative study. In a future study, it would be of interest to implement such methods and compare their results as far as the mesoscale is concerned.

An example of implementing these methods is discussed in [43]. Li et al.[43] in order to capture the correct temperature-dependence of a fluid, developed an energy-conserving dissipative particle dynamics (eDPD) model by expressing the weighting terms of the dissipative force and the random force as functions of temperature. They found that for non-isothermal fluid systems, the present model can predict the diffusivity and viscosity consistent with available experimental data of liquid water at various temperatures.

3.2.2. Flow velocity

Average velocity profiles for all protrusion lengths are presented in Fig.9. We observe that the average velocity for all grooved channels is smaller than in the channel with flat walls and actually it decreases as the protrusion length also decreases. This behavior is also detected in the nanochannel case and is described in detail in Kasiteropoulou et al [6]. Moreover, similar behavior is observed in many previous works performed with different simulation methods [44]. For example, Sofos et al [44] studied the effects of wall roughness on the flow in nanochannels by a molecular dynamics simulation and found that as the rectangular wall cavities become narrower, velocity values inside the cavities decrease and fluid atoms tend to be trapped inside them.

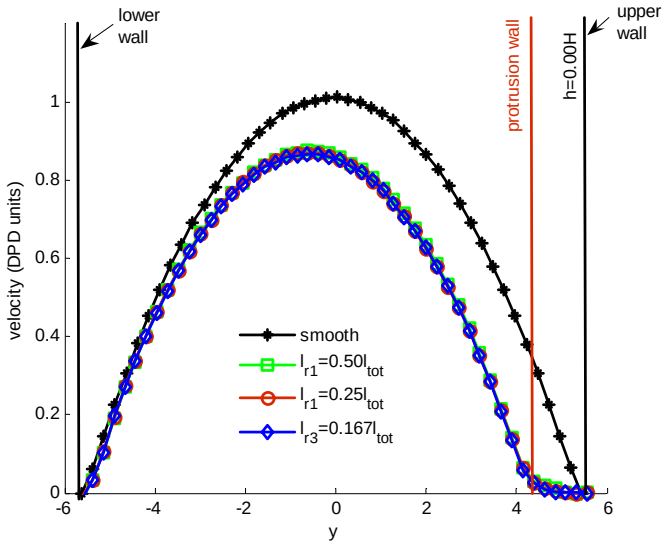
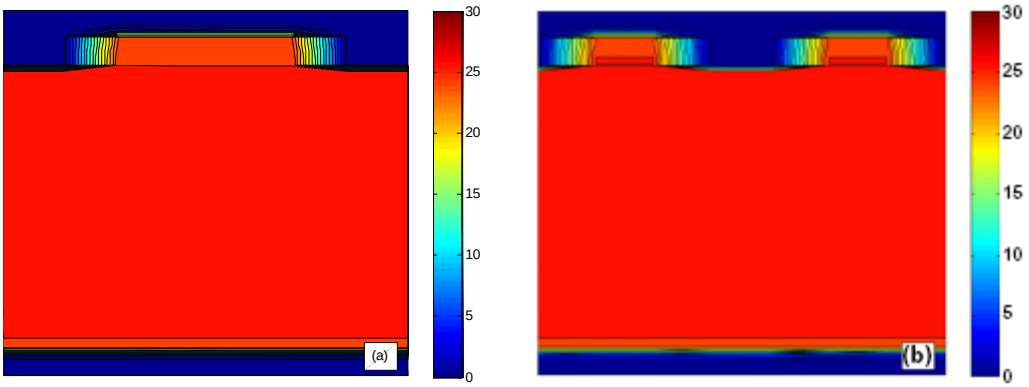


Fig. 9. Average velocity profiles obtained for N_m=2.5x10^6 and $a_{wf} = a_{ff} = a_{ww}$ = 7.5 (parameters kept constant: the cut-off radius r_c = 1.0 and the external driving force F_{ext} = 0.02) for cases a) l_{r1}=0.5l_{tot}, b) l_{r2}=0.25l_{tot}, c) l_{r3}=0.167l_{tot} . Dash-dot lines denote solid wall limits and y is in r_c units.

3.2.3. Isobars

Pressure behaviour also reveals interesting characteristics (Fig.7, 10). Both in the nanochannel case (Fig.7) as well as in the microchannel case (Fig.10), pressure remains constant in the central part of the channel and decreases as we move towards to the walls.

High pressure regions are observed inside the grooves and their magnitude depends on the groove shape (Figs. 10a, 10b, 10c). Smaller protrusion lengths lead to pressure reduction and shorten (across the y-direction) high pressure regions. For case lr_1 (Fig. 10a), the length of the high pressure region inside the cavity is about 2% of the protrusion height, while for cases lr_2 (Fig. 10b) and lr_3 (Fig. 10c) it is 25% and 50% respectively. Pressure profile towards the lower flat wall is not affected by the presence of protrusion elements in the upper wall.



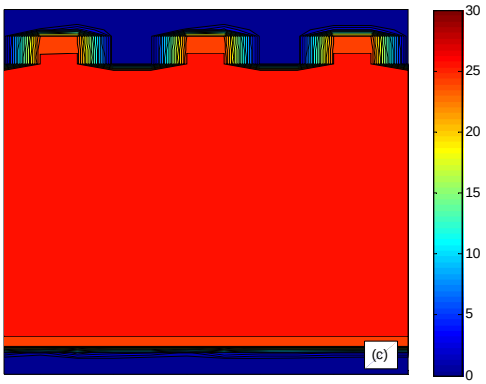


Fig. 10. Isobars obtained for $N_m=2.5x10^6$ and $a_{wf} = a_{ff} = a_{ww} = 7.5$ (parameters kept constant: the cut-off radius $r_c = 1.0\sigma$ and the external driving force $F_{ext} = 0.02$) for cases a) $l_{r1}=0.5l_{tot}$, b) $l_{r2}=0.25l_{tot}$, c) $l_{r3}=0.167l_{tot}$. Dash-dot lines denote solid wall limits and y is in r_c units.

4 CONCLUDING REMARKS

In this paper, we have presented DPD simulations of flow in periodically grooved nano- and microchannels. The rough wall was designed by placing a number of orthogonal protrusions and cavities of equal length in a periodic pattern. Calculations of density, velocity and pressure maps or profiles show clearly that the fluid/wall interaction, the range of interaction of DPD particles, the external driving force and the coarse-graining parameter affect considerably the fluid motion. Number density values in the core of the channel and inside the cavities decreases as the conservative force parameter decreases. The cut-off radius affects density values near the walls and inside the cavities. The value of the external driving force has no significant effect on fluid ordering.

Velocity values in the core of the channel and inside the cavities increases as the conservative force parameter and the external driving force increases and the cut-off radius decreases for all channel cases studied here.

Pressure is strongly affected by the cut-off radius and the conservative force parameter amplitude. For all parameters, as the parameter increases pressure also increases both in the core of the channel and inside the cavities. Inside the cavities high pressure regions are detected and their shape and topology depend on the protrusion length.

Interesting behavior is presented in the microscale channel cases. Fluid particle localization is similar for the nanochannel and the microchannel case, as high density peaks are detected for both cases inside the cavities and near the solid walls. Differences are observed in the core of the channel and in the cavity and protrusion midplanes. In particular, in the nanochannel and the protrusion midplane density is slightly lower than its average value and at cavity is higher, in addition to the microchannel case where at the protrusion midplane it is equal to its average value and at the cavity it is lower. Density peaks inside the cavities reveal trapping of fluid inside the rectangular cavities. This particle trapping affects macroscopic quantities considered here such as velocity and pressure distribution inside and close to the cavities. Velocity reduces systematically and this reduction becomes more pronounced as the protrusion length decreases. For the microchannel, the pressure remains almost constant in the core of the channel and its pattern near and inside the cavities depend on the protrusion shape.

Different materials for the wall and fluid particles, several fluid and wall interface conditions and different fluid types on nano and microchannels with grooves on the upper

wall were investigated. The effect on density, velocity, and pressure distributions were outlined. Although significant progress has been made towards the development and research of grooved channels investigated with different numerical frameworks, the current state of the art does not represent in detail the situation close to the wall in the mesoscale case and remains on the qualitative flow behavior under special circumstances. Future research and development is required to address methods for the near wall region behavior at the mesoscale case and compare their results with other research methods. One has to take into consideration the applied research interest of solid grooved surfaces and their involve in micro/nano electromechanical systems (known as MEMS/NEMS), the cooling of micro and nano devices, the lab-on-chip devices, the drug delivery and the micro and nanofliters. It should be noted that modeling of heat transfer processes requires the implementation of an extension of the DPD method that incorporates the energy transport equation, known as eDPD [43].

REFERENCES

[1] P. Koumoutsakos, Multiscale flow simulations using particles. *Annual Review of Fluid Mechanics,* **37**, 457-487, 2005.

[2] J. Davies, D. Maynes, B.W. Webb, B. Woolford, Laminar flow in a microchannel with superhydrophobic walls exhibiting transverse ribs. *Physics of Fluids*, **18**, 087110-1-11, 2006.

[3] A. Niavarani, N.V. Priezjev, Modelling the combined effect of surface roughness and shear rate on slip flow of simple fluids. *Physical Review E,* **81**, 011606-1-32, 2010.

[4] R. Kamali, A. Kharazmi, Molecular dynamics simulation of surface roughness effects on nanoscale flows. *International Journal of Thermal Sciences,* **50**, 226-232, 2011. doi:10.1016/j.ijthermalsci.2010.05.004

[5] N. Asproulis, D.Drikakis, Surface Roughness Effects in Micro and Nanofluidic Devices. *Journal of Computational and Theoretical Nanoscience*, **7(9)**,1825-1830, 2010.

[6] D. Kasiteropoulou, T. Karakasidis, A. Liakopoulos, A Dissipative Particle Dynamics study of flow in periodically grooved nanochannel. *International Journal for Numerical Methods in Fluid*s, **68(9)**,1156-1172, 2011.

[7] J. Sun, Y.L. He, W.Q. Tao, J.W. Rose, H.S. Wang, Multi-scale study of liquid flow in micro/nanochannels of surface wettability and topology. *Microfluidics and Nanofluidics,* **12**, 991-1008, 2012. DOI 10.1007/s10404-012-0933-7

[8] N.V. Priezjev, Interfacial friction between semiflexible polymers and crystalline surfaces. *The Journal of Chemical Physics*, **136**, 224702-1-10, 2012.

[9] W. Chen, R. Zhang, J. Koplik, Velocity slip on curved surfaces. *Condensed Matter,* **1**, 1-5, 2013. axix.org/pdf/1309.1423v1.

[10] N. Priezjev, Molecular dynamic simulations of oscillatory Couette flows with slip boundary conditions. *Microfluidics and Nanofluidics*, **14**, 225-233, 2013.

[11] F. Sofos, T.E. Karakasidis, A. Liakopoulos, How wall properties control diffusion in grooved nanochannels: a molecular dynamic study. *Heat and Mass Transfer*, **49**, 1081-1088, 2013.

[12] R.A. Sahan, A. Liakopoulos, H. Gunes, Reduced Dynamical Models of Nonisothermal Transitional Grooved-Channel Flow. *Physics of Fluids,* **9**, 551-565, 1997.

[13] A. Liakopoulos, P.H. Blythe, H. Gunes, A reduced dynamical model of convective flows in tall laterally heated cavities. *Proceeding of the Royal Society of London*, **453**, 663-672, 1997.

[14] M. Greiner, R.J. Faulkner, V.T Van, H.M. Tufo, P.F. Fischer, Simulations of Three Dimensional Flow and Augmented Heat Transfer in a Symmetrically Grooved Channel, *Journal of Heat Transfer,* **122**, 653-660, 2000.

[15] Y.M. Chung, P.G. Tucker, D.G. Roychowdhury, Unsteady laminar flow and convective heat transfer in a sharp 180° bend. *International Journal of Heat and Fluid Flow,* **24**, 67-76, 2003.

[16] Y.M. Chung, P.G. Tucker, K.H. Luo, Large-eddy simulation of complex internal flows, *in Direct and Large-Eddy Simulation IV, Eds. B. Geurts, R. Friedrich, O. Metais*, London, 2001.

[17] T. Adachi, S. Hasegawa, Transition of the Flow in a Symmetric Channel with Periodically Expanded Grooves. *Chemical Engineering Science* , **61**, 2721-2729, 2006.

[18] M. Hahn, D. Drikakis, Large eddy simulation of compressible turbulence using high-resolution methods. *International Journal for Numerical Methods in Fluids*, **47**, 971-977, 2005.

[19] B. Thornber, D. Drikakis, Implicit Large Eddy Simulation of a Deep Cavity Using High-Resolution Methods. *AIAA Journal,* **46(10)**, 2634-2685, 2008.

[20] M. Kalweit, D. Drikakis, Coupling strategies for hybrid molecular-continuum simulation methods, *Procceedings of IMechE Collection, Part C: J. Mechanical Engineering Science*, **222**, 797-806, 2008.

[21] N. Asproulis, M. Kalweit, D. Drikakis, A Hybrid Molecular Continuum Method using Point Wise Coupling. *Advances in Engineering Software*, **46, Issue 1**, 85-92, 2012.

[22] E.E. Keaveny, I.V. Pivkin, M. Maxey, G. Karniadakis, A comparative study between dissipative particle dynamics and molecular dynamics for simple- and complex-geometry flows. *The Journal of Chemical Physics*, **123**, 104107-1-104107-9, 2005.

[23] R.D Groot, P.B. Warren, Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulation. *The Journal of Chemical Physics,* **107**, 4423-4435, 1997.

[24] D.A. Fedosov, I.V. Pivkin, G.E. Karniadakis, Velocity limit in DPD simulations of wall-bounded flows. *Journal of Computational Physics,* **227**, 2540-2559, 2008.

[25] P. Espanol, P. Warren, Statistical mechanics of dissipative particle dynamics. *Europhysics Letters,* **30**, 191-196, 1995.

[26] G. Karniadakis, A. Beskok, N. Aluru, *Microflows and Nanoflows: Fundamentals and Simulation*, New York :Springer, 2002.

[27] Y. Trofimov, *Thermodynamic consistency in dissipative particle dynamics*. Technische Universiteit Eindhoven, Ph.D. Thesis, 2003.

[28] P. Nikunen, M. Karttunen, I. Vattulainen, How would you integrate the equations of motion in dissipative particle dynamics simulations?. *Computer Physics Communications,* **, 153**, 407-423, 2003.

[29] I.V. Pivkin, G.E. Karniadakis, A new method to impose no-slip boundary conditions in dissipative particle dynamics. *Journal of Computational Physics,* **207**, 114-128, 2005.

[30] S.J. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics,* **117**, 1-19, 1995.

[31] D. Kasiteropoulou, T.E. Karakasidis, A. Liakopoulos, Dissipative Particle Dynamics investigation of parameters affecting planar nanochannel flows. *Material Science and Engineering B,* **176(9)**, 1574-1579, 2011.

[32] A. Kumar, Y. Asako, E. Abu-Nada, M. Krafczyk, M. Faghri, From dissipative particle dynamics scales to physical scales: a coarse-graining study for water flow in microchannel. *Microfluidics and Nanofluidics*, **7**, 467-477, 2009.

[33] M. Revenga, I. Zuniga, P. Espanol, Boundary conditions in dissipative particle dynamics. *Computer Physics Communications*, **121–122**, 309-311, 1999.

[34] J.H. Irving, J.G. Kirkwood, The statistical mechanics of transport processes. IV. The equation of hydrodynamics. *The Journal of Chemical Physics*, **18**, 817-830, 1950.
[35] X. Fan, N. Phan-Thien, N. Yong, X. Wu, D. Xu, Microchannel flow of a macromolecular suspension. *Physics of Fluids*, **15 (1)**, 11-21, 2003.
[36] F. Sofos, T.E. Karakasidis, A. Liakopoulos, Surface wettability effects on flow in rough wall nanochannels. *Microfluidics and Nanofluidics,* **12**, 25-31, 2012. DOI 10.1007/s10404-011-0845-y
[37] S.C. Young, Effects of surface roughness and interface wettability on nanoscale flow in a nanochannel. *Microfluidics and Nanofluidics,* **2**, 501-511, 2006.
[38] N.V. Priezjev, Effects of surface roughness on rate-dependent slip in simple fluids. *The Journal of Chemical Physics,* **127**, 144708, 2007.
[39] D. Kasiteropoulou, T. Karakasidis, A. Liakopoulos, Mesoscopic simulation of fluid flow in periodically grooved microchannels. *Computers & Fluids,* **74**, 91-101, 2013.
[40] D. Kasiteropoulou, T. Karakasidis, A. Liakopoulos, Study of fluid flow in grooved micro and nano-channels via dissipative particle dynamic: a tool for desalination membrane design. *Desalination and Water Treatment,* 2016. doi: 10.1080/19443994.2016.1141118
[41] I.V. Pivkin, G. Em. Karniadakis, Controlling density fluctuations in wall-bounded dissipative particle dynamics systems. *Physical Review Letters*, **96**, 206001, 2006. DOI:http://dx.doi.org/10.1103/PhysRevLett.96.206001
[42] T. Werder, J.H. Walther, P. Koumoutsakos, Hybrid atomistic–continuum method for the simulation of dense fluid flows. *Journal of Computational Physics*, **205**, 373–390, 2005.
[43] Z. Li, Y.-H. Tang, H. Lei, B. Caswell, G. Karniadakis, Energy-conserving dissipative particle dynamics with temperature-dependent properties. *Journal of Computational Physics,* **265**, 113–127, 2014.
[44] F.D. Sofos, T.E. Karakasidis, A. Liakopoulos, Effects of wall roughness on flow in nanochannels. Physical Review E, **79**, 026305:1-026305:7, 2009. doi: 10.1103/PhysRevE.79.026305.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

HIGH COMPUTATIONAL EFFICIENCY THROUGH GENERIC ANALYTICAL FORMULATION FOR LINEAR SOIL PRESSURE DISTRIBUTION OF RIGID SPREAD RECTANGULAR FOOTINGS

John Bellos[1] **and Nikolaos P. Bakas**[2]

[1] Assistant Professor
School of Architecture, Engineering, Land & Environmental Sciences, Neapolis University Pafos
j.bellos@nup.ac.cy

[2] Lecturer
School of Architecture, Engineering, Land & Environmental Sciences, Neapolis University Pafos
n.bakas@nup.ac.cy

Keywords: Rectangular footing, soil pressure distribution, analytical solution, vertical loading, biaxial bending, eccentricity regions, compression zone, computational efficiency.

Abstract. *The primary objective of this paper is to resolve and provide generic analytical formulas concerning the linear pressure distribution of rigid spread rectangular footings and, consequently, limit computational costs. All five distinct regions of the eccentricity diagram are related to five possible forms of footing deformations and five discrete shapes of compression zone. For each region, the linear soil pressure distribution in soil-footing interface, the neutral axis position, the maximum pressure and the pressure values at the four corners, are expressed in closed forms as functions of biaxial eccentricities, mean soil pressure and footing dimensions. Several special cases are also presented, verifying the correctness and the consistency of the developed analytical formulas and revealing the physical meaning of the eccentricity diagram. The explicit expressions for responses and resultants enable algorithmic implications without iterations, providing high computational efficiency with low computational cost when forming envelopes for shear forces and bending moment or optimizing the design of footing geometry and footing reinforcement, etc. Through developed computer software, a provided simple example of a rigid spread footing under variable eccentric loading, demonstrates how the theoretical content of this article is used to perform numerical calculations. The software itself comprises 3D visualization technology to facilitate visual examination and validation of the results.*

1 INTRODUCTION

It is well known that rigid spread rectangular footings resting on elastic soils tend to deform in a linearly elastic manner under the action of vertical load and biaxial bending [1, 12]. Such loading conditions may occur either due to eccentric vertical loading or high horizontal excitation as those used for structural design in high seismic risk areas. Thus, assuming footing rigidity of adequately high magnitude, it is reasonable to consider a planar settling of footing base and a linear soil pressure distribution due to constant ratio of pressure to settlement.

The footing-soil interaction generates compression underneath the entire or part of the footing base, depending on the area being in contact with the soil. This area, usually called *active area* or *effective area* or *compression zone* [3, 11, 13], should not be less than a certain percentage of the entire base area. In fact, when the effective eccentricities, computed by dividing bending moments with vertical force, lay outside a certain geometric ellipse called *secondary core*, the active area is smaller than half the base area. As strictly as required by regulations and usually recommended by good practice, such cases are not acceptable in footing design.

The most important issue concerning rectangular footings is optimal geometric and reinforcement design [2, 3, 6, 7, 14]. Initially, a set of suitable values is chosen so that, under serviceability conditions, the maximum pressure exerted at any footing position does not exceed the soil bearing capacity. Moreover, constraints set for enveloped bending, shearing and punching should not be violated under ultimate loading conditions, while requirements imposed by regulations should be fully met. All the above involve an immense amount of intensive calculations through complex iterative processes. Several procedures were developed in the past for handling efficiently such footing-soil interaction problems, by using either analytical [2, 4, 5, 6, 7, 13, 14], numerical [10, 11], graphical [3, 15], or hybrid techniques [8, 9]. Although accurate and useful, some of them are rather computationally expensive, while others are not so wide-ranging or not easily implementable in designing footings of such specific shape, yet participating in a complex structure assembly subject to various loading actions. On the other hand, the generic explicit formulas presented in this article increase the efficiency of the required design processes by reducing considerably the computational costs.

Another advantage of the presented analytical method is that, instead of relating eccentricities to external loading conditions, they are evaluated by using reaction resultants at the respective support node as derived from the space frame model of the structure. Hence, these effective eccentricities express a clear relation of biaxial bending moments to vertical load for the specific footing, since they are not directly associated with external loading and influences from neighboring columns, footings and connecting beams.

This paper attempts to enlighten the issues concerning the rigid spread rectangular footings by providing easily programmable analytical formulas and, through a developed intelligent software and 3D visualization technology, facilitate results inspection, confirm engineering intuition, unshed light to foundation settlement and encourage future research in related fields.

2 THE MATHEMATICAL MODEL

As we stated previously, a rigid spread rectangular footing is expected to deform in a planar manner. Therefore, it is reasonable to consider a linear soil pressure distribution underneath the base, that is

$$p(x, y) = p_{\max} \cdot \left(-\frac{x}{x_n} - \frac{y}{y_n} + 1 \right) \tag{1}$$

where p_{max} is the maximum soil pressure on the base and x_n, y_n are the intercepts of the neutral axis determining the active area of the footing (see Figure 1).

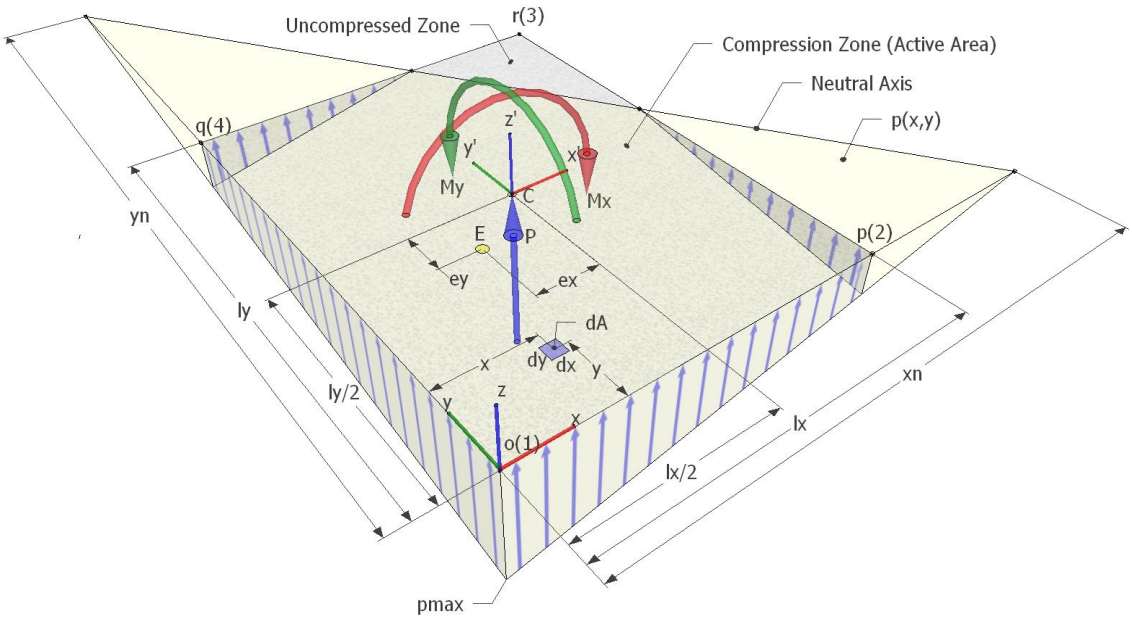


Figure 1: Linear soil pressure distribution on a rigid spread rectangular footing.

The effective eccentricities from the footing center, derived from certain reaction resultants at the central support node, i.e. the calculated values of biaxial moments M_x, M_y and vertical force P (see Figure 1), are expressed as

$$e_x = \frac{M_x}{P}, \qquad e_y = \frac{M_y}{P} \tag{2}$$

Equilibrium of actions and reactions at the base center (x_c, y_c) requires

$$P = \iint_A p(x,y) \cdot dA \tag{3}$$

$$M_x = \iint_A (x_c - x) \cdot p(x,y) \cdot dA \tag{4}$$

$$M_y = \iint_A (y_c - y) \cdot p(x,y) \cdot dA \tag{5}$$

Substituting equation (3) into equations (4) and (5), they are transformed into

$$M_x = P \cdot x_c - \iint_A x \cdot p(x,y) \cdot dA \Rightarrow P \cdot (x_c - e_x) = \iint_A x \cdot p(x,y) \cdot dA \tag{6}$$

$$M_y = P \cdot y_c - \iint_A y \cdot p(x,y) \cdot dA \Rightarrow P \cdot (y_c - e_y) = \iint_A y \cdot p(x,y) \cdot dA \tag{7}$$

The mean soil pressure on a footing with dimensions l_x, l_y is defined as

$$p_m = \frac{P}{l_x \cdot l_y} \tag{8}$$

If, by convention, one uses absolute values for eccentricities and thus presets the maximum soil pressure at the lower left footing corner, i.e. $p_o = p_{max}$ (see Figure 1). Then equations (3), (6) and (7) can be written as

$$p_m \cdot l_x \cdot l_y = \iint_A p(x, y) \cdot dx \cdot dy \tag{9}$$

$$p_m \cdot l_x \cdot l_y \cdot \left(\frac{l_x}{2} - |e_x| \right) = \iint_A x \cdot p(x, y) \cdot dx \cdot dy \Rightarrow p_m \cdot l_x^2 \cdot l_y \cdot \left(\frac{1}{2} - \frac{|e_x|}{l_x} \right) = \iint_A x \cdot p(x, y) \cdot dx \cdot dy \tag{10}$$

$$p_m \cdot l_x \cdot l_y \cdot \left(\frac{l_y}{2} - |e_y| \right) = \iint_A y \cdot p(x, y) \cdot dx \cdot dy \Rightarrow p_m \cdot l_x \cdot l_y^2 \cdot \left(\frac{1}{2} - \frac{|e_y|}{l_y} \right) = \iint_A y \cdot p(x, y) \cdot dx \cdot dy \tag{11}$$

There is a direct relation of pressure values at actual footing corners with those computed from at rearranged footing corners, with the order strictly depending on the eccentricities signs, i.e.

$$\begin{aligned}
&e_x \geq 0 \,\&\, e_y \geq 0 \Rightarrow p_1 = p_o = p_{\max},\ p_2 = p_p,\ p_3 = p_r,\ p_4 = p_q \\
&e_x < 0 \,\&\, e_y > 0 \Rightarrow p_1 = p_p,\ p_2 = p_o = p_{\max},\ p_3 = p_q,\ p_4 = p_r \\
&e_x \leq 0 \,\&\, e_y < 0 \Rightarrow p_1 = p_r,\ p_2 = p_q,\ p_3 = p_o = p_{\max},\ p_4 = p_p \\
&e_x > 0 \,\&\, e_y < 0 \Rightarrow p_1 = p_q,\ p_2 = p_r,\ p_3 = p_p,\ p_4 = p_o = p_{\max}
\end{aligned} \tag{12}$$

Solving the system of equations (9), (10) and (11) and determining the unknown parameters x_n, y_n and p_{max}, then the equation (1) depends on the signs of effective eccentricities and, therefore, it can be rewritten as

$$p(x, y) = p_{\max} \cdot \left[-sign(e_x) \cdot \frac{x}{x_n} - sign(e_y) \cdot \frac{y}{y_n} + 1 \right] \tag{13}$$

and by relocating the origin of the coordinate system at the footing centroid, the linear soil pressure distribution can be written in an equivalent to equation (1) expression, that is

$$p(x', y') = p_{\max} \cdot \left[-sign(e_x) \cdot \frac{x'}{x_n} - sign(e_y) \cdot \frac{y'}{y_n} + 1 - \frac{1}{2} \left(\frac{l_x}{x_n} + \frac{l_y}{y_n} \right) \right] \tag{14}$$

3 THE ECCENTRICITY DIAGRAM

As effective eccentricities, evaluated using equations (2), sweep the rectangular footing base certain types of deformation occur, thus generating an *eccentricity diagram* for each footing, as presented in Figure 2. This diagram, also obtained by other authors using different approaches [13], constitutes a very strong practical tool for determining the form of pressure distribution expected under certain loading conditions. In this work, the border lines on the eccentricity diagram are provided explicitly while special cases reveal their physical meaning.

There are five distinct regions in this diagram, namely A-E, corresponding to different types of deformation and active areas of unique geometric shapes (see Figure 2). Each regional case is subdivided into four subcases, each associated with one primary footing corner of maximum soil pressure, as determined by eccentricities signs in equations (12). Notice that there is a rhombus domain, usually called *main core*, having semi-diagonals equal to 1/6 of the respective footing lengths and a surrounding elliptic domain, usually called *secondary*

core, bounded by an ellipse with major and minor semi-axes equal to 1/3 of the respective footing lengths (dashed line in Figure 2). The first contains the entire regions C, while the second comprises regions B and certain parts of regions D and E. The remaining parts of regions D and E and the entire regions A are excluded, since corresponded to active areas smaller than half the footing base area. Regulations as well as good practice require that effective eccentricities should remain inside the secondary core in order to maintain an acceptable footing design.

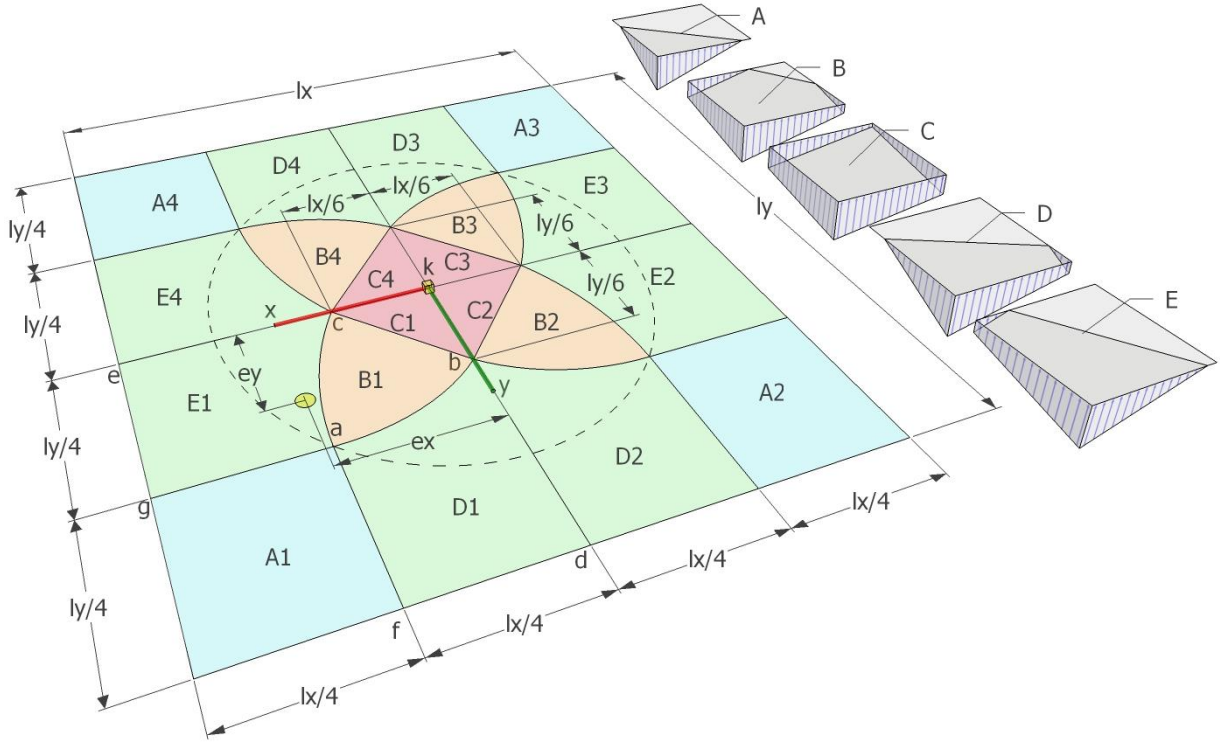


Figure 2: The eccentricity diagram of a rigid spread rectangular footing.

4 THE ECCENTRICITY REGIONS

The analytical formulas for all five regions are presented in the subsections below. They are produced by integrating the equations (9), (10) and (11) over the active area of the footing base corresponding to each regional case. Nevertheless, for saving space and time, the detailed development process of the formulas is not provided here.

4.1 Region C: Full Compression Zone

Regions C corresponds to four equal right triangles composing a rhombus shape with semi-diagonals equal to 1/6 of the respective footing lengths (see Figure 3). This rhombus area, usually called main core, represents small eccentricities generated by vertical forces applied at points in it and capable of setting the entire footing base under compression. Maximum pressure occurs at the footing corner corresponding to the right triangle enclosing the eccentricity point.

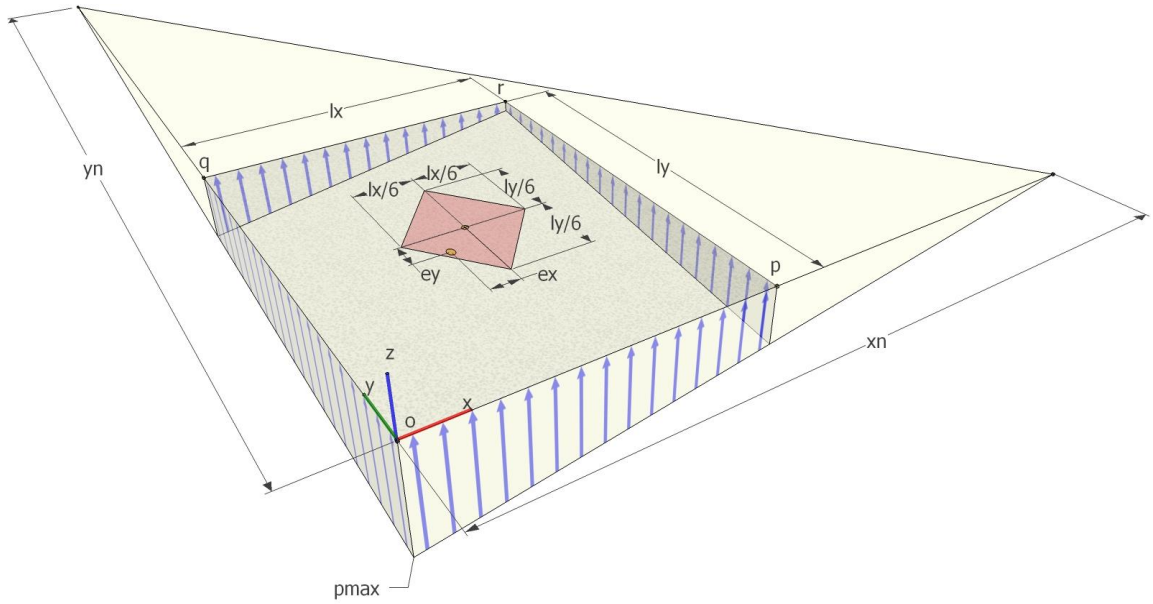


Figure 3: Full compression zone.

The intercepts of the neutral axis are given by

$$x_n = l_x \cdot \left(1+6\cdot\frac{|e_x|}{l_x}+6\cdot\frac{|e_y|}{l_y}\right)\Bigg/\left(12\cdot\frac{|e_x|}{l_x}\right), \qquad y_n = l_y \cdot \left(1+6\cdot\frac{|e_x|}{l_x}+6\cdot\frac{|e_y|}{l_y}\right)\Bigg/\left(12\cdot\frac{|e_y|}{l_y}\right) \quad (15)$$

The maximum soil pressure at the primary corner of the footing base (coordinate system origin O) along with the soil pressure at the rest corners, are provided by the equations

$$\begin{aligned} p_o &= p_{\max} = p_m \cdot \left(1+6\cdot\frac{|e_x|}{l_x}+6\cdot\frac{|e_y|}{l_y}\right), & p_p &= p_m \cdot \left(1-6\cdot\frac{|e_x|}{l_x}+6\cdot\frac{|e_y|}{l_y}\right), \\ p_q &= p_m \cdot \left(1+6\cdot\frac{|e_x|}{l_x}-6\cdot\frac{|e_y|}{l_y}\right), & p_r &= p_{\min} = p_m \cdot \left(1-6\cdot\frac{|e_x|}{l_x}-6\cdot\frac{|e_y|}{l_y}\right) \end{aligned} \quad (16)$$

In this particular case, since the active area is the entire footing base, all footing corners should be under compression, i.e. $p_o \geq 0$ & $p_p \geq 0$ & $p_q \geq 0$ & $p_r \geq 0$. Since p_r represents the minimum compression value, the last of equations (16) yields the rhombus shape boundary for regions C shown in Figure 3, i.e.

$$\frac{|e_x|}{l_x}+\frac{|e_y|}{l_y} \leq \frac{1}{6} \quad (17)$$

4.2 Region A: Triangular Compression Zone

Regions A corresponds to four smaller corner rectangles, each having sides equal to 1/4 of the respective footing lengths (see Figure 4). They represent excessive eccentricities generated by vertical forces applied at points inside these corner rectangles, imposing high pressure to respective triangular compression zones of the footing base. Maximum pressure occurs at the footing corner associated with the corner rectangle enclosing the eccentricity point.

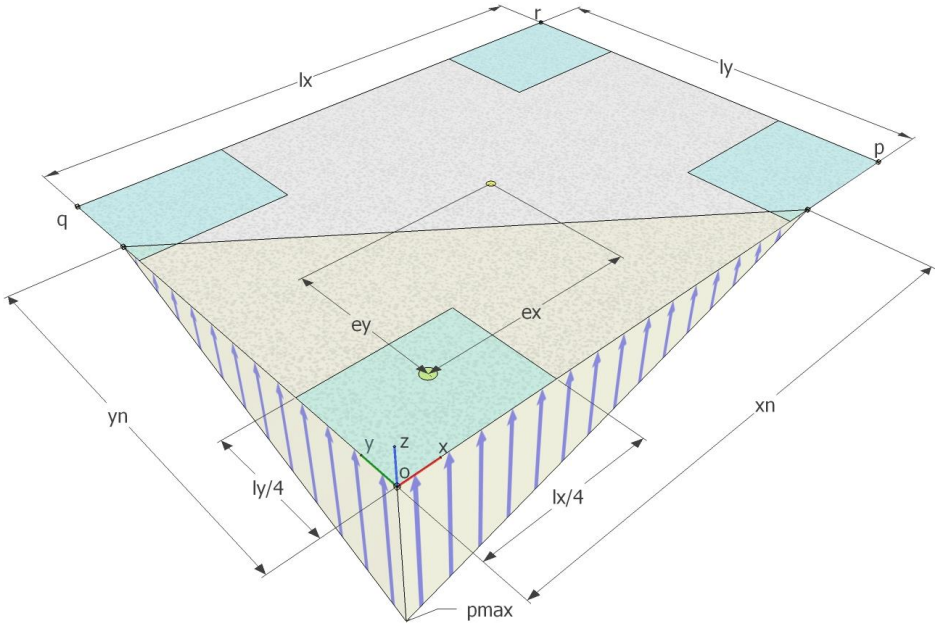


Figure 4: Triangular compression zone.

The neutral axis intercepts are given by

$$x_n = 4 \cdot l_x \cdot \left(\frac{1}{2} - \frac{|e_x|}{l_x} \right), \qquad y_n = 4 \cdot l_y \cdot \left(\frac{1}{2} - \frac{|e_y|}{l_y} \right) \tag{18}$$

The soil pressure values at the footing corners are provided by

$$p_o = p_{\max} = \frac{3 \cdot p_m}{8 \cdot \left(\frac{1}{2} - \frac{|e_x|}{l_x} \right) \cdot \left(\frac{1}{2} - \frac{|e_y|}{l_y} \right)}, \qquad p_p = p_q = p_r = p_{\min} = 0 \tag{19}$$

For triangular compression zone the intercepts should be confined within the neighboring sides of the footing base. Hence, equations (18) yield the rectangular shape boundary for region A (see Figure 4), expressed as the lower and upper bounds in the two inequalities

$$\begin{aligned} 0 < x_n \le l_x \Rightarrow \qquad & \frac{1}{4} \le \frac{|e_x|}{l_x} < \frac{1}{2} \\ 0 < y_n \le l_y \Rightarrow \qquad & \frac{1}{4} \le \frac{|e_y|}{l_y} < \frac{1}{2} \end{aligned} \tag{20}$$

where the upper bounds denote that eccentricities outside the footing base are unbalanced.

4.3 Region D: Trapezoidal Compression Zone in x-Direction

Regions D corresponds to four curved trapezoids attached to the x-direction footing sides in pairs, each having bases equal to 1/3 and 1/4 of the y-direction footing length respectively, altitude 1/4 of the x-direction footing length and a curved side (see Figure 5). They represent large eccentricities generated by vertical forces applied at points inside these curved trapezoids, imposing pressure to respective trapezoidal zones covering the entire x-direction foot-

ing side. Maximum pressure occurs at the footing corner corresponding to the curved trapezoid enclosing the eccentricity point.

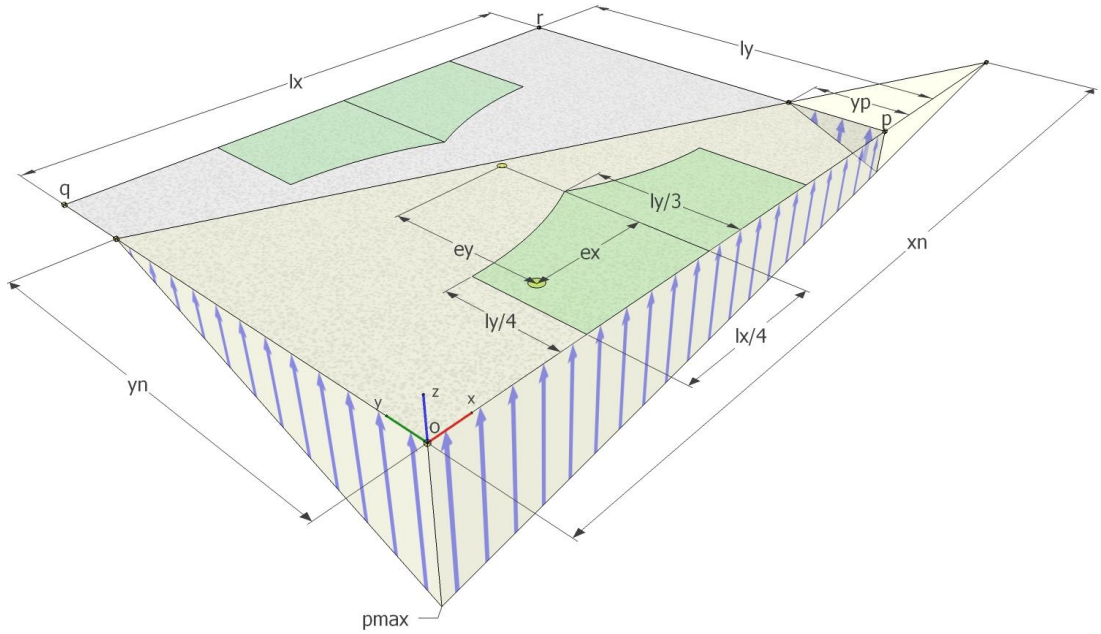


Figure 5: Trapezoidal compression zone covering the entire side in x-direction.

The neutral axis position, determined by its intercepts and its trace y_p on the parallel side (see Figure 5), is provided by

$$x_n = l_x \cdot \left(1+6\cdot\frac{|e_x|}{l_x}+\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\Bigg/\left(12\cdot\frac{|e_x|}{l_x}\right),$$

$$y_n = \frac{1}{2}\cdot l_y \cdot \left(\frac{1}{2}-\frac{|e_y|}{l_y}\right)\cdot\left(2+\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\cdot\left(1+2\cdot\frac{|e_x|}{l_x}-\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\Bigg/\left[\frac{|e_x|}{l_x}\cdot\left(1+4\cdot\frac{e_x^2}{l_x^2}\right)\right], \quad (21)$$

$$y_p = y_n\cdot\left(1-\frac{l_x}{x_n}\right)$$

The soil pressure values at the footing corners are

$$p_o = p_{\max} = \frac{p_m}{3}\cdot\left(2-\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\cdot\left(1+6\cdot\frac{|e_x|}{l_x}+\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\Bigg/\left(\frac{1}{2}-\frac{|e_y|}{l_y}\right),$$

$$p_p = p_{\max}\cdot\left(1-\frac{l_x}{x_n}\right), \qquad p_q = p_r = p_{\min} = 0 \quad (22)$$

Requiring the x-intercept to lie on the extension of the respective footing side, the bases of the trapezoidal region are given as the lower and upper bounds of an inequality, i.e.

$$1 \le \frac{x_n}{l_x} \Rightarrow \qquad 0 \le \frac{|e_x|}{l_x} \le \frac{1}{4} \quad (23)$$

In fact, the upper bound of the inequality guaranties that the discriminant, appearing under the square roots of equations (21) and (22), is always positive definite.

Confining the y-intercept within the respective footing side, the lateral sides of the trapezoidal region are given as the lower and upper bounds of another inequality (see Figure 5)

$$0<\frac{y_n}{l_y}\le 1\Rightarrow \qquad \frac{1}{2}-\frac{1}{6}\cdot\left(2-\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\cdot\left(1+2\cdot\frac{|e_x|}{l_x}+\sqrt{1-12\cdot\frac{e_x^2}{l_x^2}}\right)\Bigg/\left(1+4\cdot\frac{|e_x|}{l_x}\right)\le\frac{|e_y|}{l_y}<\frac{1}{2} \quad (24)$$

Here, the lower bound of the inequality expresses the curved side of trapezoid while the upper bound states that eccentricities outside the footing base are unbalanced.

4.4 Region E: Trapezoidal Compression Zone in y-Direction

Regions E corresponds to four curved trapezoids attached to the y-direction footing sides in pairs, each having bases equal to 1/3 and 1/4 of the y-direction footing length respectively, altitude 1/4 of the y-direction footing length and a curved side (see Figure 6). They represent large eccentricities generated by vertical forces applied at points inside these curved trapezoids, imposing pressure to respective trapezoidal compression zones covering the entire y-direction footing side. Maximum pressure occurs at the footing corner corresponding to the curved trapezoid enclosing the eccentricity point.

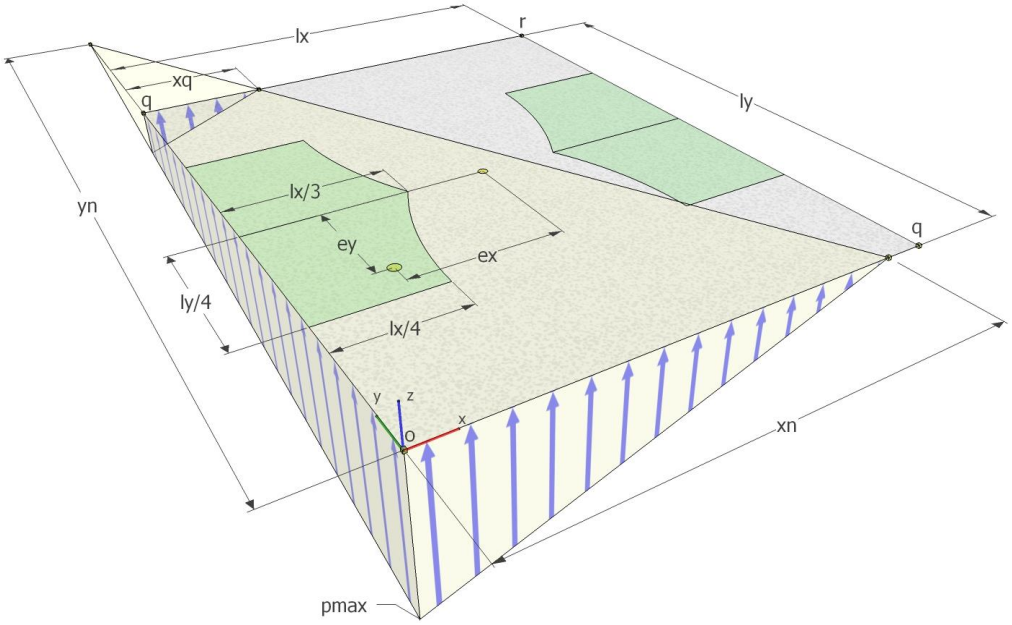


Figure 6: Trapezoidal compression zone covering entire side in y-direction.

The position of the neutral axis is provided by

$$x_n=\frac{1}{2}\cdot l_x\cdot\left(\frac{1}{2}-\frac{|e_x|}{l_y}\right)\cdot\left(2+\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\cdot\left(1+2\cdot\frac{|e_y|}{l_y}-\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\Bigg/\left[\frac{|e_y|}{l_y}\cdot\left(1+4\cdot\frac{e_y^2}{l_y^2}\right)\right],$$

$$y_n=l_y\cdot\left(1+6\cdot\frac{|e_y|}{l_y}+\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\Bigg/\left(12\cdot\frac{|e_y|}{l_y}\right),\qquad x_q=x_n\cdot\left(1-\frac{l_y}{y_n}\right) \quad (25)$$

The soil pressure values at the footing corners are

$$p_o = p_{\max} = \frac{p_m}{3}\cdot\left(2-\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\cdot\left(1+6\cdot\frac{|e_y|}{l_y}+\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\Bigg/\left(\frac{1}{2}-\frac{|e_x|}{l_x}\right),$$
$$p_q = p_{\max}\cdot\left(1-\frac{l_y}{y_n}\right), \qquad p_p = p_r = p_{\min} = 0 \tag{26}$$

Confining properly the intercepts of neutral axis with respect to footing sides in equations (25), the curved trapezoid is determined by the following two inequalities

$$0<\frac{x_n}{l_x}\le 1 \Rightarrow \qquad \frac{1}{2}-\frac{1}{6}\cdot\left(2-\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\cdot\left(1+2\cdot\frac{|e_y|}{l_y}+\sqrt{1-12\cdot\frac{e_y^2}{l_y^2}}\right)\Bigg/\left(1+4\cdot\frac{|e_y|}{l_y}\right)\le\frac{|e_x|}{l_x}<\frac{1}{2}$$
$$1\le\frac{y_n}{l_y}\Rightarrow \qquad 0\le\frac{|e_y|}{l_y}\le\frac{1}{4} \tag{27}$$

The bounds of the first inequality express the lateral sides of the curved trapezoid while the bounds of the second its bases.

4.5 Region B: Pentagonal Compression Zone

Regions B corresponds to four diagonally placed curved triangles, each having two curved sides and one flat base, as shown in Figure 7. These curved triangles form part of the secondary core, inscribed in its elliptic boundary and surrounding the main core (see Figure 2). They represent moderate eccentricities generated by vertical forces applied at points inside them, imposing pressure to respective pentagonal compression zones while preserving a triangular uncompressed zone. Maximum pressure occurs at the footing corner corresponding to the curved triangle enclosing the eccentricity point.

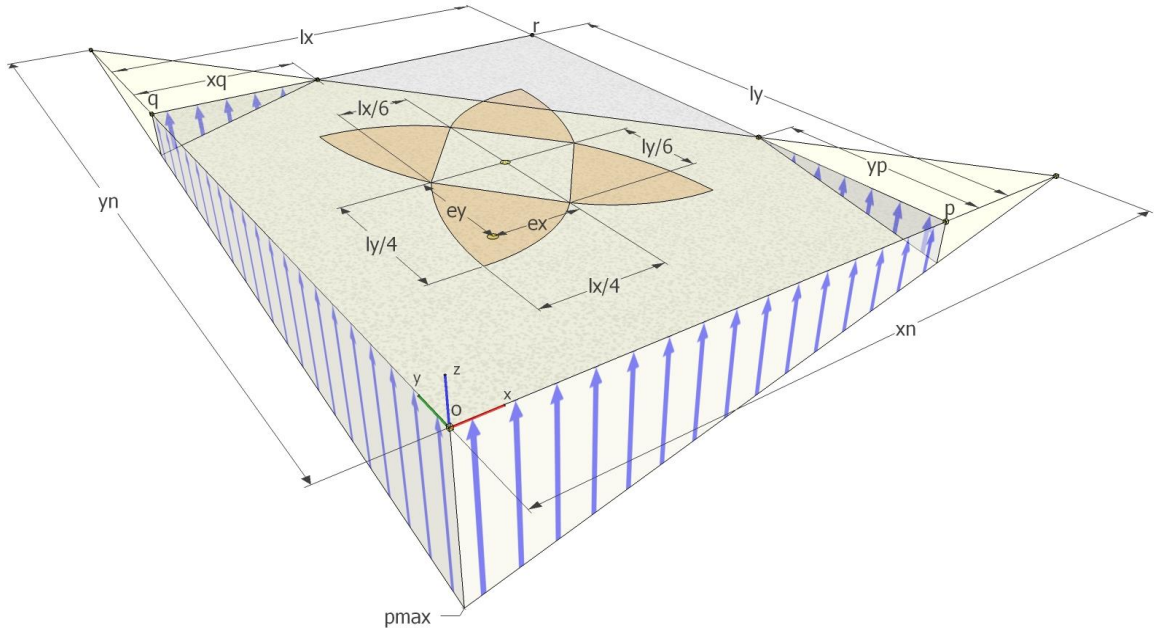


Figure 7: Pentagonal compression zone (triangular uncompressed zone).

As shown below, in full agreement with other sources [13], in this case the solution cannot be expressed in closed form. In fact, the intercepts of the neutral axis are provided by recursive formulas, i.e.

$$x_n = 4 \cdot l_x \cdot \left(\frac{1}{2} - \frac{|e_x|}{l_x} \right) \cdot \frac{1 - \left(1 - \frac{l_x}{x_n}\right)^3 - \left(1 - \frac{l_y}{y_n}\right)^3}{1 - \left(1 - \frac{l_x}{x_n}\right)^4 - \left(1 - \frac{l_y}{y_n}\right)^4 - 4 \cdot \left(1 - \frac{l_x}{x_n}\right)^3 \cdot \frac{l_x}{x_n}},$$
$$y_n = 4 \cdot l_y \cdot \left(\frac{1}{2} - \frac{|e_y|}{l_y} \right) \cdot \frac{1 - \left(1 - \frac{l_x}{x_n}\right)^3 - \left(1 - \frac{l_y}{y_n}\right)^3}{1 - \left(1 - \frac{l_x}{x_n}\right)^4 - \left(1 - \frac{l_y}{y_n}\right)^4 - 4 \cdot \left(1 - \frac{l_y}{y_n}\right)^3 \cdot \frac{l_y}{y_n}} \tag{28}$$

One can easily prove, either numerically or algebraically, that the set of equations (27) constitutes a fast converging iterative procedure. Consequently, once the intercept values found they are substituted to all consequent equations.

The traces of the neutral axis on the two opposite footing sides, as shown in Figure 7, are

$$y_p = y_n \cdot \left(1 - \frac{l_x}{x_n}\right), \qquad x_q = x_n \cdot \left(1 - \frac{l_y}{y_n}\right) \tag{29}$$

The soil pressure values at the footing corners are

$$p_o = p_{\max} = \frac{6 \cdot p_m \cdot \frac{l_x}{x_n} \cdot \frac{l_y}{y_n}}{1 - \left(1 - \frac{l_x}{x_n}\right)^3 - \left(1 - \frac{l_y}{y_n}\right)^3},$$
$$p_p = p_{\max} \cdot \left(1 - \frac{l_x}{x_n}\right), \qquad p_q = p_{\max} \cdot \left(1 - \frac{l_y}{y_n}\right), \qquad p_r = p_{\min} = 0 \tag{30}$$

Since, on its limits, region B degenerates to regions C, D and E respectively, the boundary of the curved triangle is clearly determined by the set of three inequalities

$$\frac{|e_x|}{l_x} + \frac{|e_y|}{l_y} \geq \frac{1}{6}$$
$$\frac{|e_x|}{l_x} \leq \frac{1}{2} - \frac{1}{6} \cdot \left(2 - \sqrt{1 - 12 \cdot \frac{e_y^2}{l_y^2}}\right) \cdot \left(1 + 2 \cdot \frac{|e_y|}{l_y} + \sqrt{1 - 12 \cdot \frac{e_y^2}{l_y^2}}\right) \Bigg/ \left(1 + 4 \cdot \frac{|e_y|}{l_y}\right) \tag{31}$$
$$\frac{|e_y|}{l_y} \leq \frac{1}{2} - \frac{1}{6} \cdot \left(2 - \sqrt{1 - 12 \cdot \frac{e_x^2}{l_x^2}}\right) \cdot \left(1 + 2 \cdot \frac{|e_x|}{l_x} + \sqrt{1 - 12 \cdot \frac{e_x^2}{l_x^2}}\right) \Bigg/ \left(1 + 4 \cdot \frac{|e_x|}{l_x}\right)$$

The first inequality represents eccentricities outside the main core of rhombus region C, while the rest represent eccentricities inside the secondary zone bound by the curved sides of the trapezoidal regions D and E respectively (see Figure 2), as derived in the respective chapters.

5 SPECIAL CASES

Several borderline cases are represented here, verifying the correctness and the consistency of the developed analytical formulas, reassuring the regional continuity and revealing the physical meaning of the eccentricity diagram.

5.1 One-way Eccentricity in Region C

Imposing x-direction eccentricity in Region C by setting $e_y = 0$, as shown in Figure 8, equations (15) and (16) yield the well-known formulas for soil pressure distribution of single slope

$$x_n = l_x \cdot \left(1+6\cdot\frac{|e_x|}{l_x}\right) \Big/ \left(12\cdot\frac{|e_x|}{l_x}\right), \qquad y_n = \infty \tag{32}$$

$$p_o = p_q = p_{\max} = p_m \cdot \left(1+6\cdot\frac{|e_x|}{l_x}\right), \qquad p_p = p_r = p_{\min} = p_m \cdot \left(1-6\cdot\frac{|e_x|}{l_x}\right) \tag{33}$$

while inequality (17) diminishes to representing segment ‘kc’, that is

$$\frac{|e_x|}{l_x} < \frac{1}{6} \tag{34}$$

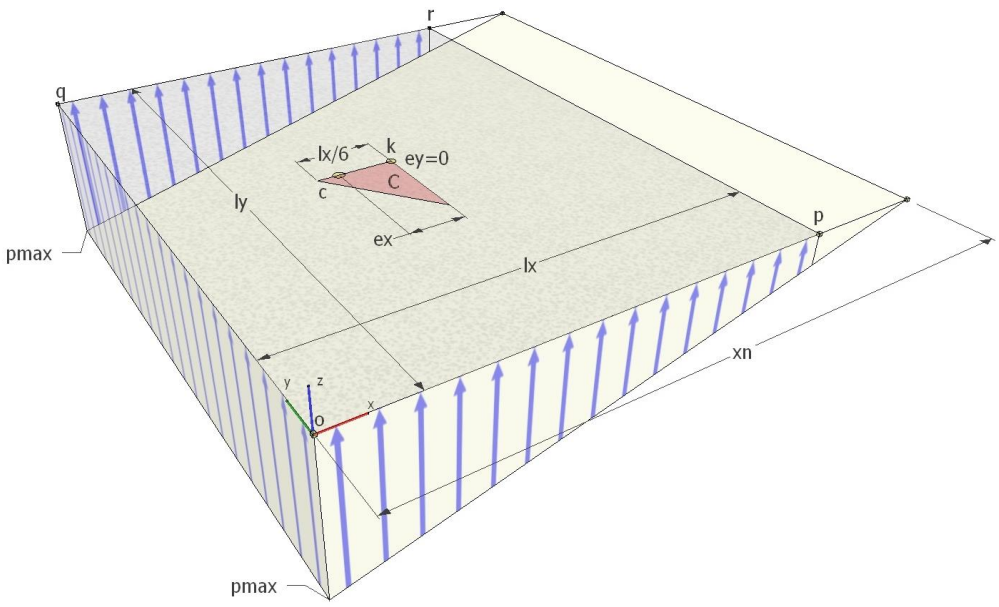


Figure 8: Soil pressure distribution for one-way eccentricity in segment 'kc'.

5.2 One-way Eccentricity in Region C

Imposing x-direction eccentricity in Region E by setting $e_y = 0$, as shown in Figure 9, equations (25) and (26) yield the well-known formulas for soil pressure distribution of single slope and limited active area

$$x_n = 3\cdot l_x \cdot \left(\frac{1}{2}-\frac{|e_x|}{l_x}\right), \qquad y_n = \infty \tag{35}$$

$$p_o = p_q = p_{\max} = \frac{2}{3} \cdot p_m \Big/ \left(\frac{1}{2} - \frac{|e_x|}{l_x} \right), \qquad p_p = p_r = p_{\min} = 0 \tag{36}$$

while inequality (27) diminishes to representing segment ‘ce’, that is

$$\frac{1}{6} \le \frac{|e_x|}{l_x} < \frac{1}{2} \tag{37}$$

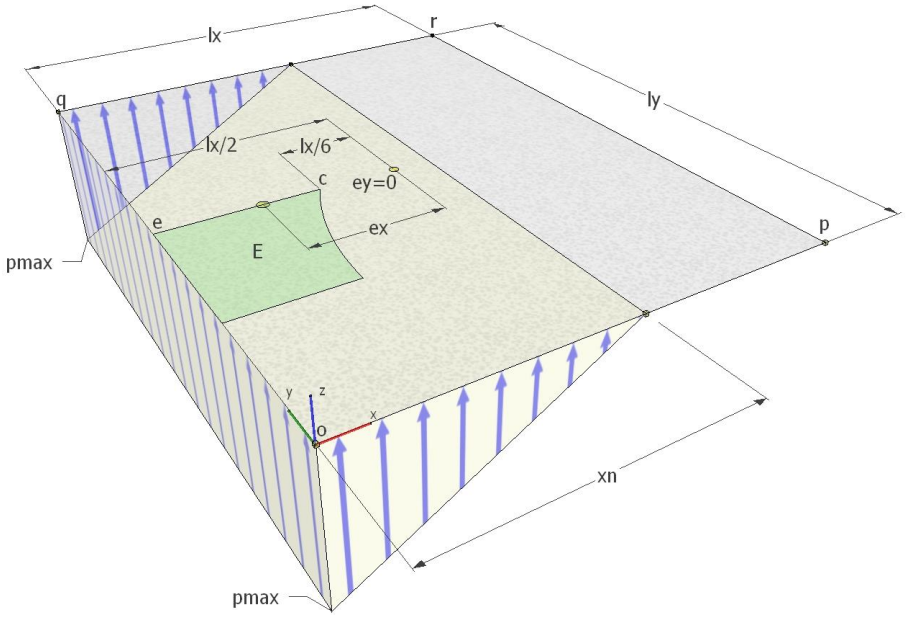


Figure 9: Soil pressure distribution for one-way eccentricity in segment 'ce'.

5.3 One-way Eccentricity in Junction of Regions C, E and B

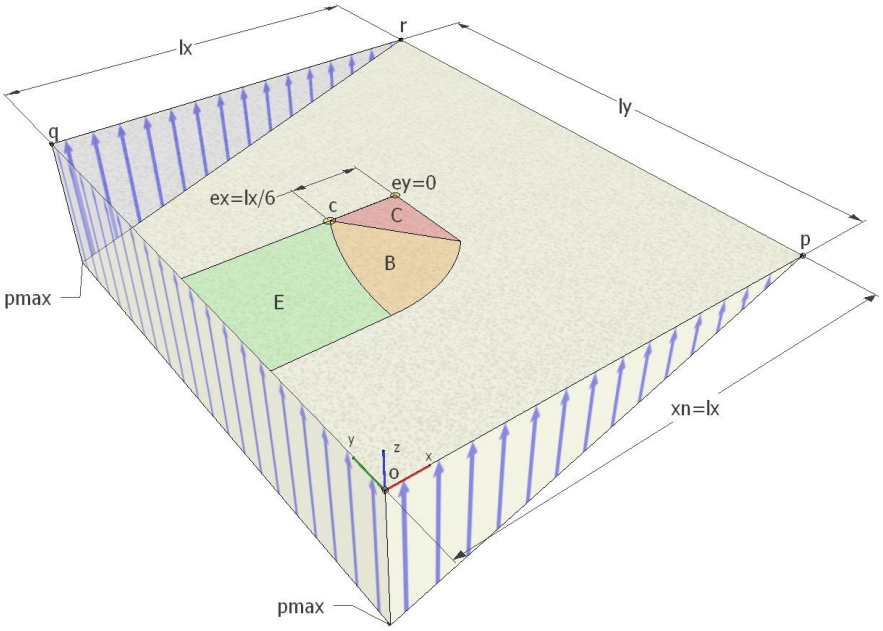


Figure 10: Soil pressure distribution for one-way eccentricity in point 'c'.

Applying x-direction eccentricity in point ‘c’, as shown in Figure 10, by setting

$$\frac{|e_x|}{l_x} = \frac{1}{6}, \qquad e_y = 0 \tag{38}$$

the equations (15) and (16) of region C, (25) and (26) of region E, and (28) and (30) of region B, yield identical results

$$x_n = l_x, \qquad y_n = \infty \tag{39}$$

$$p_o = p_q = p_{\max} = 2 \cdot p_m, \qquad p_p = p_r = p_{\min} = 0 \tag{40}$$

In such a case, the soil pressure distribution of single slope has its neutral axis coincide with the footing base side in y-direction, as illustrated in Figure 10.

5.4 Biaxial Eccentricity in Junction of Regions A, D, E and B

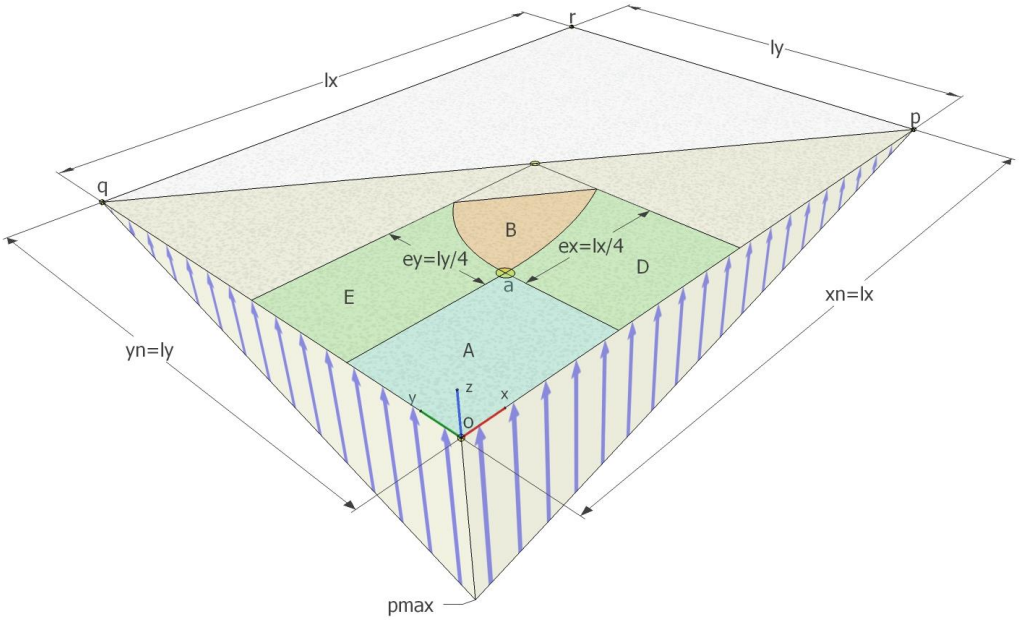


Figure 11: Soil pressure distribution for eccentricity in point 'a'.

Applying biaxial eccentricity in point ‘a’, as shown in Figure 11, by setting

$$\frac{|e_x|}{l_x} = \frac{1}{4}, \qquad \frac{|e_y|}{l_y} = \frac{1}{4} \tag{41}$$

the equations (18) and (19) of region A, (21) and (22) of region D, (25) and (26) of region E, and (28) and (30) of region B, yield identical results

$$x_n = l_x, \qquad y_n = l_y \tag{42}$$

$$p_o = p_{\max} = 6 \cdot p_m, \qquad p_p = p_q = p_r = p_{\min} = 0 \tag{43}$$

In such a case, the soil pressure distribution has its neutral axis coincide with the respective diagonal of the orthogonal base, as illustrated in Figure 11.

6 APPLICATION AND DISCUSSION

A simple example is provided here to demonstrate how the generic analytical formulas of this article are used to perform numerical calculations via developed computer software, thus offering high computational efficiency by limiting computational costs. The software itself comprises a 3D visualization module, facilitating visual examination and validation of the results. Note that all figures and tables of this example have been produced through this software and can be reproduced by rerunning it, practically offering an infinite number of variations to the researcher/engineer.

The structural model concerns a 0.25mx0.25m square column of height 3.00m fixed to a 1.00mx1.00m rigid spread square footing of depth 0.50m. The material used is concrete C30/37 reinforced with steel B500C, with safety factors 1.50 and 1.15 for concrete and steel respectively, and applied with covering of 0.03m for the column and 0.065 for the footing.

The central footing studied for variable eccentric loading required by building codes. Eurocode 1 (EN1991, EC1), Eurocode 2 (EN1992, EC2) and Eurocode 8 (EN1998, EC8) applied in our case, complemented by Greek National Annexes [16]. Accordingly, 33 combinations of actions required in total, as presented in Table 1, where G stands for dead loads, Q for imposed loads and E for seismic loads to primary directions with accidental eccentricities to perpendicular directions.

$A = 1.35G + 1.50Q,$	$I = 1, 2, 3, 4$
${}_iB = 1.00G + 0.30Q + 1.00E_{x\pm eccy} + 0.30E_{y\pm eccx},$	${}_iC = 1.00G + 0.30Q + 1.00E_{x\pm eccy} - 0.30E_{y\pm eccx},$
${}_iD = 1.00G + 0.30Q + 0.30E_{x\pm eccy} + 1.00E_{y\pm eccx},$	${}_iE = 1.00G + 0.30Q - 0.30\,E_{x\pm eccy} + 1.00E_{y\pm eccx},$
${}_iF = 1.00G + 0.30Q - 1.00E_{x\pm eccy} - 0.30E_{y\pm eccx},$	${}_iG = 1.00G + 0.30Q - 1.00E_{x\pm eccy} + 0.30E_{y\pm eccx},$
${}_iH = 1.00G + 0.30Q - 0.30E_{x\pm eccy} - 1.00E_{y\pm eccx},$	${}_iI = 1.00G + 0.30Q + 0.30\,E_{x\pm eccy} - 1.00\,E_{y\pm eccx},$

Table 1: The 33 combinations of actions required.

Comb.	$\mathbf{e_x}$ (m)	$\mathbf{e_y}$ (m)	$\mathbf{p_m}$ (kPa)	$\mathbf{p_1}$ (kPa)	$\mathbf{p_2}$ (kPa)	$\mathbf{p_3}$ (kPa)	$\mathbf{p_4}$ (kPa)	$\mathbf{V_{xf}}$ (kN)	$\mathbf{V_{yf}}$ (kN)	$\mathbf{M_{xf}}$ (kNm)	$\mathbf{M_{yf}}$ (kNm)
A	0.000	0.000	23.2	23.2	23.2	23.2	23.2	8.7	8.7	1.6	1.6
${}_{1,2,3,4}B$	-0.076	-0.003	17.2	7.2	22.5	27.1	11.8	9.1	9.7	1.8	1.8
${}_{1,2,3,4}C$	-0.076	0.003	17.2	11.8	27.1	22.5	7.2	9.1	9.7	1.8	1.8
${}_{1,2,3,4}D$	-0.003	-0.076	17.2	7.2	11.8	27.1	22.5	9.7	9.1	1.8	1.8
${}_{1,2,3,4}E$	0.003	-0.076	17.2	11.8	7.2	22.5	27.1	9.7	9.1	1.8	1.8
${}_{1,2,3,4}F$	0.076	0.003	17.2	27.1	11.8	7.2	22.5	9.1	9.7	1.8	1.8
${}_{1,2,3,4}G$	0.076	-0.003	17.2	22.5	7.2	11.8	27.1	9.1	9.7	1.8	1.8
${}_{1,2,3,4}H$	0.003	0.076	17.2	27.1	22.5	7.2	11.8	9.7	9.1	1.8	1.8
${}_{1,2,3,4}I$	-0.003	0.076	17.2	22.5	27.1	11.8	7.2	9.7	9.1	1.8	1.8

Table 2: Effective eccentricities and resultants for the square central footing for all combinations.

Since no imposed loads exist (Q=0) and no accidental eccentricities considered in this example (ecc_x=ecc_y=0), only 9 out of the 33 combinations are different. The effective eccentricities e_x, e_y, the mean soil pressure p_m, the soil pressure at the corners p_1, p_2, p_3, p_4 and the maximum shearing forces V_{xf}, V_{yf} and bending moments M_{xf}, M_{yf}, are presented in Table 2. It must be underlined that the effective eccentricities have been calculated using equations (2),

the soil pressure values via the explicit formulas applicable in different eccentricity regions along with mapping equations (12), while the rest of the resultants through algebraic integration of soil pressure distribution provided by equation (14). Notice that all combinations, but 'A', impose effective eccentricities despite the fact the footing itself has no structural eccentricity. This was expected since 'A' is the only combination formed exclusively by vertical loads while only horizontal loads can generate rotational reactions in central footings. The soil pressure distributions for different combinations and their envelope are illustrated in Figure 12, while the resultants presented in Table 2 and their envelopes are also shown in Figure 13.

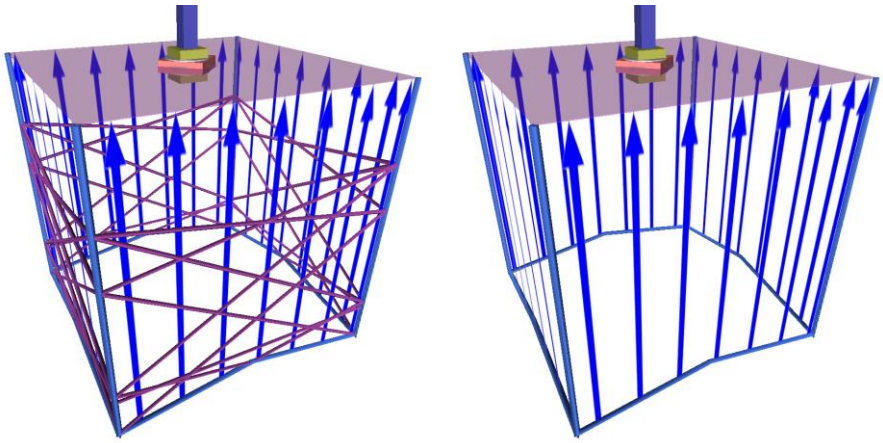


Figure 12: Soil pressure distributions for different combinations and envelope.

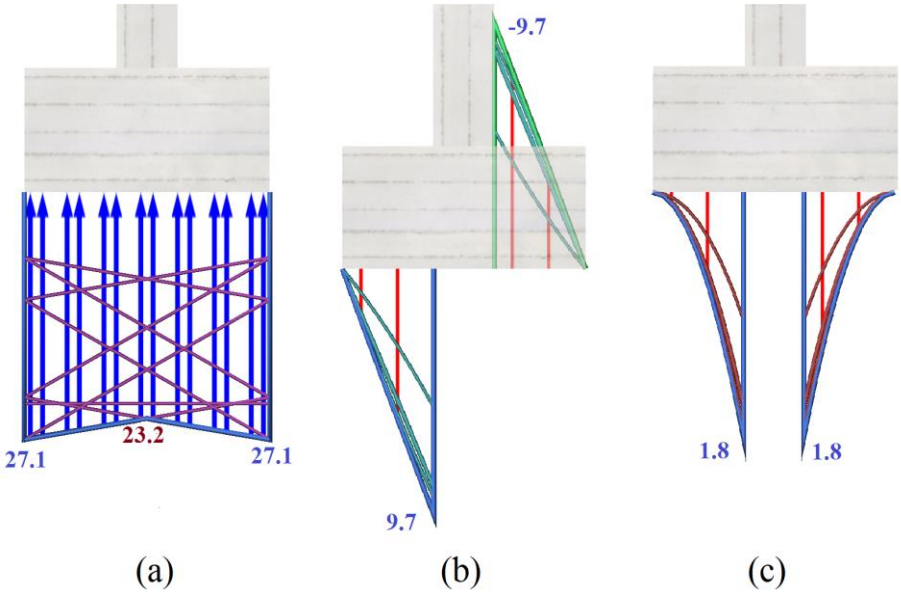


Figure 13: Resultants for different combinations and envelopes:
(a) Soil pressure, (b) Shear forces, (c) Bending moments.

The footing is also studied for variable structural eccentricity with respect to the column. The enveloped maxima of bending moments, namely M_{xd}, M_{yd}, are presented in Figure 14 as a function of structural eccentricity in x-direction. They are also called design bending moments since they are associated with footing design. Notice that only for central footings the two bending components are identical. In fact, by increasing structural eccentricity bending is affected substantially in the direction of structural eccentricity and slightly in the perpendicular direction. This reflects directly on the calculated two-way reinforcement A_{sx}, A_{sy}, shown in

Figure 15 as a function of structural eccentricity in x-direction, where required reinforcement increases substantially in the direction of structural eccentricity.

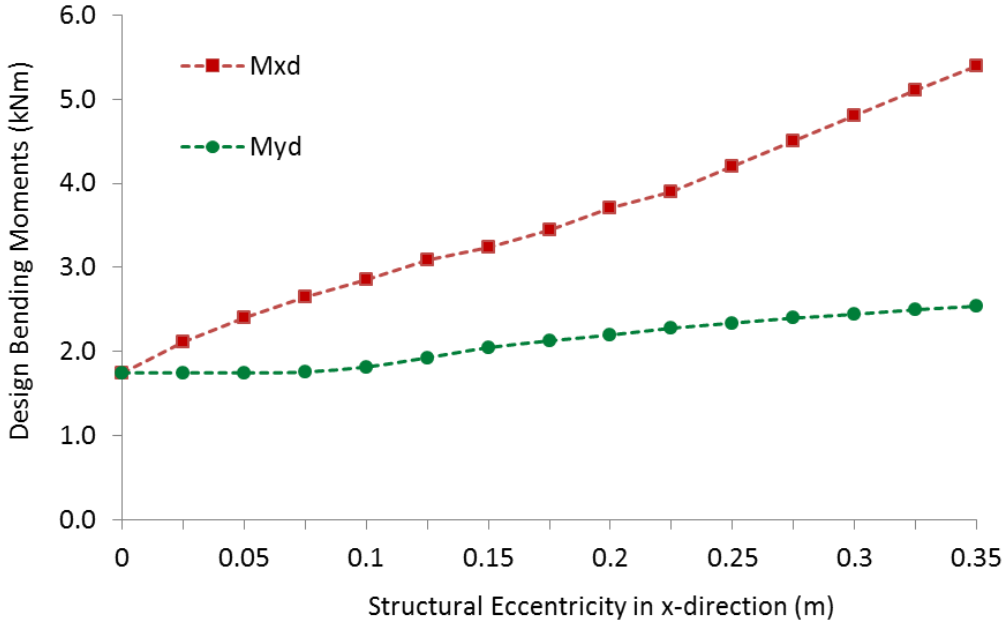


Figure 14: Design bending moments in relation to structural eccentricity in x-direction.

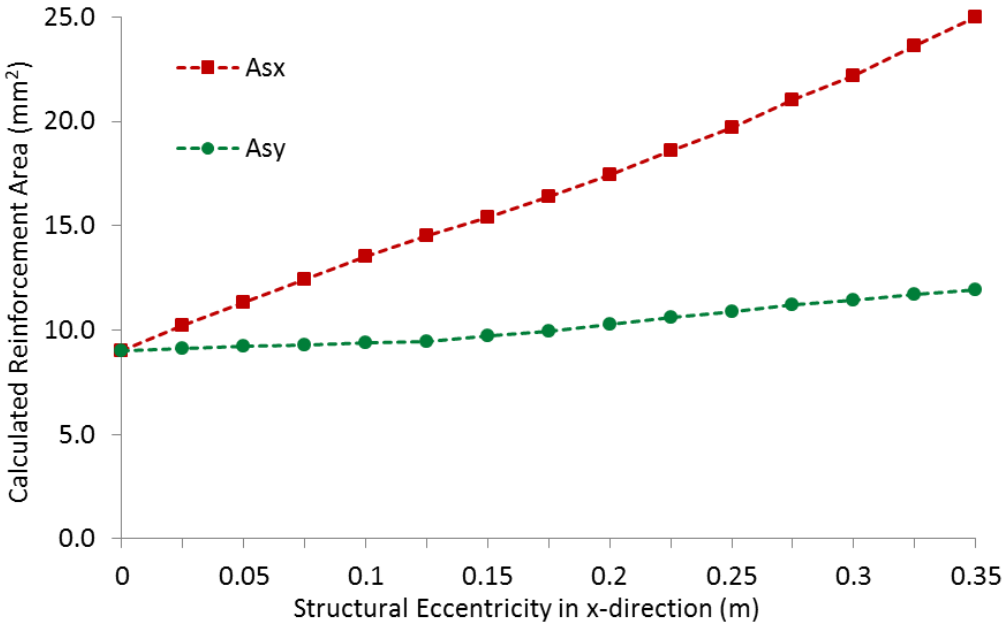


Figure 15: Calculated two-way reinforcement in relation to structural eccentricity in x-direction.

By analyzing the above outcomes, it is clear that there is a substantial amount of calculations involved in the design of rigid spread footings geometry and reinforcement, even if a limited number of combinations is taken into account. Of course, the numbers increase dramatically when optimization is attempted via iterative procedures or when one seeks resultants for multi footings foundations (see Figure 16). In such cases, the explicit formulas revealed in this article are certainly preferable over numerical methods or similar techniques, since they undoubtedly limit computational costs. Furthermore, the proposed here analytical

method is applicable to either a single footing case or a multi footings and connecting beams assembly. Only resultant reactions retrieved from the space frame model are used as input in equations (2), and therefore the process is independent from linked neighboring components.

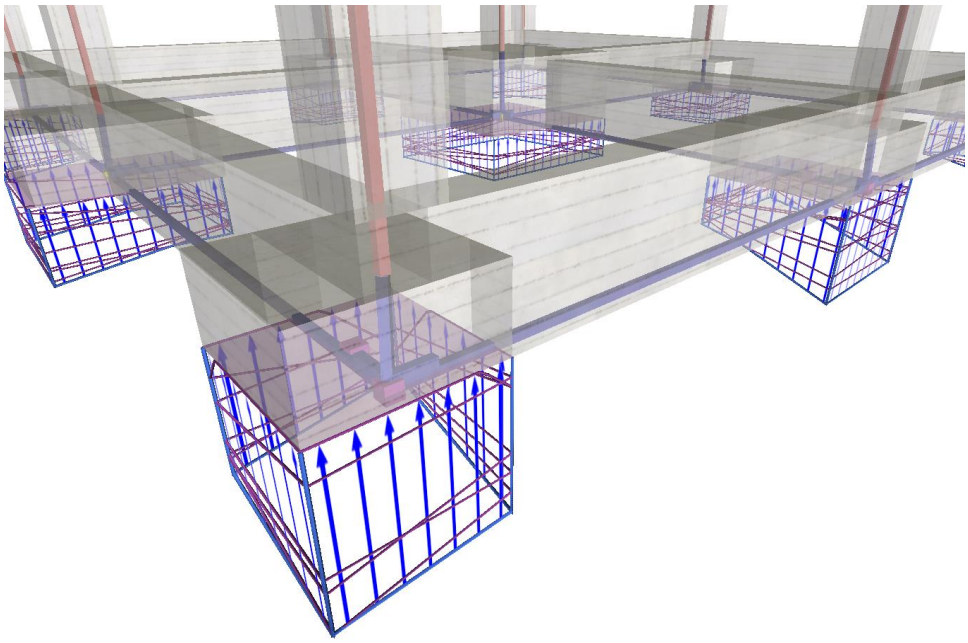


Figure 16: Soil pressure distributions and envelopes for a multi rigid spread rectangular footings foundation.

7 CONCLUSIONS

This article adds value in studying rigid spread rectangular footings resting on elastic soils by offering:

1. High computational efficiency through developed explicit formulas for soil pressure values in different eccentricity regions.
2. A generic, robust and effective solution, independent of the foundation layout, by using as input the reactions computed at the supports instead of actual loads on the footing.
3. An eccentricity diagram per footing, as a very practical qualitative tool for predetermining the form of soil pressure distribution and footing deformation expected.
4. An integrated virtual reality software, facilitating visual examination and validation of the results.

The innovative and straightforward research approach followed here will assist future researchers by providing a valuable reference through a powerful research software tool. Potentially, it could be extended to accounting for more generic footing geometries and non-linear soil pressure distributions.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to pi-Systems International S.A. for the constant financial support provided during the project, the Institute of Structural Analysis and Seismic Research of National Technical University of Athens for the contribution in developing the computational algorithms for the space frame model and all associates involved in the production of the integrated virtual reality software.

NOTATION

A	Active area of footing
P	Vertical force reaction on footing
M_x, M_y	Rotational moment reactions on footing in x and y directions
$p(x, y)$	Soil pressure distribution
p_1, p_2, p_3, p_4	Soil pressure values on footing corners
p_o, p_p, p_q, p_r	Soil pressure values on rearranged footing corners
p_{max}	Maximum soil pressure value on footing
p_{min}	Minimum soil pressure value on footing
p_m	Mean soil pressure value on footing
l_x, l_y	Footing dimensions in x and y directions
e_x, e_y	Effective eccentricities from footing center in x and y directions
x_c, y_c	Coordinates of footing base centroid
x_n, y_n	Neutral axis intercepts on x and y axes
x, y, z	Coordinate system with origin on footing corner
x', y', z'	Coordinate system with origin on footing centroid
$sign( )$	Sign of a real number
V_{xf}, V_{yf}	Maximum values of shear forces on footing in x and y directions
M_{xf}, M_{yf}	Maximum values of bending moments on footing in x and y directions
M_{xd}, M_{yd}	Design bending moments of footing in x and y directions
A_{sx}, A_{sy}	Calculated steel reinforcement of footing in x and y directions

REFERENCES

[1] J. E. Bowles, *Foundation Analysis and Design, 3rd Edition*. McGraw Hill, New York, 1977.

[2] R. Jarquio, V. Jarquio, Design of Footing Area with Biaxial Bending. *Journal of Geotechnical Engineering*, 109, 1337-1341, 1983.

[3] W. H. Highther, J. C. Anders, Dimensioning Footings Subjected to Eccentric Loads. *Journal of Geotechnical Engineering*, 111, 659-665, 1985.

[4] D. M. Vitone, A. J. Valsangkar, Stresses from Loads over Rectangular Areas. *Journal of Geotechnical Engineering Division*, 112, 961-964, 1986.

[5] R. Irles, F. Irles, Explicit Stresses under Rectangular Footings. *Journal of Geotechnical Engineering*, 120, 444–450, 1994.

[6] H. M. Algin, Stresses from Linearly Distributed Pressures over Rectangular Areas. *International Journal of Numerical Analysis Methods in Geomechanics*, 24, 681–692, 2000.

[7] H. M. Algin, Practical Formula for Dimensioning a Rectangular Footing. *Engineering Structures*, 29, 1128-1134, 2007.

[8] J. P. Smith-Pardo, A. Bobet, Behavior of Rigid Footings on Gravel under Axial Load and Moment. *Journal of Geotechnical and Geoenvironmental Engineering*, 133, 1203–1215, 2007.

[9] N., Yamamoto, M. F. Randolph, I. Einav, Simple Formulas for the Response of Shallow Foundations on Compressible Sands. *International Journal of Geomechanics*, 8, 230–239, 2008.

[10] M. H. T. Rayhani, M. H. El Naggar, Numerical Modeling of Seismic Response of Rigid Foundation on Soft Soil. *International Journal of Geomechanics*, 8, 336–346, 2008.

[11] A. Ivan, Pressures Distribution for Eccentrically Loaded Rectangular Footings on Elastic Soils. *Proceedings of the 2010 International Conference on Mathematical Models for Engineering Science*, 213-216, 2010.

[12] N. S. V. Kameswara Rao, *Analysis of Footings on Elastic Foundations, Chapter 5, Foundation Design: Theory and Practice*. John Wiley & Sons (Asia) Pte Ltd, 165–202, 2011.

[13] G. Özmen, Determination of Base Stresses in Rectangular Footings under Biaxial Bending. *Teknik Dergi*, 22, 5659-5674, 2011.

[14] A. L. Rojas, J. G. F. Herrera, R. A. A. Vallejo and M. A. C. Alvarez, Design of Isolated Footings of Rectangular Form Using a New Model. *International Journal of Innovative Computing, Information and Control*, 9, 4001-4021, 2013.

[15] J. A. Rodriguez-Gutierrez, J. D. Aristizabal-Ochoa, Rigid Spread Footings Resting on Soil Subjected to Axial Load and Biaxial Bending. II: Design Aids. *International Journal of Geomechanics*, 13, 109-119, 2013.

[16] A. Konstantinides, J. Bellos, *Earthquake Resistant Buildings made of Reinforced Concrete: Static and Dynamic Analysis according to Eurocodes, Volume B*. Alta Grafico, Athens, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A REAL-SPACE MODAL ANALYSIS METHOD FOR NON-PROPORTIONAL DAMPED STRUCTURES

E. Stanoev[1]
[1] Chair of Wind Energy Technology,
Faculty of Mechanical Engineering and Marine Technology,
University of Rostock, Albert-Einstein-Str. 2, 18059 Rostock, Germany
e-mail: evgueni.stanoev@uni-rostock.de

Keywords: modal decomposition of the equations of motion, complex eigenvalue problem, proportional/non-proportional damping, synchronous free vibrations

Abstract. *The inclusion of damping in the equations of motion of FEM-based structural models yields a complex (quadratic) eigenvalue problem. In this paper is presented a variant of a general method [4], [5] for real-space modal transformation of damped multi-degree-of-freedom-systems (MDOFS) with non-modal (non-proportional) symmetric damping matrix. The method is based on the conjugated complex right eigenvectors of the system, normalized relative to the general mass matrix. After state-space formulation of the equations of motion a real modal transformation matrix is built by a combination of two complex transformations, which is the main advantage of the presented method. Analytically expressions for the modal transformation basis are developed be the aid of computer algebra software (MATLAB).*

Applying the suggested method to the special case of proportionally damped system, an analytical expression for the constant phase lag of the free vibration modes has been derived. The conversion of the developed general real transformation matrix into the modal matrix of the undamped problem is analytically proved by taking into account the synchronous free oscillations in this special case.

The derived formulas for the modal transformation basis contain the real and the imaginary parts of the eigenvectors and the associated eigenvalues.

A numerical example – vibration of a rotor blade of a wind turbine - demonstrates the performance of the presented modal decomposition method for the general case of non-proportional damped system. The damping matrix of this example contains structural and aerodynamic damping. The initial computation of the complex eigensolution of the FEM beam model in the presented example and all subsequent computations are done by the aid of the Symbolic Math Toolbox of MATLAB. The suggested procedure can be applied in structural systems containing different damping and energy-loss mechanism in various parts of the structure and also in structure-environment interaction problems, where a non-modal damping matrix is occurring.

1 INTRODUCTION

The modal decomposition of the equations of motion of multi-degree-of-freedom-systems (MDOFS) is usually applied to systems without damping. The associated eigenvalue problem has real eigenvectors and real free frequencies. The inclusion of damping in the equations of MDOFS leads to a quadratic eigenvalue problem with complex conjugate pairs of eigenvalues and eigenmodes. The modal decomposition of the equations has to be performed in complex space. Aiming to avoid the computation in complex arithmetic, a new modal decomposition method, presented in details in [2] – [5], is briefly outlined in Sec. 2. This procedure is based on a real modal transformation matrix, derived from the complex eigenvalue solution of a MDOFS with symmetric non-proportional (non-modal) damping matrix.

In the suggested procedure the complex eigenvectors and eigenvalues of the structural model should be computed first. In the presented example in Sec. 4 – vibration of a rotor blade of a wind generator - computer algebra software was applied to solve the eigenvalue problem. In real life applications of the presented method to high dimensional problems it must be available a reliable eigenmode solver for large complex eigenvalue computations. There are many literature references for large scaled problems with various solution strategies, see [10] – [12]. The author has used an implicitly restarted Arnoldi/Lanczos method [11], [12] to solve the complex eigenvalue problem in an application of the method to a fluid-structure-foundation interaction problem, see in [1],[2].

Another topic of this paper is to show an analytical proof of the statement for the constant phase lag/lead of free vibrations in the proportional damping case – see the introduction in Sec. 1.2. The analytical proof in an indirect manner is based on the procedure, summarized in Sec. 2. A formula for computing of the constant ratio $\frac{Im(\mathbf{X})_k}{Re(\mathbf{X})_k}$ has been derived in Sec. 3.1.

In Sec. 4 the proposed modal analysis method, presented in Section 2, has been applied to a rotor blade beam structure with 54 DOF. The numerical example demonstrates the performance of the method for the general case of non-proportional damping. In this case the damping matrix of the system contains a stiffness proportional (Rayleigh) damping and aerodynamic (non-proportional) damping parts. In the second variant of the solution – with proportional damping matrix, the formula for the constant phase of the resonance modes is verified numerically.

1.1 Free vibrations of a viscously damped system

The equations of motion of a damped MDOFS are

$$\mathbf{M}\ddot{\mathbf{V}} + \mathbf{D}\dot{\mathbf{V}} + \mathbf{K}\mathbf{V} = \mathbf{p}(t) \tag{1.1}$$

where **M**, **D** and **K** are, respectively the (n x n) mass, damping and stiffness matrices, and **V**, $\dot{\mathbf{V}}$ are the (n x 1) displacement and velocity vectors and **p**(t) is the (n x 1) excitation vector.

In structural mechanics problems we consider the **M** and **K** matrices to be real, symmetric and positive definite, excluding the presence of rigid body modes. The **D** matrix is assumed to be symmetric, non-negative, she presents a non-proportional damping.

With the assumed free vibration in the form

$$\mathbf{V} = \mathbf{X}e^{\lambda t}, \qquad \dot{\mathbf{V}} = \lambda \mathbf{X}e^{\lambda t}, \tag{1.2}$$

the associated quadratic eigenvalue problem is

$$\left(\lambda_j^2 \mathbf{M} + \lambda_j \mathbf{D} + \mathbf{K}\right)\mathbf{X}_j = \mathbf{0} \qquad \forall\,(j = 1, \cdots, n) \tag{1.3}$$

In Eq. (1.3) the jth eigenvalue λ_j and the corresponding eigenmode $\mathbf{X}_j$ appear in complex conjugate pairs (index j omitted):

$$\lambda = \lambda_r + i\lambda_i, \qquad \overline{\lambda} = \lambda_r - i\lambda_i \tag{1.4a}$$

$$\mathbf{X} = \mathbf{X}_r + i\mathbf{X}_i, \qquad \overline{\mathbf{X}} = \mathbf{X}_r - i\mathbf{X}_i \tag{1.4b}$$

The dynamic equilibrium of a viscously damped single oscillator is governed by

$$m\ddot{v}(t) + c\dot{v}(t) + kv(t) = q(t) \qquad \text{resp.} \tag{1.5a}$$

$$\ddot{v}(t) + 2\eta\omega\dot{v}(t) + \omega^2 v(t) = p(t) \tag{1.5b}$$

where $\ddot{v}$ is acceleration,
$\dot{v}$ - velocity,
$\omega = \sqrt{\frac{k}{m}}$ - free vibration frequency,
$\eta = \frac{c}{2m\omega}$ - Lehr's damping ratio and $p(t) = \frac{q(t)}{m}$.

The exponential solution $x\,e^{\lambda t}$, introduced into the homogenous form of the differential equation (1.5b), yields the eigenvalue problem

$$\lambda^2 + 2\eta\omega\,\lambda + \omega^2 = 0 \tag{1.6}$$

The eigenvalue solution (assuming that $\eta \ll 1$, subcritical damping) of Eq. (1.6) is a complex conjugate pair:

$$\lambda_{1/2} = \underbrace{-\eta\omega}_{\lambda_r} \pm i\underbrace{\omega\sqrt{1-\eta^2}}_{\lambda_i = \omega_D} = \lambda_r \pm i\lambda_i \tag{1.7}$$

1.2 The constant phase lag problem

Interpreting the eigenvalue pair (1.4a) as the single-oscillator-eigenvalues (1.7), we can express the jth free vibration of the MDOFS as linear combination of the two complex conjugate eigenpairs (1.4a,b):

$$\mathbf{V} = \mathbf{X}\,e^{\lambda t} = \mathbf{X}\,e^{\left(-\eta\omega \pm i\omega\sqrt{1-\eta^2}\right)t} =$$
$$= e^{-\eta\omega t}[(\mathbf{X}_r + i\mathbf{X}_i)(cos\,\omega_D t + i\sin\omega_D t) + (\mathbf{X}_r - i\mathbf{X}_i)(cos\,\omega_D t - i\sin\omega_D t)]$$

$$= e^{-\eta\omega t}\begin{bmatrix} \mathbf{2}(\mathbf{X}_r\,cos\,\omega_D t - \mathbf{X}_i\sin\omega_D t) + \\ i(\mathbf{X}_i\,cos\,\omega_D t + \mathbf{X}_r\sin\omega_D t) \;-\; i\,(\mathbf{X}_i\,cos\,\omega_D t + \mathbf{X}_r\sin\omega_D t) \end{bmatrix}$$

$$= e^{-\eta\omega t}\left[\underbrace{2\mathbf{X}_r}_{\mathbf{F}\cos\varphi}\ cos\,\omega_D t - \underbrace{2\mathbf{X}_i}_{\mathbf{F}\sin\varphi}\ \sin\omega_D t\right] \tag{1.8}$$

The last relation leads to the real form of a damped free oscillation for every k[th] DOF:

$$V_k = e^{-\eta\omega t}[F_k\, cos(\omega_D t + \varphi_k)] \tag{1.9}$$

where $\varphi_k = arctan\frac{(X_i)_k}{(X_r)_k}$: phase lag/lead for the kth DOF (1.10)

Since the viscous damping is assumed to be non-proportional, the free vibration solution (1.9) represents non-synchronous damped oscillation (i.e. the phase φ_k is different for each DOF). In the case of proportionally damped system we have to deal with synchronous free oscillation – i.e. the phase φ_k is constant (the same for all DOF), for undamped systems φ_k is zero – see [6], [7], p.118.

The features, showed in Eq.(1.9), (1.10) are well known and used in modal analysis, see for example [6]. In the present paper the statement of synchronous free oscillations in the proportional damping case should be proved analytically in Sec. 3.1.

2 MODAL DECOMPOSITION METHOD INCLUDING THE COMPLEX RIGHT EIGENVECTORS

2.1 The single mass oscillator

The equation of motion of a damped single degree of freedom system (SDOFS) (1.5b) can be written in the form

$$\underbrace{\begin{bmatrix}1 & 0\\ 0 & -\omega^2\end{bmatrix}}_{\mathbf{m}}\underbrace{\begin{bmatrix}\dot{w}(t)\\ \dot{v}(t)\end{bmatrix}}_{\dot{\mathbf{q}}}+\underbrace{\begin{bmatrix}2\eta\omega & \omega^2\\ \omega^2 & 0\end{bmatrix}}_{\mathbf{k}}\underbrace{\begin{bmatrix}w(t)\\ v(t)\end{bmatrix}}_{\mathbf{q}}=\underbrace{\begin{bmatrix}p(t)\\ 0\end{bmatrix}}_{\mathbf{p}(t)} \tag{2.1a}$$

$$\mathbf{m}\dot{\mathbf{q}}+\mathbf{k}\,\mathbf{q}=\mathbf{p} \quad \rightarrow \quad \dot{\mathbf{q}}+\underbrace{\left(\mathbf{m}^{-1}\mathbf{k}\right)}_{\underline{\mathbf{a}}}\mathbf{q}=\mathbf{m}^{-1}\mathbf{p} \tag{2.1b}$$

where the velocity is

$$w=\dot{v} \tag{2.2}$$

The exponential solution $\mathbf{q}=\mathbf{x}e^{\lambda t}$, $\dot{\mathbf{q}}=\lambda\mathbf{x}e^{\lambda t}$, introduced into the homogenous form of the differential equation (2.1), gives the quadratic eigenvalue problem

$$\left(\lambda\underbrace{\begin{bmatrix}1 & 0\\ 0 & -\omega^2\end{bmatrix}}_{\mathbf{m}}+\underbrace{\begin{bmatrix}2\eta\omega & \omega^2\\ \omega^2 & 0\end{bmatrix}}_{\mathbf{k}}\right)\underbrace{\begin{bmatrix}\lambda\\ 1\end{bmatrix}}_{\mathbf{x}}=\mathbf{0} \quad \rightarrow \quad (\lambda\mathbf{m}+\mathbf{k})\mathbf{x}=\mathbf{0} \tag{2.3}$$

The two complex conjugate eigenvalues ($\eta << 1$, subcritical damped system) are:

$$\lambda_{1/2} = -\eta\omega \pm \sqrt{(\eta\omega)^2 - \omega^2} = \underbrace{-\eta\omega}_{\lambda_r} \pm i\underbrace{\omega\sqrt{1-\eta^2}}_{\lambda_i} = \lambda_r \pm i\lambda_i \tag{2.4}$$

where $\omega = \sqrt{(\lambda_r)^2 + (\lambda_i)^2}$, $\eta = -\frac{\lambda_r}{\omega}$, $\omega_D = \omega\sqrt{1-\eta^2}$ (2.5)

The two corresponding complex conjugate eigenvectors $\varphi_{1/2}$, at first normalized relative to the mass matrix

$$\varphi_{\mathbf{k}} = \frac{\mathbf{x_k}}{\sqrt{\mathbf{x_k^T m x_k}}} = \frac{1}{\sqrt{-\omega^2 + (-\lambda_r \pm i\lambda_i)^2}} \begin{bmatrix} \lambda_r \pm i\lambda_i \\ 1 \end{bmatrix}, \qquad (k = 1,2) \tag{2.6}$$

are combined into a modal matrix:

$$\varphi = [\varphi_1 \quad \varphi_2] \tag{2.7}$$

Due to normalization Eq. (2.6) the orthogonality relationships can be derived:

$$\varphi^{\mathbf{T}}\mathbf{m}\varphi = \varphi^{\mathbf{T}}\begin{bmatrix} 1 & 0 \\ 0 & -\omega^2 \end{bmatrix}\varphi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \leftrightarrow \quad \varphi^{-\mathbf{T}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\varphi^{-\mathbf{1}} = \begin{bmatrix} 1 & 0 \\ 0 & -\omega^2 \end{bmatrix} \tag{2.8}$$

$$\varphi^{\mathbf{T}}\mathbf{k}\varphi = \varphi^{\mathbf{T}}\begin{bmatrix} 2\eta\omega & \omega^2 \\ \omega^2 & 0 \end{bmatrix}\varphi = \begin{bmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{bmatrix} \leftrightarrow \varphi^{-\mathbf{T}}\begin{bmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{bmatrix}\varphi^{-\mathbf{1}} = \begin{bmatrix} 2\eta\omega & \omega^2 \\ \omega^2 & 0 \end{bmatrix} \tag{2.9}$$

The inverse of the complex modal matrix $\varphi(\omega, \eta)$ can be expressed analytically using computer algebra software:

$$\varphi^{-1} = \frac{1}{2\sqrt{1-\eta^2}} \begin{bmatrix} (-Z_1 - iZ_2) & P - iQ \\ (-Z_1 + iZ_2) & P + iQ \end{bmatrix} \tag{2.10}$$

where

$$Z_1 = \sqrt{\sqrt{1-\eta^2} + (1-\eta^2)} \qquad Z_2 = \sqrt{\sqrt{1-\eta^2} - (1-\eta^2)} \tag{2.11a}$$

$$P = \omega\left(\sqrt{1-\eta^2}\, Z_2 - \eta Z_1\right) \qquad Q = \omega\left(\sqrt{1-\eta^2}\, Z_1 + \eta Z_2\right) \tag{2.11b}$$

2.2 The damped multi-degree-of-freedom-system

The equations of motion (1.1) of damped MDOFS (n DOF) will be written in the state-space form:

$$\underbrace{\begin{bmatrix} \mathbf{M} & \\ & -\mathbf{K} \end{bmatrix}}_{\mathbf{M_G}} \underbrace{\begin{bmatrix} \dot{\mathbf{W}} \\ \dot{\mathbf{V}} \end{bmatrix}}_{\dot{\mathbf{Q}}} + \underbrace{\begin{bmatrix} \mathbf{D} & \mathbf{K} \\ \mathbf{K} & \end{bmatrix}}_{\mathbf{K_G}} \underbrace{\begin{bmatrix} \mathbf{W} \\ \mathbf{V} \end{bmatrix}}_{\mathbf{Q}} = \underbrace{\begin{bmatrix} \mathbf{p}(t) \\ \end{bmatrix}}_{\mathbf{P}} \quad ; \qquad \dot{\mathbf{V}} = \mathbf{W} \tag{2.12a}$$

$$\mathbf{M_G}\,\dot{\mathbf{Q}} + \mathbf{K_G}\,\mathbf{Q} = \mathbf{P}\ , \tag{2.12b}$$

where $\mathbf{M_G}$ and $\mathbf{K_G}$ are, respectively the (2n x 2n) symmetric generalized mass and the generalized stiffness matrices. The symmetric damping matrix $\mathbf{D}$ is non-negative and represents a non-proportional damping.

The exponential solution (1.2), substituted into the homogenous form of Eq.(2.12), leads to the 2n-dimensional eigenvalue problem

$$\left(\lambda \mathbf{M_G} + \mathbf{K_G}\right)\begin{bmatrix}\lambda \mathbf{X} \\ \mathbf{X}\end{bmatrix} = \mathbf{0} \tag{2.13}$$

The solution of Eq. (2.13) is given by n complex conjugate eigenpairs (1.4), now written in the form:

$$\lambda^{(j)} = \lambda_r^{(j)} + i\lambda_i^{(j)} \quad \rightarrow \begin{bmatrix}\lambda^{(j)}\mathbf{X}^{(j)} \\ \mathbf{X}^{(j)}\end{bmatrix}; \qquad \overline{\lambda}^{(j)} = \lambda_r^{(j)} - i\lambda_i^{(j)} \quad \rightarrow \begin{bmatrix}\overline{\lambda}^{(j)}\overline{\mathbf{X}}^{(j)} \\ \overline{\mathbf{X}}^{(j)}\end{bmatrix}; \quad (j = 1,2\ldots n) \tag{2.14}$$

Each jth eigenvector-pair $\mathbf{X}^{(j)}$, $\overline{\mathbf{X}}^{(j)}$ is normalized (index (j) omitted) relative to the general mass matrix $\mathbf{M_G}$:

$$\mathbf{\Phi} = \frac{\mathbf{X}}{\sqrt{A+iB}} = \mathbf{\Phi_r} + i\,\mathbf{\Phi_i}, \qquad A + iB = \begin{bmatrix}\lambda \mathbf{X} \\ \mathbf{X}\end{bmatrix}^T \mathbf{M_G} \begin{bmatrix}\lambda \mathbf{X} \\ \mathbf{X}\end{bmatrix} \tag{2.15a}$$

$$\overline{\mathbf{\Phi}} = \frac{\overline{\mathbf{X}}}{\sqrt{A-iB}} = \mathbf{\Phi_r} - i\,\mathbf{\Phi_i}, \qquad A - iB = \begin{bmatrix}\overline{\lambda}\,\overline{\mathbf{X}} \\ \overline{\mathbf{X}}\end{bmatrix}^T \mathbf{M_G} \begin{bmatrix}\overline{\lambda}\,\overline{\mathbf{X}} \\ \overline{\mathbf{X}}\end{bmatrix} \tag{2.15b}$$

Subject to the normalization (2.15) follow the orthogonality relationships –expressed in terms of the jth eigenvector-pair (index (j) omitted):

$$\begin{bmatrix}\lambda \mathbf{\Phi} & \overline{\lambda}\,\overline{\mathbf{\Phi}} \\ \mathbf{\Phi} & \overline{\mathbf{\Phi}}\end{bmatrix}^T \underbrace{\begin{bmatrix}\mathbf{M} & \\ & -\mathbf{K}\end{bmatrix}}_{\mathbf{M_G}} \begin{bmatrix}\lambda \mathbf{\Phi} & \overline{\lambda}\,\overline{\mathbf{\Phi}} \\ \mathbf{\Phi} & \overline{\mathbf{\Phi}}\end{bmatrix} = \begin{bmatrix}1 & \\ & 1\end{bmatrix} \tag{2.16}$$

$$\begin{bmatrix}\lambda \mathbf{\Phi} & \overline{\lambda}\,\overline{\mathbf{\Phi}} \\ \mathbf{\Phi} & \overline{\mathbf{\Phi}}\end{bmatrix}^T \underbrace{\begin{bmatrix}\mathbf{D} & \mathbf{K} \\ \mathbf{K} & \end{bmatrix}}_{\mathbf{K_G}} \begin{bmatrix}\lambda \mathbf{\Phi} & \overline{\lambda}\,\overline{\mathbf{\Phi}} \\ \mathbf{\Phi} & \overline{\mathbf{\Phi}}\end{bmatrix} = \begin{bmatrix}-\lambda & \\ & -\overline{\lambda}\end{bmatrix} \tag{2.17}$$

The (2n x 2n) complex square modal matrix, denoted by $\mathbf{\Phi_G}$, is made up of the n eigenvector-pairs, see Eqn.(2.15):

$$\mathbf{\Phi_G} = \begin{bmatrix}\lambda^{(1)}\mathbf{\Phi}^{(1)} & \overline{\lambda}^{(1)}\overline{\mathbf{\Phi}}^{(1)} & \ldots & \ldots & \lambda^{(n)}\mathbf{\Phi}^{(n)} & \overline{\lambda}^{(n)}\overline{\mathbf{\Phi}}^{(n)} \\ \\ \mathbf{\Phi}^{(1)} & \overline{\mathbf{\Phi}}^{(1)} & \ldots & \ldots & \mathbf{\Phi}^{(n)} & \overline{\mathbf{\Phi}}^{(n)}\end{bmatrix} \tag{2.18}$$

The orthogonality properties – see Eq. (2.16), (2.17), are used to perform a modal decomposition of the equations of motion (2.12):

$$\underbrace{\mathbf{\Phi}_G^T \begin{bmatrix}\mathbf{M} & \\ & -\mathbf{K}\end{bmatrix} \mathbf{\Phi_G}}_{\mathbf{E}} \dot{\mathbf{A}} \quad + \quad \underbrace{\mathbf{\Phi}_G^T \begin{bmatrix}\mathbf{D} & \mathbf{K} \\ \mathbf{K} & \end{bmatrix} \mathbf{\Phi_G}}_{diag\{-\lambda^{(j)}\}} \mathbf{A} \quad = \quad \mathbf{\Phi}_G^T \begin{bmatrix}\mathbf{p}(t) \\ \end{bmatrix} \tag{2.19}$$

where

$$\begin{bmatrix}\mathbf{W}\\ \mathbf{V}\end{bmatrix} = \mathbf{\Phi_G}\cdot\mathbf{A} = \mathbf{\Phi_G}\cdot\left[a^{(1)}\ b^{(1)}\quad\cdots\quad a^{(n)}\ b^{(n)}\right]^T \tag{2.20}$$

is a coordinate transformation, and $a^{(j)}$, $b^{(j)}$ are new complex variables.

Introducing real modal coordinates $x^{(j)}$, $y^{(j)}$ for each jth eigenpair, i.e.:

$$\begin{bmatrix}a^{(j)}\\ \\ b^{(j)}\end{bmatrix} = \left(\varphi^{(j)}\right)^{-1}\begin{bmatrix}x^{(j)}\\ \\ y^{(j)}\end{bmatrix}, \tag{2.21}$$

the differential equations (2.19) can be transformed **in pairs** into the real form of SDOFS-equation (index (j) omitted), regarding Eqs.(2.8), (2.9) and using (2.5):

$$\begin{aligned}
&\underbrace{\varphi^{-\mathbf{T}}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\varphi^{-\mathbf{1}}}\cdot\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} + \underbrace{\varphi^{-\mathbf{T}}\begin{bmatrix}-\lambda & 0\\ 0 & -\overline{\lambda}\end{bmatrix}\varphi^{-\mathbf{1}}}\cdot\begin{bmatrix}x\\ y\end{bmatrix} = \underbrace{\varphi^{-\mathbf{T}}\begin{bmatrix}\lambda\mathbf{\Phi}^T\mathbf{p}(t)\\ \overline{\lambda}\overline{\mathbf{\Phi}}^T\mathbf{p}(t)\end{bmatrix}}\\
&\begin{bmatrix}1 & 0\\ 0 & -\omega^2\end{bmatrix}\cdot\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} + \begin{bmatrix}2\eta\omega & \omega^2\\ \omega^2 & \end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}g(t)\\ h(t)\end{bmatrix}
\end{aligned} \tag{2.22}$$

For each eigenvalue pair $\lambda^{(j)} = \lambda_r^{(j)} \pm i\lambda_i^{(j)}$ the matrix $[\varphi(\omega,\eta)]^{-1}$ of the corresponding SDOFS can be computed by Eqs. (2.4),(2.5),(2.10), (2.11).

2.3 The real modal transformation basis

Using both transformations (2.19) and (2.22), the equations of motion (2.12) will be uncoupled into n real SDOFS block equations as follows:

$$\underbrace{\mathbf{Y}^{\mathrm{T}}\cdot\begin{bmatrix}\mathbf{M} & \\ & -\mathbf{K}\end{bmatrix}\cdot\mathbf{Y}}_{\begin{bmatrix}1 & & & & \\ & -\omega_1^2 & & & \\ & & \cdots & & \\ & & & 1 & \\ & & & & -\omega_n^2\end{bmatrix}}\cdot\underbrace{\begin{bmatrix}\dot{x}_1\\ \dot{y}_1\\ \cdots\\ \dot{x}_n\\ \dot{y}_n\end{bmatrix}}_{\dot{\mathbf{X}}} + \underbrace{\mathbf{Y}^{\mathrm{T}}\cdot\begin{bmatrix}\mathbf{D} & \mathbf{K}\\ \mathbf{K} & \end{bmatrix}\cdot\mathbf{Y}}_{\begin{bmatrix}2\eta_1\omega_1 & \omega_1^2 & & & \\ -\omega_1^2 & 0 & & & \\ & & \cdots & & \\ & & & 2\eta_n\omega_n & \omega_n^2\\ & & & \omega_n^2 & 0\end{bmatrix}}\cdot\underbrace{\begin{bmatrix}x_1\\ y_1\\ \cdots\\ x_n\\ y_n\end{bmatrix}}_{\mathbf{X}} = \underbrace{\mathbf{Y}^{\mathrm{T}}\cdot\begin{bmatrix}\mathbf{p}\\ \end{bmatrix}}_{\begin{bmatrix}g_1\\ h_1\\ \cdots\\ g_n\\ h_n\end{bmatrix}} \tag{2.23}$$

The new (2n x 2n) transformation basis $\mathbf{Y}$ is defined by combination of two complex transformations (2.19), (2.22):

$$\begin{bmatrix}\mathbf{W}\\ \mathbf{V}\end{bmatrix} = \mathbf{\Phi_G}\cdot\underbrace{\begin{bmatrix}\left(\varphi^{(\mathbf{1})}\right)^{-\mathbf{1}} & & \\ & \cdots & \\ & & \left(\varphi^{(\mathbf{n})}\right)^{-\mathbf{1}}\end{bmatrix}}_{\mathbf{\Psi}^{-1}}\cdot\underbrace{\begin{bmatrix}x_1\\ y_1\\ \cdots\\ x_n\\ y_n\end{bmatrix}}_{\mathbf{X}} = \underbrace{\mathbf{\Phi_G}\cdot\mathbf{\Psi}^{-\mathbf{1}}}_{\mathbf{Y}}\cdot\mathbf{X} = \mathbf{Y}\cdot\mathbf{X} \tag{2.24}$$

It can be shown that the $\mathbf{Y}$-matrix in Eq. (2.24) and all „load"-vectors $[g(t)\quad h(t)]^T$, see Eq.(2.22), are purely real. After component multiplication of the analytically expressed terms

of $\boldsymbol{\Phi_G}$ and of $\boldsymbol{\Psi^{-1}}$ all imaginary parts cancel each other, see details in [4]. This is briefly sketched below by developing the two columns of $\mathbf{Y}$, belonging to the jth eigenvector-pair:

$$\left[\begin{array}{c|cc|c} \cdots & \mathbf{Y}_{\mathbf{x}}^{(j)\mathbf{W}} & \mathbf{Y}_{\mathbf{y}}^{(j)\mathbf{W}} & \cdots \\ \hline \cdots & \mathbf{Y}_{\mathbf{x}}^{(j)\mathbf{V}} & \mathbf{Y}_{\mathbf{y}}^{(j)\mathbf{V}} & \cdots \end{array}\right] = \underbrace{\left[\begin{array}{c|cc|c} \cdots & \lambda^{(j)}\,\boldsymbol{\Phi}^{(j)} & \overline{\lambda}^{(j)}\,\overline{\boldsymbol{\Phi}}^{(j)} & \cdots \\ & & & \\ \cdots & \boldsymbol{\Phi}^{(j)} & \overline{\boldsymbol{\Phi}}^{(j)} & \cdots \end{array}\right]}_{\boldsymbol{\Phi_G}} \cdot \underbrace{\begin{bmatrix} \cdots & & \\ & \left(\varphi^{(j)}\right)^{-1} & \\ & & \cdots \end{bmatrix}}_{\boldsymbol{\Psi^{-1}}} \tag{2.25}$$

With regard to Eq. (2.5), (2.10), (2.11) the multiplication

$$\begin{bmatrix} (\lambda_r + i\lambda_i)(\boldsymbol{\Phi}_{\mathbf{r}} + i\,\boldsymbol{\Phi}_{\mathbf{i}}) & (\lambda_r - i\lambda_i)(\boldsymbol{\Phi}_{\mathbf{r}} - i\,\boldsymbol{\Phi}_{\mathbf{i}}) \\ \boldsymbol{\Phi}_{\mathbf{r}} + i\,\boldsymbol{\Phi}_{\mathbf{i}} & \boldsymbol{\Phi}_{\mathbf{r}} - i\,\boldsymbol{\Phi}_{\mathbf{i}} \end{bmatrix} \cdot \underbrace{\frac{1}{2\sqrt{1-\eta^2}} \begin{bmatrix} (-Z_1 - iZ_2) & P - iQ \\ (-Z_1 + iZ_2) & P + iQ \end{bmatrix}}_{\left(\varphi^{(j)}\right)^{-1}} \tag{2.26}$$

leads to purely real components of the two columns of the transformation basis $\mathbf{Y}$:

$$\begin{aligned} \mathbf{Y}_{\mathbf{x}}^{\mathbf{W}} &= \frac{1}{\sqrt{1-\eta^2}} \left\{ (\omega_D\, Z_2 + \eta\omega\, Z_1)\boldsymbol{\Phi}_{\mathbf{r}} + (\omega_D\, Z_1 - \eta\omega\, Z_2)\boldsymbol{\Phi}_{\mathbf{i}} \right\} \\ \mathbf{Y}_{\mathbf{y}}^{\mathbf{W}} &= \frac{1}{\sqrt{1-\eta^2}} \left(\omega^2\, Z_1\, \boldsymbol{\Phi}_{\mathbf{r}} - \omega^2\, Z_2\, \boldsymbol{\Phi}_{\mathbf{i}} \right) \\ \mathbf{Y}_{\mathbf{x}}^{\mathbf{V}} &= \frac{1}{\sqrt{1-\eta^2}} \left(-Z_1\, \boldsymbol{\Phi}_{\mathbf{r}} + Z_2\, \boldsymbol{\Phi}_{\mathbf{i}} \right) \\ \mathbf{Y}_{\mathbf{y}}^{\mathbf{V}} &= \frac{1}{\sqrt{1-\eta^2}} \left\{ (\omega_D\, Z_2 - \eta\omega\, Z_1)\boldsymbol{\Phi}_{\mathbf{r}} + (\omega_D\, Z_1 + \eta\omega\, Z_2)\boldsymbol{\Phi}_{\mathbf{i}} \right\} \end{aligned} \tag{2.27a-d}$$

In the same manner we develop the "load"-vector in Eq. (2.22)

$$\begin{bmatrix} g(t) \\ h(t) \end{bmatrix} = \frac{1}{2\sqrt{1-\eta^2}} \begin{bmatrix} (-Z_1 - iZ_2) & (-Z_1 + iZ_2) \\ P - iQ & P + iQ \end{bmatrix} \begin{bmatrix} (\lambda_r + i\lambda_i)\left(\boldsymbol{\Phi}_{\mathbf{r}}^{\mathbf{T}} + i\boldsymbol{\Phi}_{\mathbf{i}}^{\mathbf{T}}\right) \\ (\lambda_r - i\lambda_i)\left(\boldsymbol{\Phi}_{\mathbf{r}}^{\mathbf{T}} - i\boldsymbol{\Phi}_{\mathbf{i}}^{\mathbf{T}}\right) \end{bmatrix} \mathbf{p}(t) \tag{2.28}$$

$$g(t) = \frac{\omega}{\sqrt{1-\eta^2}} \left\{ \left(Z_2\sqrt{1-\eta^2} + Z_1\eta \right)\boldsymbol{\Phi}_{\mathbf{r}}^{\mathbf{T}} + \left(Z_1\sqrt{1-\eta^2} - Z_2\,\eta \right)\boldsymbol{\Phi}_{\mathbf{i}}^{\mathbf{T}} \right\} \mathbf{p}(t) \tag{2.29a}$$

$$h(t) = \frac{\omega^2}{\sqrt{1-\eta^2}} \left\{ Z_1 \boldsymbol{\Phi}_{\mathbf{r}}^{\mathbf{T}} - Z_2\, \boldsymbol{\Phi}_{\mathbf{i}}^{\mathbf{T}} \right\} \mathbf{p}(t) \tag{2.29b}$$

Each jth SDOFS block equation in (2.23) can be easily solved, eliminating the modal coordinate $x^{(j)}$ to obtain the usual form of the SDOFS equation of motion (index (j) omitted):

$$x = \dot{y} + \frac{1}{\omega^2} h(t) \tag{2.30a}$$

$$\ddot{y} + 2\eta\omega\,\dot{y} + \omega^2 y = g(t) - \frac{2\eta}{\omega} h(t) - \frac{1}{\omega^2} \dot{h}(t) \tag{2.30b}$$

The dynamic response $y^{(j)}(t)$ can be obtained by step-by-step integration, applied to Eqs. (2.30b,a). The final time response of the original n DOFs is calculated by superposition of the modal coordinates $x^{(j)}$, $y^{(j)}$ in accordance to Eq. (2.24).

The major advantage of the suggested method is the developed new modal transformation matrix $\mathbf{Y}$, see (2.24), in real space for damped MDOFS with symmetric non-diagonalisable (i.e. non-modal) damping matrix. The method has also the usual modal superposition advantage - an uncompleted transformation employing only a few modes (k<<n) in the $\mathbf{Y}$ -basis leads with sufficient numerical accuracy - after the final back coordinate transformation - to the dynamic response of all n DOF.

3 THE PROPORTIONAL DAMPED SYSTEM

3.1 Modal transformation of the equations of motion

A simple method to construct a damping matrix $\mathbf{D}_p$, presenting a proportional damping, is the Rayleigh damping assumption:

$$\mathbf{D_p} = (\alpha\,\mathbf{M} + \beta\,\mathbf{K}) \tag{3.1a}$$

where α, β : unknown weighting parameter, see Eq.(3.7),(3.8a,b)

The modal damping matrix is a particular case of a more general proportional damping assumption, see [7] p.105, in the form:

$$\mathbf{D_p} = \sum_{k=1}^{n} a_k \mathbf{M}\,(\mathbf{M^{-1}K})^{k-1} \tag{3.1b}$$

The matrix (3.1b) turns for $n = 2$ to $\mathbf{D_p} = a_1\mathbf{M} + a_2\mathbf{K}$, which is the Rayleigh approach (3.1a).

The eigenvalue problem

$$\left(\lambda_j^2\,\mathbf{M} + \mathbf{K}\right)\mathbf{X}_j = \mathbf{0}\,, \tag{3.2}$$

corresponding to the equations of motion of MDOFS without damping

$$\mathbf{M\ddot{V}} + \mathbf{KV} = \mathbf{p}(t), \tag{3.3}$$

has the solution: $\lambda_j = i\omega_{0j}$ with the free frequency ω_{0j}
$\mathbf{X}_j$, $(j = 1,2\ \cdots, n)$ real eigenvectors

The modal matrix $\mathbf{U}$, belonging to (3.2)

$$\mathbf{U} = [\mathbf{U}_1 \quad \mathbf{U}_2 \quad \cdots \quad \mathbf{U}_n] \tag{3.4}$$

comprises n real, mass normalized eigenvectors

$$\mathbf{U}_j = \frac{\mathbf{X}_j}{\sqrt{\mathbf{X}_j^T\,\mathbf{M}\,\mathbf{X}_j}} \qquad (j = 1,2,\cdots,n). \tag{3.5}$$

The mass normalisation (3.5) leads to the orthogonality relationships

$$\mathbf{U}^{\boldsymbol{T}}\mathbf{M}\mathbf{U} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \dots & \\ & & & 1 \end{bmatrix} \tag{3.6a}$$

$$\mathbf{U}^{\boldsymbol{T}}\mathbf{K}\mathbf{U} = \begin{bmatrix} \omega_{01}^2 & & & \\ & \omega_{02}^2 & & \\ & & \dots & \\ & & & \omega_{0n}^2 \end{bmatrix} = \mathbf{\Omega} \tag{3.6b}$$

The Rayleigh damping matrix $\mathbf{D_p}$ can be diagonalized applying (3.6a,b) to (3.1a):

$$\mathbf{U}^{\boldsymbol{T}}\mathbf{D_p}\mathbf{U} = \begin{bmatrix} \alpha+\beta\omega_{01}^2 & & & \\ & \dots & & \\ & & \dots & \\ & & & \alpha+\beta\omega_{0n}^2 \end{bmatrix} = \begin{bmatrix} 2\eta_1\,\omega_{01} & & & \\ & \dots & & \\ & & \dots & \\ & & & 2\eta_n\,\omega_{0n} \end{bmatrix} \tag{3.7}$$

The general form of the $\mathbf{D_p}$ matrix (3.1b) can also be transformed in diagonal form by use of the real modal matrix $\mathbf{U}$, as shown in [7] p. 105.

The terms in the main diagonal of the right side of (3.7) are set to be equal to the modal damping term $2\eta\omega_0$ of the equation of motion for SDOFS (2.1). The two unknown parameter α and β can be calculated by solving a system of two equations $\alpha+\beta\omega_{0j}^2 = 2\eta_j\,\omega_{0j}$, $(j = 1{,}2)$, using the first two lowest free frequencies ω_{0j} and by arbitrary choose of two appropriate damping ratios η_j :

$$\begin{aligned} \alpha &= \frac{2\omega_{01}\omega_{02}(\omega_{01}\eta_2 - \omega_{02}\eta_1)}{\omega_{01}^2-\omega_{02}^2} \\ \beta &= \frac{2(\omega_{01}\eta_1 - \omega_{02}\eta_2)}{\omega_{01}^2-\omega_{02}^2} \end{aligned} \tag{3.8a,b}$$

Typically for the viscous damping is evidently the frequency related damping parameters. Below the parameter α and β are set to be known. In the general case (3.1b) the unknown coefficients a_k are to be determined by solving a system of n linear equation, see [7] for details.

We consider now the proportionally damped system (1.1) with $\mathbf{D} = \mathbf{D_p}$. Assuming the solution (1.2), the associated quadratic eigenvalue problem (1.3) gives

$$\lambda_j^2\,\mathbf{M}\,\mathbf{\Phi}_j + \lambda_j \underbrace{(\alpha\,\mathbf{M} + \beta\,\mathbf{K})}_{\mathbf{D_p}}\,\mathbf{\Phi}_j + \mathbf{K}\,\mathbf{\Phi}_j = \mathbf{0} \tag{3.9}$$

The mass normalized eigenvectors $\mathbf{\Phi}_j$ in Eq. (3.9) are generally complex conjugate, see Eq. (2.13) – (2.15). But the eigenvalue problem (3.9) possess also "classical" real eigenmodes, identical to the eigenmodes $\mathbf{U}_j$, belonging to the eigenvalue problem without damping, see Eq. (3.2) – (3.5).

Using the $\mathbf{U}_j$ eigenmodes instead of $\mathbf{\Phi}_j$**,** the eigenvalue problem (3.9) can be transformed, with regard to (3.6), (3.7), to

$$\left\{\lambda_j^2\,\mathbf{M}+\ \lambda_j\underbrace{\left(\alpha\ +\beta\omega_{0j}^2\ \right)\mathbf{M}}_{\mathbf{D_p}}+\underbrace{\omega_{0j}^2\mathbf{M}}_{\mathbf{K}}\right\}\mathbf{U}_j=\mathbf{0} \tag{3.11a}$$

The corresponding complex eigenvalue λ_j is then computed from

$$\lambda_j^2\ +\ \lambda_j\left(\alpha\ +\beta\omega_{0j}^2\ \right)+\omega_{0j}^2=0$$

$$\rightarrow\quad \lambda_{j1,2}=\ -\frac{1}{2}\underbrace{\left(\alpha\ +\beta\omega_{0j}^2\ \right)}_{2\eta_j\omega_{0j}}\ \pm\ \sqrt{\frac{1}{4}\left(\alpha\ +\beta\omega_{0j}^2\ \right)^2-\omega_{0j}^2}\ =$$

$$\rightarrow\quad \lambda_{j1,2}=\underbrace{-\eta_j\omega_{0j}}_{\lambda_r}\ \pm\ i\ \underbrace{\omega_{0j}\sqrt{1-\eta_j^2}}_{\lambda_i} \tag{3.11b}$$

By comparing (3.11b) to (1.7) is evidently, that the free frequency ω, computed according to (2.4), (2.5)

$$\omega=\sqrt{\lambda_r^2+\lambda_i^2}=\left\{(\eta\omega)^2+\left(\omega\sqrt{1-\eta^2}\right)^2\right\}^{\frac{1}{2}} \tag{3.11c}$$

by use of the jth conjugate complex eigenvalues λ_j of the proportional damped system, is identic to the free frequency ω_{0j} of the corresponding system without damping. With the eigenvalue λ_j , Eq.(3.11b), the relationship (3.11a) proves that $\mathbf{U}_j$ is a eigenvector of the proportional damped system (3.9).

In the considered case the equations of motion (1.1) can be transformed in modal space – Eq. (3.12), with regard to (3.6), (3.7):

$$\mathbf{U}^T\mathbf{M}\mathbf{U}\,\ddot{\mathbf{y}}+\mathbf{U}^T\mathbf{D_p}\mathbf{U}\,\dot{\mathbf{y}}+\mathbf{U}^T\mathbf{K}\mathbf{U}\,\mathbf{y}=\mathbf{U}^T\mathbf{p}(t)\ , \tag{3.12}$$

In (3.12) the modal superposition of the original DOF is supposed by use of the "classical" modal matrix **U** of the undamped problem, see Eq.(3.4), :

$$\mathbf{V}=[\mathbf{U}_1\quad\mathbf{U}_2\quad\cdots\quad\mathbf{U}_n]\cdot\begin{bmatrix}y_1\\y_2\\\cdots\\y_n\end{bmatrix}=\mathbf{U}\cdot\mathbf{y} \tag{3.13}$$

In order to transform the state space form of the equations of motion (2.12) we construct a (2n x2n) transformation matrix $\mathbf{Y}_\mathrm{U}$ by the mass normalized eigenvectors $\mathbf{U}_j$ (3.4), (3.5) in the form

$$\mathbf{Y}_\mathrm{U}=\begin{bmatrix}\mathbf{U}_1&\mathbf{0}&\mathbf{U}_2&\mathbf{0}&\cdots&\mathbf{U}_n&\mathbf{0}\\\mathbf{0}&\mathbf{U}_1&\mathbf{0}&\mathbf{U}_2&\cdots&\mathbf{0}&\mathbf{U}_n\end{bmatrix} \tag{3.14}$$

Eq. (2.12) can be transformed into n uncoupled real SDOFS block equations by the aid of $\mathbf{Y}_\mathrm{U}$, with regard to Eqs. (3.6), (3.7):

$$\underbrace{\mathbf{Y}_U^{\mathrm{T}} \cdot \begin{bmatrix} \mathbf{M} & \\ & -\mathbf{K} \end{bmatrix} \cdot \mathbf{Y}_{\mathrm{U}}}_{\begin{bmatrix} 1 & & & & \\ & -\omega_1^2 & & & \\ & & \cdots & & \\ & & & 1 & \\ & & & & -\omega_n^2 \end{bmatrix}} \cdot \begin{bmatrix} \dot{x}_1 \\ \dot{y}_1 \\ \cdots \\ \dot{x}_n \\ \dot{y}_n \end{bmatrix} + \underbrace{\mathbf{Y}_U^{\mathrm{T}} \cdot \begin{bmatrix} \mathbf{D_p} & \mathbf{K} \\ \mathbf{K} & \end{bmatrix} \cdot \mathbf{Y}_{\mathrm{U}}}_{\begin{bmatrix} 2\eta_1\,\omega_1 & \omega_1^2 & & & \\ -\omega_1^2 & 0 & & & \\ & & \cdots & & \\ & & & 2\eta_n\,\omega_n & \omega_n^2 \\ & & & \omega_n^2 & 0 \end{bmatrix}} \cdot \begin{bmatrix} x_1 \\ y_1 \\ \cdots \\ x_n \\ y_n \end{bmatrix} = \underbrace{\mathbf{Y}_U^{\mathrm{T}} \cdot \begin{bmatrix} \mathbf{p} \end{bmatrix}}_{\begin{bmatrix} g_1 \\ 0 \\ \cdots \\ g_n \\ 0 \end{bmatrix}} \tag{3.15}$$

where the modal velocity is

$$x_j = \dot{y}_j \tag{3.16}$$

Equation (3.15) is another form of the modal decomposition (3.12). Note the difference of Eq. (3.16) from Eq.(2.30a) in the general case of non-proportional damping, see further Eq. (3.17).

We compare now the transformed equations (3.15) for the case of proportional damping with the general form (2.23) where $\mathbf{D} = \mathbf{D}_p$. For both of the compared SDOFS block equations to be identic, it is evidently that each "load" term h_j from (2.23) must be equal to zero, see (2.29b):

$$h = \frac{\omega^2}{\sqrt{1-\eta^2}} \left\{ Z_1 \mathbf{\Phi_r^T} - Z_2\, \mathbf{\Phi_i^T} \right\} \mathbf{p} = 0 \tag{3.17}$$

All terms in Eq. (3.17) exclusive of $\mathbf{p}$ belong to the considered j[th] eigenmode. Thus, with regard to (2.11a), (2.5)

$$\left(Z_1 \mathbf{\Phi_r^T} - Z_2 \mathbf{\Phi_i^T}\right) = \mathbf{0}$$

$$\rightarrow \quad \frac{\Phi_{i(k)}}{\Phi_{r(k)}} = \frac{Z_1}{Z_2} = \frac{\sqrt{\sqrt{1-\eta^2}+(1-\eta^2)}}{\sqrt{\sqrt{1-\eta^2}-(1-\eta^2)}} = \frac{\eta}{1-\sqrt{1-\eta^2}} = const. \tag{3.18}$$

for all k[th] DOF of the jth eigenmode pair $(\mathbf{\Phi_r} \pm i\mathbf{\Phi_i})$ with corresponding eigenvalue $(\lambda_r \pm i\lambda_i)$. Equation (3.18) proves the statement of a constant phase lag/lead, see Eq. (1.10), i.e. in the case of proportionally damped system each free vibration is a synchronous motion of all DOF.

3.2 The transformation matrix Y

The modal equations (3.15) demonstrate, that for the investigated case of proportional damping the modal transformation matrix $\mathbf{Y}_{\mathrm{U}}$, Eq. (3.14), must be identical to the matrix $\mathbf{Y}$, Eq. (2.23), (2.24), derived for the case $\mathbf{D} = \mathbf{D}_p$. By comparing the two columns of $\mathbf{Y}$, see Eq. (2.27b,c), to the corresponding zero-columns of $\mathbf{Y}_{\mathrm{U}}$, it follows

$$\mathbf{Y_y^W} = \frac{1}{\sqrt{1-\eta^2}} (\omega^2 Z_1 \mathbf{\Phi_r} - \omega^2 Z_2 \mathbf{\Phi_i}) = \mathbf{0} \tag{3.19a}$$

$$\mathbf{Y_x^V} = \frac{1}{\sqrt{1-\eta^2}} (-Z_1 \mathbf{\Phi_r} + Z_2 \mathbf{\Phi_i}) = \mathbf{0} \tag{3.19b}$$

$$\rightarrow \quad -Z_1\boldsymbol{\Phi}_\mathbf{r} + Z_2\boldsymbol{\Phi}_\mathbf{i} = \mathbf{0} \tag{3.20}$$

The relationship (3.20) leads again to $\frac{\Phi_{i(k)}}{\Phi_{r(k)}} = \frac{Z_1}{Z_2} = const.$ (Eq.(3.18)) for all k[th] DOF of the considered eigenmode.

Thus the modal transformation matrix **Y**, see Eq. (2.25) and (3.14), has in this case the form

$$\mathbf{Y} = \begin{bmatrix} \cdots & \mathbf{Y}_\mathbf{x}^{(j)\mathbf{W}} & \mathbf{0} & \cdots \\ \cdots & \mathbf{0} & \mathbf{Y}_\mathbf{y}^{(j)\mathbf{V}} & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \mathbf{U}_j & \mathbf{0} & \cdots \\ \cdots & \mathbf{0} & \mathbf{U}_j & \cdots \end{bmatrix} \tag{3.21}$$

4 NUMERICAL EXAMPLE

4.1 Structural system, stiffness and geometry data

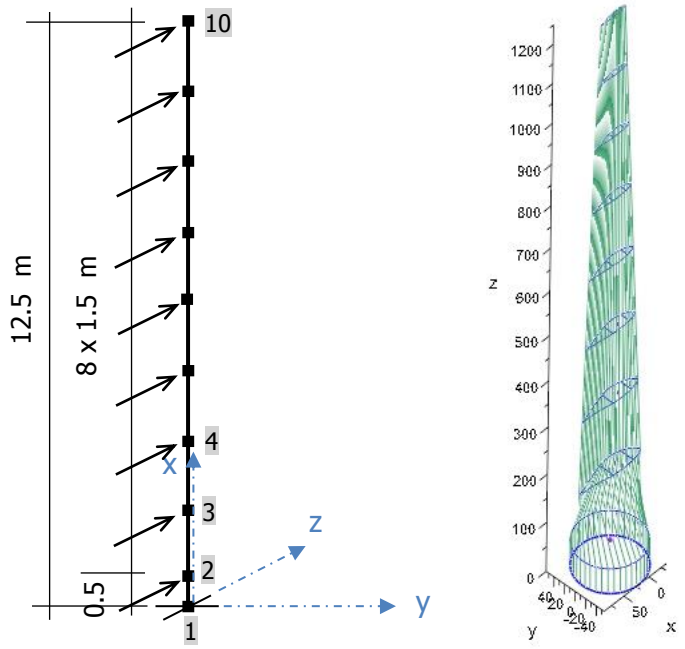


Fig. 1 Rotor blade beam model subjected to wind loads

The stiffness data of the blade cross sections have been calculated in [14]. The generic aerodynamic blade geometry has been derived from real blade data. Below are given for instance the stiffness data, referred to the origin of the coordinate system of the cross section, at the distance of 2.0 m from the blade root – see Fig. 2:

Center of mass F (0.124, -0.0119) [m] distributed mass 73.835 [kg/m]
EA = 947410000.0 [N] (axial stiffness)
EA_y = 86101990.0 [Nm] EA_z = -4408355.0 [Nm]
EA_{yy} = 48655550.0 [Nm^2] EA_{zz} = 16441220.0 [Nm^2]
EA_{yz} = 281046.0 [Nm^2] GI_T = 6500099.0 [Nm^2] (torsional stiffness)

The finite element solution is based on the numerical integration of the system of differential equations for the Bernoulli-beam. The reference axis of the beam model coincides with the centre of the circular-section at the root – it is the real rotational axis of the rotor blade. Thereby the differential equations and all cross section stiffness data are refered to this axis, accounting for the eccentric mass application.

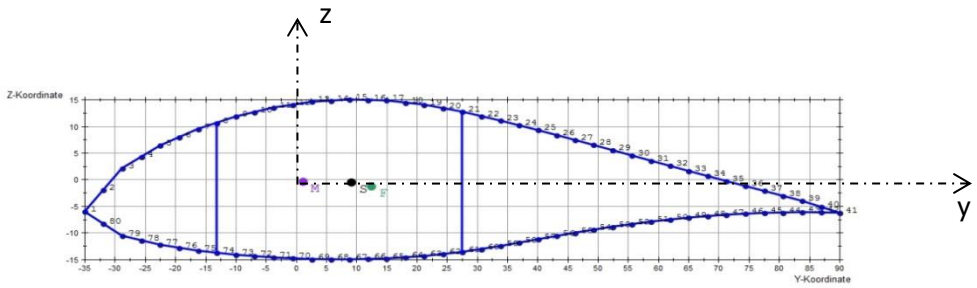


Fig. 2 Rotor blade sections at 2.0 m – thin wall cross section model

4.2 Wind loads

The wind loads are calculated according to the formula for the aerodynamic lift force per unit length of an aerofoil, see [13] p.59:

$$L = \frac{1}{2}\rho \cdot c(r) \cdot W^2 \cdot C_L \tag{4.1}$$

where:
- W : air velocity relative to the aerofoil
- ρ : air density = 1.225 [kg/m^3]
- $c(r)$: chord of the aerofoil
- C_L : lift coefficient $C_L = 2\pi\,\alpha = 2\pi\left(\frac{\pi}{180}6.0\right) = 0.658$, the flow angle α is assumed to be 6.0 [deg]

The air velocity W is the vector sum of the rotational speed Ω (with assumed 60 rpm) and the wind speed u, incident on the aerofoil in accordance with the Betz-theory:

$$W = \sqrt{(\Omega\, r)^2 + \left(\frac{2}{3}u\right)^2} \qquad \text{where } \Omega = \left(\frac{60}{30}\pi\right) \text{ in [rad/s]} \tag{4.2}$$

The wind speed function is assumed to be

$$u(t) = 16.0 + 8.0\, sin(2\pi f t), \quad \text{where} \quad f = 1 \text{ [Hz]} \tag{4.3}$$

The resulting wind thrust loads per unit length along the x-axis of the rotor blade are given below. In the structural model the wind thrust loads are acting as summarized nodal forces.

r [m]	c [m]	wind thrust F(t) [N/m]
0	1.1	0
0.5	1.1	$0.440881\,(5.33333\sin(6.28319\,t)+10.6667)^2+4.35132$
2.0	1.25	$0.501001\,(5.33333\sin(6.28319\,t)+10.6667)^2+79.115$
3.5	1.15	$0.460921\,(5.33333\sin(6.28319\,t)+10.6667)^2+222.906$
5.0	1.05	$0.420841\,(5.33333\sin(6.28319\,t)+10.6667)^2+415.354$
6.5	0.95	$0.380761\,(5.33333\sin(6.28319\,t)+10.6667)^2+635.095$
8.0	0.85	$0.340681\,(5.33333\sin(6.28319\,t)+10.6667)^2+860.771$
9.5	0.75	$0.300601\,(5.33333\sin(6.28319\,t)+10.6667)^2+1071.02$
11.0	0.65	$0.260521\,(5.33333\sin(6.28319\,t)+10.6667)^2+1244.48$
12.5	0.55	$0.220441\,(5.33333\sin(6.28319\,t)+10.6667)^2+1359.79$

The wind thrust functions F(t) are acting on the rotor blade as shown in Fig. 3 for 10 sec.

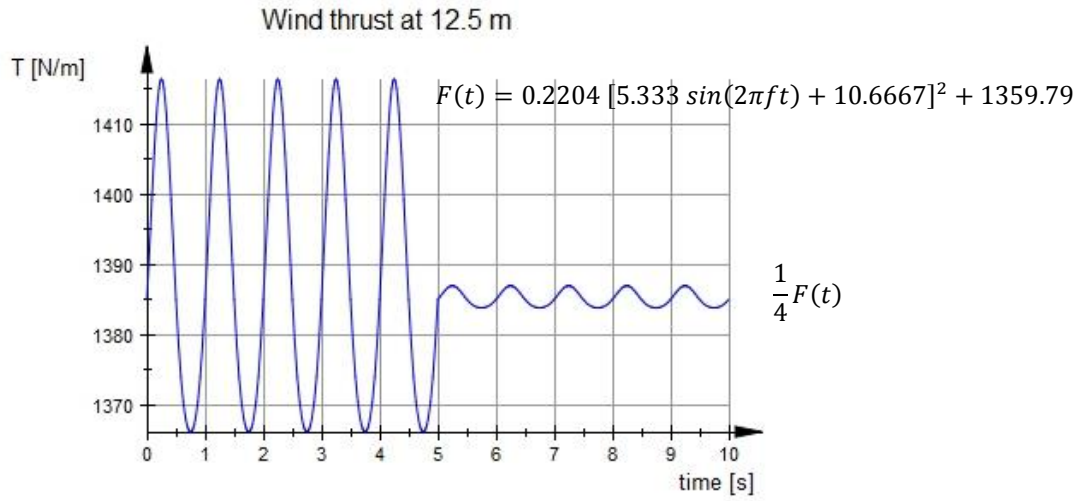


Fig. 3 Wind thrust function at 12.5 m

4.3 Relationships and data for the damping approach

Starting point of the computation are the equations of motion

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & -\mathbf{K} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{W}} \\ \dot{\mathbf{V}} \end{bmatrix} + \begin{bmatrix} \mathbf{D} & \mathbf{K} \\ \mathbf{K} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \mathbf{V} \end{bmatrix} = \begin{bmatrix} \mathbf{P}(t) \\ \mathbf{0} \end{bmatrix} \tag{4.4}$$

where $\mathbf{P}(t)$ is the nodal force vector, representing the wind thrust according to Sec. 4.2.

The system equations (4.4) will be solved applying the proposed modal analysis method in Sec. 2 for two cases: non-proportional and proportional damping.

The lowest four free-vibration frequencies and associated periods for the undamped system are calculated to

$$\begin{aligned} f_1 &= 2.643\ [s^{-1}] & T_1 &= 0.378\ [s] \\ f_2 &= 4.622\ [s^{-1}] & T_2 &= 0.216\ [s] \\ f_3 &= 7.942\ [s^{-1}] & T_3 &= 0.126\ [s] \\ f_4 &= 16.650\ [s^{-1}] & T_4 &= 0.060\ [s] \end{aligned} \tag{4.5}$$

Assuming stiffness proportional damping in accordance with Eq.(3.1), the damping system matrix is

$$\mathbf{D_p} = \beta\ \mathbf{K} \tag{4.6}$$

With an assumed damping ratio $\eta = 0.008$, see [13] p.249, for the first natural period T_1, we obtain with regard to Eq.(3.8a,b)

$$\beta = \frac{2\eta}{\omega_1} = \frac{\eta\, T_1}{\pi} = 0.000964[s] \tag{4.7}$$

The non-proportional symmetric damping matrix $\mathbf{D_{np}}$ is build adding to the $\mathbf{D_p}$-matrix a new matrix $\mathbf{D_a}$, which represents the aerodynamic damping. The formulation is based on a simple expression for the aerodynamic damping coefficient per unit length $c(r)$, given in [13], p. 247:

$$c(r) = \frac{1}{2}\rho \cdot\ \Omega r\ \cdot c(r) \cdot \frac{dC_L}{d\alpha} \quad \left[\frac{kg}{s}\frac{1}{m}\right], \quad \text{where} \quad \frac{dC_L}{d\alpha} = 2\pi \tag{4.8}$$

With Eq. (4.1), (4.2), the corresponding damping coefficients along the x-axis of the rotor blade are calculated to

r [m]	c [m]	$c(r)$ [kg/s.m]
0	1.1	0
0.5	1.1	13.2993
2.0	1.25	60.4513
3.5	1.15	97.3266
5.0	1.05	126.948
6.5	0.95	149.315
8.0	0.85	164.428
9.5	0.75	172.286
11.0	0.65	172.891
12.5	0.55	166.241

The coefficients $c(r)$, which represent the aerodynamic damping, are active for vibration in z-direction of the cross-section coordinate system, see Fig. 2. The associate symmetric damping matrix for the Bernoulli-beam element is derived by analogy with the method used to derive the finite element mass matrix, see [15]. Finally the symmetric system damping matrix, $\mathbf{D_{np}}$, is assembled in a finite-element manner, including structural (proportional) and aerodynamic damping:

$$\mathbf{D_{np}} = \mathbf{D_p} + \mathbf{D_a} \tag{4.9}$$

4.4 Non-proportional damped system

We use here the matrix $\mathbf{D_{np}}$ – Eq.(4.9). The vector of the first ten complex conjugate eigenvalue pairs of the matrix $\mathbf{A} = \mathbf{M_G^{-1}} \cdot \mathbf{K_G}$, see Eq.(2.13), is

$$\begin{pmatrix} -0.40981+29.0336\,i \\ -0.40981-29.0336\,i \\ -5.43041+105.219\,i \\ -5.43041-105.219\,i \\ -5.56181+15.7652\,i \\ -5.56181-15.7652\,i \\ -6.33469+49.2454\,i \\ -6.33469-49.2454\,i \\ -9.53814+104.542\,i \\ -9.53814-104.542\,i \\ -20.7608+185.185\,i \\ -20.7608-185.185\,i \\ -22.1068+207.402\,i \\ -22.1068-207.402\,i \\ -27.8796+238.91\,i \\ -27.8796-238.91\,i \\ -45.8047+292.379\,i \\ -45.8047-292.379\,i \\ -63.5216+353.962\,i \\ -63.5216-353.962\,i \end{pmatrix} \tag{4.10}$$

The number of modes considered in the modal transformation is limited to the first four eigenvector pairs – this are in ascending order the #3, 1, 4, 5 of the vector in (4.10). The structural system has in Fig. 1 has 54 DOF. The corresponding (108x8) modal matrix $\mathbf{\Phi_G}$ with mass normalized eigenvectors – Eq.(2.18), is computed to (only the first ten rows are printed)

$$\begin{pmatrix} 1.57\,10^{-5}+1.1\,10^{-7}\,i & 1.57\,10^{-5}-1.1\,10^{-7}\,i & -6.12\,10^{-5}-1.46\,10^{-6}\,i & -6.12\,10^{-5}+1.46\,10^{-6}\,i & -5.28\,10^{-8}-1.12\,10^{-7}\,i & -5.28\,10^{-8}+1.12\,10^{-7}\,i & -2.41\,10^{-8}+1.31\,10^{-7}\,i & -2.41\,10^{-8}-1.31\,10^{-7}\,i \\ 1.98\,10^{-5}+1.39\,10^{-7}\,i & 1.98\,10^{-5}-1.39\,10^{-7}\,i & -7.82\,10^{-5}-1.85\,10^{-6}\,i & -7.82\,10^{-5}+1.85\,10^{-6}\,i & -6.65\,10^{-8}-1.41\,10^{-7}\,i & -6.65\,10^{-8}+1.41\,10^{-7}\,i & -3.03\,10^{-8}+1.67\,10^{-7}\,i & -3.03\,10^{-8}-1.67\,10^{-7}\,i \\ 4.34\,10^{-7}+1.33\,10^{-7}\,i & 4.34\,10^{-7}-1.33\,10^{-7}\,i & -1.98\,10^{-6}-6.65\,10^{-6}\,i & -1.98\,10^{-6}+6.65\,10^{-6}\,i & 6.1\,10^{-6}+1.75\,10^{-6}\,i & 6.1\,10^{-6}-1.75\,10^{-6}\,i & -1.99\,10^{-5}-1.67\,10^{-6}\,i & -1.99\,10^{-5}+1.67\,10^{-6}\,i \\ 1.33\,10^{-7}+5.13\,10^{-8}\,i & 1.33\,10^{-7}-5.13\,10^{-8}\,i & -3.94\,10^{-6}-5.33\,10^{-6}\,i & -3.94\,10^{-6}+5.33\,10^{-6}\,i & 7.47\,10^{-7}+8.79\,10^{-7}\,i & 7.47\,10^{-7}-8.79\,10^{-7}\,i & -8.12\,10^{-6}-1.22\,10^{-6}\,i & -8.12\,10^{-6}+1.22\,10^{-6}\,i \\ -1.72\,10^{-6}-5.24\,10^{-7}\,i & -1.72\,10^{-6}+5.24\,10^{-7}\,i & 7.74\,10^{-6}+2.58\,10^{-5}\,i & 7.74\,10^{-6}-2.58\,10^{-5}\,i & -2.42\,10^{-5}-6.9\,10^{-6}\,i & -2.42\,10^{-5}+6.9\,10^{-6}\,i & 7.8\,10^{-5}+6.46\,10^{-6}\,i & 7.8\,10^{-5}-6.46\,10^{-6}\,i \\ 7.85\,10^{-5}+5.52\,10^{-7}\,i & 7.85\,10^{-5}-5.52\,10^{-7}\,i & -3.05\,10^{-4}-7.23\,10^{-6}\,i & -3.05\,10^{-4}+7.23\,10^{-6}\,i & -2.63\,10^{-7}-5.59\,10^{-7}\,i & -2.63\,10^{-7}+5.59\,10^{-7}\,i & -1.31\,10^{-7}+6.56\,10^{-7}\,i & -1.31\,10^{-7}-6.56\,10^{-7}\,i \\ 2.1\,10^{-4}+1.27\,10^{-6}\,i & 2.1\,10^{-4}-1.27\,10^{-6}\,i & -6.7\,10^{-4}-8.21\,10^{-6}\,i & -6.7\,10^{-4}+8.21\,10^{-6}\,i & -1.1\,10^{-5}-4.28\,10^{-6}\,i & -1.1\,10^{-5}+4.28\,10^{-6}\,i & 2.76\,10^{-5}+3.53\,10^{-6}\,i & 2.76\,10^{-5}-3.53\,10^{-6}\,i \\ 0.0018+1.22\,10^{-5}\,i & 0.0018-1.22\,10^{-5}\,i & -0.00609-1.26\,10^{-4}\,i & -0.00609+1.26\,10^{-4}\,i & -2.76\,10^{-5}-1.88\,10^{-5}\,i & -2.76\,10^{-5}+1.88\,10^{-5}\,i & 5.66\,10^{-5}+1.86\,10^{-5}\,i & 5.66\,10^{-5}-1.86\,10^{-5}\,i \\ 2.56\,10^{-5}+2.72\,10^{-5}\,i & 2.56\,10^{-5}-2.72\,10^{-5}\,i & -1.34\,10^{-4}-0.00113\,i & -1.34\,10^{-4}+0.00113\,i & 0.00133+3.66\,10^{-4}\,i & 0.00133-3.66\,10^{-4}\,i & -0.00386-2.82\,10^{-4}\,i & -0.00386+2.82\,10^{-4}\,i \\ 5.85\,10^{-6}+2.27\,10^{-6}\,i & 5.85\,10^{-6}-2.27\,10^{-6}\,i & -1.71\,10^{-4}-2.31\,10^{-4}\,i & -1.71\,10^{-4}+2.31\,10^{-4}\,i & 3.3\,10^{-5}+3.88\,10^{-5}\,i & 3.3\,10^{-5}-3.88\,10^{-5}\,i & -3.57\,10^{-4}-5.3\,10^{-5}\,i & -3.57\,10^{-4}+5.3\,10^{-5}\,i \end{pmatrix} \tag{4.11}$$

The matrix $\mathbf{\Psi}^{-1}$ is now calculated in the case of four involved eigenmodes according to Eq. (2.24):

$$\mathbf{\Psi}^{-1} = \begin{bmatrix} \left(\varphi^{(1)}\right)^{-1} & & & \\ & \left(\varphi^{(2)}\right)^{-1} & & \\ & & \left(\varphi^{(3)}\right)^{-1} & \\ & & & \left(\varphi^{(4)}\right)^{-1} \end{bmatrix} =$$

$$\begin{pmatrix}
-0.717706-0.122888\,i & -2.05439-11.9983\,i & 0 & 0 & 0 & 0 & 0 & 0\\
-0.717706+0.122888\,i & -2.05439+11.9983\,i & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & -0.707124-0.0049903\,i & -0.1449-20.5324\,i & 0 & 0 & 0 & 0\\
0 & 0 & -0.707124+0.0049903\,i & -0.1449+20.5324\,i & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & -0.708562-0.045386\,i & -2.25347-35.1809\,i & 0 & 0\\
0 & 0 & 0 & 0 & -0.708562+0.045386\,i & -2.25347+35.1809\,i & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & -0.707841-0.032224\,i & -3.38274-74.3061\,i\\
0 & 0 & 0 & 0 & 0 & 0 & -0.707841+0.032224\,i & -3.38274+74.3061\,i
\end{pmatrix} \tag{4.12}$$

Finally the (108x8) real transformation matrix **Y** is computed according to (2.24) – here only the first ten rows:

$$\mathbf{Y} = \begin{pmatrix}
0 & -0.000002 & -0.000022 & 0 & 0 & 0.000009 & 0 & -0.001258\\
0 & -0.000003 & -0.000028 & 0 & 0 & 0.000012 & 0 & -0.001609\\
-0.000008 & 0.000017 & -0.000001 & 0.000005 & 0.000028 & -0.000027 & -0.000067 & 0.000011\\
-0.000001 & 0.000018 & 0 & 0.000002 & 0.000011 & -0.000049 & -0.000053 & 0.000259\\
0.000033 & -0.000066 & 0.000002 & -0.000021 & -0.00011 & 0.000103 & 0.000259 & -0.000027\\
0 & -0.000012 & -0.000111 & 0 & 0 & 0.000047 & 0.000001 & -0.00628\\
0.000015 & -0.000058 & -0.000297 & -0.000009 & -0.000039 & 0.000124 & 0.00008 & -0.013754\\
0.000035 & -0.000337 & -0.00255 & -0.00002 & -0.000079 & 0.001052 & 0.000202 & -0.125274\\
-0.001819 & 0.003317 & -0.000036 & 0.001111 & 0.005438 & -0.002459 & -0.011388 & -0.001937\\
-0.000038 & 0.000796 & -0.000008 & 0.000091 & 0.000501 & -0.002123 & -0.002276 & 0.011014
\end{pmatrix} \tag{4.13}$$

After the modal transformation in regard to (2.23) the time-dependent "load" vector (here for the time 0…5 sec) is calculated to be, see also Fig. 3,

$$\begin{bmatrix} g_1(t) \\ h_1(t) \\ \cdots \\ g_3(t) \\ h_3(t) \end{bmatrix} = \mathbf{Y}^{\mathrm{T}} \cdot \begin{bmatrix} \mathbf{p} \end{bmatrix} = \begin{pmatrix}
-0.1878\,(5.333\,\sin(6.283\,t)+10.67)^2-690.3\\
0.02234\,(5.333\,\sin(6.283\,t)+10.67)^2-165.7\\
-0.003661\,(5.333\,\sin(6.283\,t)+10.67)^2-14.86\\
0.07117\,(5.333\,\sin(6.283\,t)+10.67)^2+227.9\\
0.04344\,(5.333\,\sin(6.283\,t)+10.67)^2-212.1\\
1.025\,(5.333\,\sin(6.283\,t)+10.67)^2+3882.0\\
-0.06672\,(5.333\,\sin(6.283\,t)+10.67)^2-114.6\\
1871.0-0.01053\,(5.333\,\sin(6.283\,t)+10.67)^2
\end{pmatrix} \tag{4.14}$$

The resultant four uncoupled SDOFS block equations from type of Eq. (2.23), prepared in the form (2.30a,b), are solved by step-by-step integration:

$$\begin{bmatrix} 1 & & & & \\ & -\omega_1^2 & & & \\ & & \cdots & & \\ & & & 1 & \\ & & & & -\omega_n^2 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} \dot{x}_1 \\ \dot{y}_1 \\ \cdots \\ \dot{x}_n \\ \dot{y}_n \end{bmatrix}}_{\dot{\mathbf{x}}} + \begin{bmatrix} 2\eta_1\,\omega_1 & \omega_1^2 & & & \\ -\omega_1^2 & 0 & & & \\ & & \cdots & & \\ & & & 2\eta_n\,\omega_n & \omega_n^2 \\ & & & \omega_n^2 & 0 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} x_1 \\ y_1 \\ \cdots \\ x_n \\ y_n \end{bmatrix}}_{\mathbf{x}} = \begin{bmatrix} g_1 \\ h_1 \\ \cdots \\ g_n \\ h_n \end{bmatrix}, \qquad (n=4) \tag{4.15}$$

where $[\omega_i] =$ (16.7175 29.0365 49.6511 104.976)

$[\eta_i] =$ (0.332693 0.0141136 0.127584 0.0908603) (4.16a,b)

The effect of the implied additional aerodynamic damping results evidently in the large damping ratio $\eta_i = 0.33269$ for the first free vibration.
The vibration-response has been determined in the time 0…10 s, the time step length for the applied Newmark integration method is 0.005 s.

The time response of the modal coordinates $y_j(t)$, $(j=1,...4)$, are shown in the following figure 4 for the time 0…5 sec:

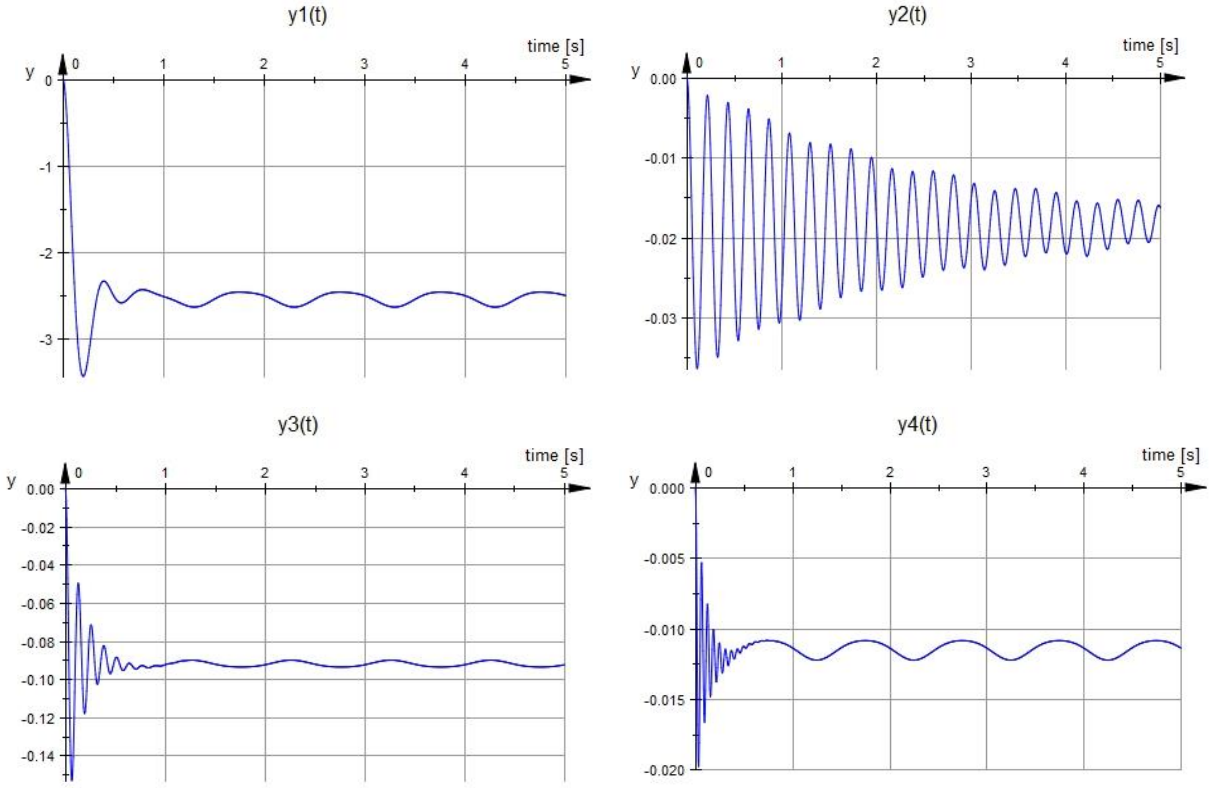


Fig. 4 Time response of the modal coordinates $y_j(t)$ for the case "non-proportional damping"

By a back transformation according to Eq. (2.24) the total response $\mathbf{V}(t)$ is obtained - see Figs. 5a-c:

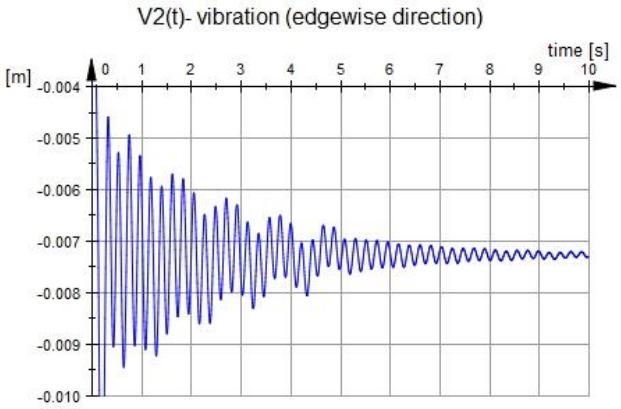


Fig.5a Total vibration $u_2(t)\ [m]$ at the rotor blade tip (y-direction at node #10)

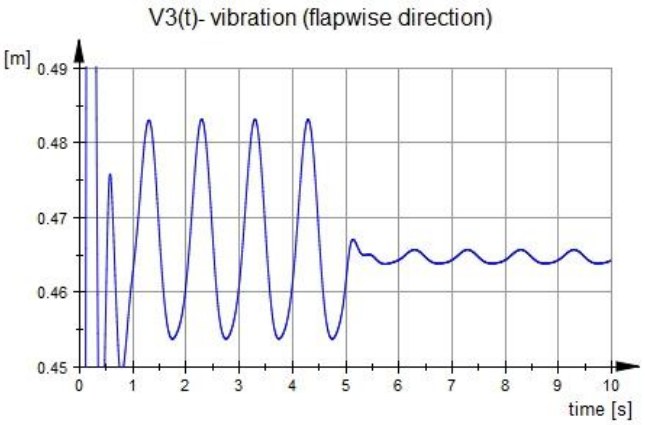


Fig.5b Total vibration $u_3(t)\ [m]$ at the rotor blade tip (z-direction at node #10)

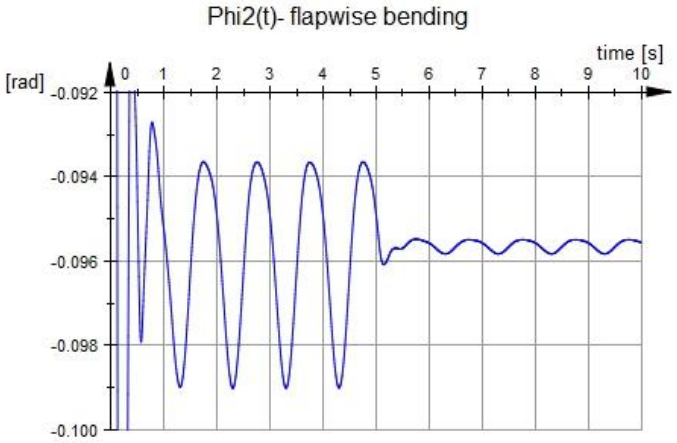


Fig.5c Total rotation $\varphi_2(t)\ [rad]$ at the rotor blade tip (y-axis at node #10)

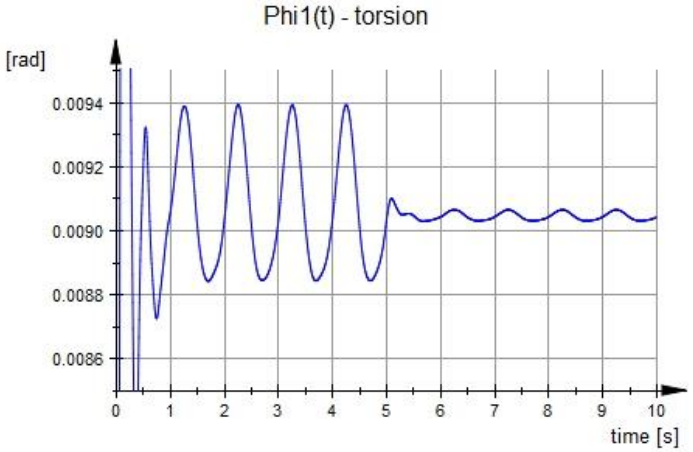


Fig.5d Total torsional rotation $\varphi_1(t)\ [rad]$ at the rotor blade tip (x-axis at node #10)

The vibration responses, computed by direct step-by-step integration of the equations (4.4), are practically identical to those in Fig. 5a-d.

4.5 Proportional damped system

In this case we use the derived symmetric damping matrix $\mathbf{D_p}$ – Eq.(4.6), (4.7). The first ten lowest complex conjugate eigenvalue pairs, resulting from Eq. (2.13), are now:

$$\begin{pmatrix} -0.132832+16.6035\,i \\ -0.132832-16.6035\,i \\ -0.406268+29.0352\,i \\ -0.406268-29.0352\,i \\ -1.19966+49.8844\,i \\ -1.19966-49.8844\,i \\ -5.27314+104.483\,i \\ -5.27314-104.483\,i \\ -5.39463+105.676\,i \\ -5.39463-105.676\,i \\ -16.7361+185.622\,i \\ -16.7361-185.622\,i \\ -20.9751+207.591\,i \\ -20.9751-207.591\,i \\ -27.8753+238.91\,i \\ -27.8753-238.91\,i \\ -42.2056+292.945\,i \\ -42.2056-292.945\,i \\ -62.3277+354.226\,i \\ -62.3277-354.226\,i \end{pmatrix} \tag{4.17}$$

The corresponding (108x8) $\mathbf{\Phi_G}$ modal matrix – Eq. (2.18), comprises the first four mass normalized complex conjugate eigenvector pairs. In order to verify the derived relationship $\frac{\Phi_{i(k)}}{\Phi_{r(k)}} = \frac{\eta}{1-\sqrt{1-\eta^2}} = const.$, see (3.14), we compute this ratio for all components of the involved $(\mathbf{\Phi_r} \pm i\mathbf{\Phi_i})^{(j)}$ $(j = 1, \ldots 4)$ eigenvectors (for instance the first ten rows only):

$$\begin{pmatrix} 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \\ 250.0 & 143.0 & 83.2 & 39.7 \end{pmatrix} \leftrightarrow \begin{pmatrix} 250.0 \\ 143.0 \\ 83.2 \\ 39.7 \end{pmatrix} = \frac{\eta_j}{1-\sqrt{1-\eta_j^2}} \quad , \quad (j = 1, \ldots 4) \tag{4.18}$$

The corresponding damping ratios η_j , see Eq. (4.21b), are computed in accordance with Eq. (2.5).

The next step is the computation of the matrix $\mathbf{\Psi}^{-1}$, Eq. (2.24). The (108x8) real transformation matrix **Y**, computed in regard with Eq. (2.24), (2.27), has now the form of (3.21):

$$\mathbf{Y} = \begin{pmatrix} 0 & 0 & -0.000022 & 0 & 0 & 0 & 0.000053 & 0 \\ 0 & 0 & -0.000028 & 0 & 0 & 0 & 0.000068 & 0 \\ -0.000009 & 0 & -0.000001 & 0 & 0.000028 & 0 & -0.00005 & 0 \\ -0.000002 & 0 & 0 & 0 & 0.000011 & 0 & -0.000038 & 0 \\ 0.000035 & 0 & 0.000003 & 0 & -0.00011 & 0 & 0.000195 & 0 \\ 0.000001 & 0 & -0.000111 & 0 & 0 & 0 & 0.000265 & 0 \\ 0.000017 & 0 & -0.000297 & 0 & -0.000039 & 0 & 0.000642 & 0 \\ 0.000053 & 0 & -0.002549 & 0 & -0.000082 & 0 & 0.005423 & 0 \\ -0.001935 & 0 & -0.000051 & 0 & 0.005447 & 0 & -0.008771 & 0 \\ -0.000069 & 0 & -0.00001 & 0 & 0.000499 & 0 & -0.001632 & 0 \end{pmatrix} \tag{4.19}$$

In (4.19) are printed again only the first ten rows of **Y.**

The time-dependent "load" vector in the general modal transformed equations (2.23) is now (here for the time 0…5 sec) - see also Eq. (3.15) and (4.12):

$$\begin{bmatrix} g_1(t) \\ h_1(t) \\ \cdots \\ g_4(t) \\ h_4(t) \end{bmatrix} = \begin{pmatrix} -0.1872\,(5.333\sin(6.283\,t)+10.67)^2 - 678.6 \\ 0 \\ -0.004824\,(5.333\sin(6.283\,t)+10.67)^2 - 18.84 \\ 0 \\ 0.03913\,(5.333\sin(6.283\,t)+10.67)^2 - 227.6 \\ 0 \\ -0.052\,(5.333\sin(6.283\,t)+10.67)^2 - 94.4 \\ 0 \end{pmatrix} \tag{4.20}$$

Eq. (4.20) implies $x_j = \dot{y}_j$, see Eq.(3.15), (3.16), contrary to the general case Eq.(2.30a).

In the resultant four uncoupled SDOFS block equations, see (4.15), the free frequencies and the modal damping ratios are resp.

$[\omega_i] =$ (16.604 29.0381 49.8988 104.616)

$[\eta_i] =$ (0.008 0.0139909 0.0240418 0.0504049) (4.21a,b)

After step-by-step integration of the four modal equations (4.15), the time series of the modal coordinates $x_j(t)$, $y_j(t)$, $(j=1,...4)$, are obtained – Fig. 6:

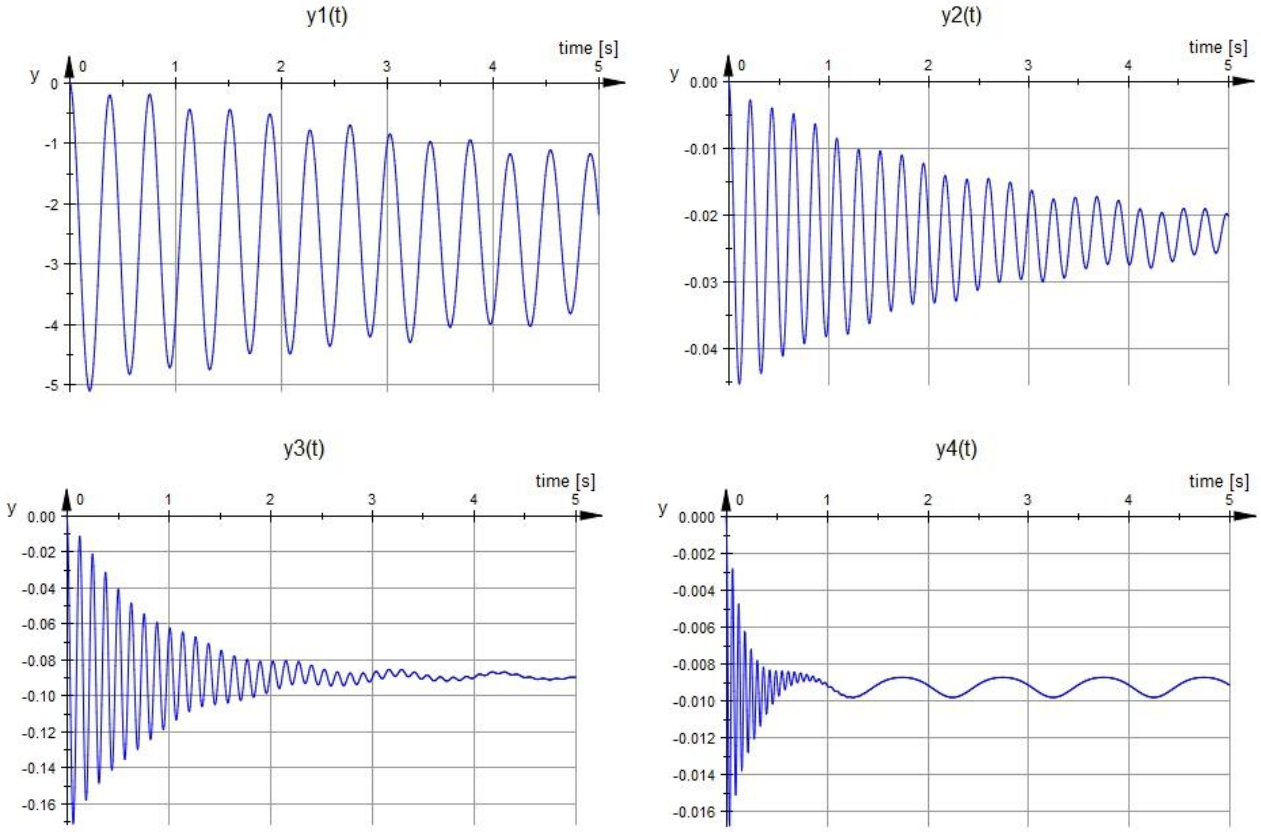


Fig. 6 Time response of the modal coordinates $y_j(t)$ for the case "proportional damping"

The total responses $\mathbf{V}(t)$ are computed by a back transformation according to Eq. (2.24) – see Figs. 7a-d:

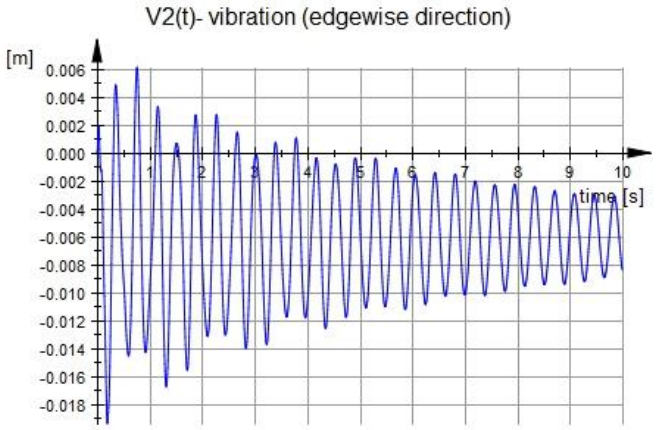


Fig 7a - Total vibration $u_2(t)\ [m]$ at the rotor blade tip (y-direction at node #10)

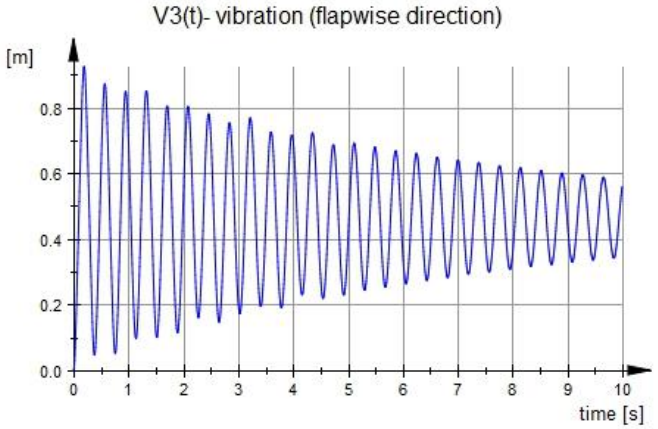


Fig 7b - Total vibration $u_3(t)\ [m]$ at the rotor blade tip (z-direction at node #10)

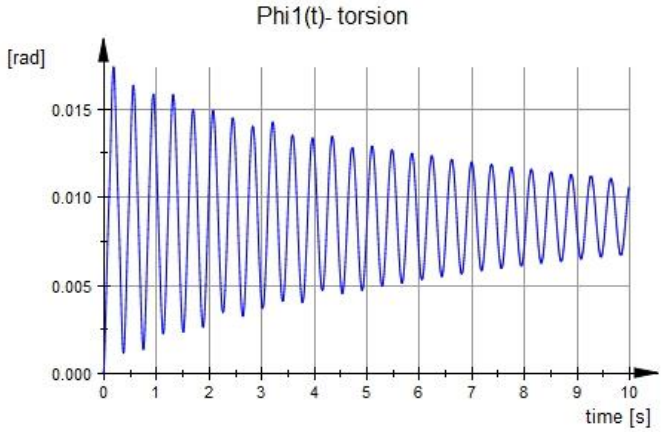


Fig 7c - Total torsion $\varphi_1(t)\ [rad]$ at the rotor blade tip (x-direction at node #10)

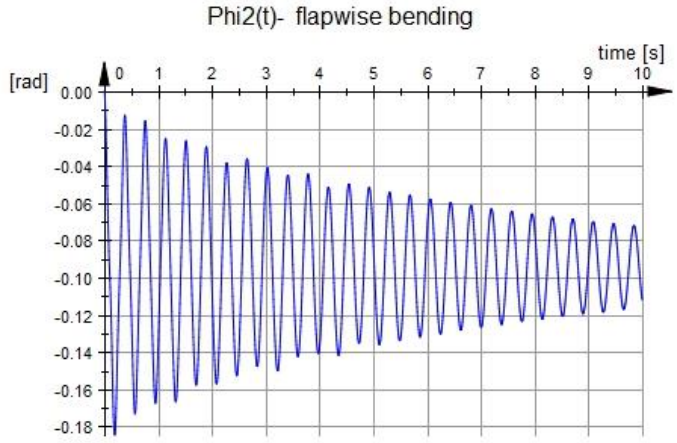


Fig 7d - Total rotation $\varphi_2(t)\ [rad]$ at the rotor blade tip (y-axis at node #10)

5 CONCLUSIONS

- A general modal decomposition method of MDOFS with non-proportional damping is briefly presented in Sec. 2. The procedure is based on the complex eigenvalue solution of a structural model with symmetric non-proportional damping matrix. By use of the right complex conjugate eigenvector pairs, normalized relative to the general mass matrix, a new real transformation matrix **Y**, see Eq. (2.24), (2.27), is developed analytically to perform a modal decomposition of the equations of motion in real arithmetic. The complex conjugate eigenpairs – eigenvalues and the corresponding eigenvectors – are to be computed first, at least for the lowest few modal shapes.
- The equations of motion are transformed into uncoupled SDOFS block equations. Employing only a few (k) eigenvector pairs in the **Y** -basis (k<<n) is leading – typical for a modal transformation procedure – with sufficient numerical accuracy to the total time response of all n DOF. The modal equations are numerically integrated and finally transformed back to the original DOF. In more details the method has been described in [3], and in [4] has been developed a similar method, based on the right and left eigenvector pairs.
- The application of the suggested method to the special case of proportional damped system is considered in details in Sec. 3. Employing a Rayleigh damping matrix, it has been shown that the modal transformation from Sec. 2 implies a ratio $\frac{\Phi_{i(k)}}{\Phi_{r(k)}} = const.$ for all kth DOF of each considered eigenmode $(\mathbf{\Phi_r} \pm i\mathbf{\Phi_i})$, i.e. the "constant phase" statement. This proves in an indirect manner that the free vibrations in the proportional damping case are synchronous. A simple formula for computing of the constant ratio has been also derived, expressing it through the associated modal damping ratio η.
- In Section 4 a numerical example – vibration of a rotor blade with 54 DOF - demonstrates the performance of the presented modal method for the two cases – non-proportional and proportional (Rayleigh) damping. In the first variant the damping matrix of the system contains a stiffness-proportional part and a simple approximated aerodynamic damping part. In the second variant the formula for the constant phase of the resonance modes is verified numerically.
- Real life applications of the proposed modal analysis method and possible numerical complications are discussed more widely in [4], [5]. The present paper studies some

known features of proportionally damped systems – the synchronous free vibrations – from a viewpoint of a new proposed modal analysis method.

REFERENCES

[1] E. Stanoev, H. Cramer. *Zur Boden-Bauwerk-Interaktion auf der Basis der modalen Analyse*. BKI 03/3, Universität der Bundeswehr München, München 2003, p. 135-146

[2] H. Cramer, E. Stanoev. Zur modalen Analyse bei gedämpften Mehrmassensystemen.

10. Dresdner Baustatik-Seminar: Neue Bauweisen – Trends in Statik und Dynamik, TU Dresden, Lehrstuhl für Statik, Dresden, 2006, p. 81-90

[3] H. Cramer, E. Stanoev. *Ein Verfahren zur modalen Analyse gedämpfter Systeme der Strukturmechanik*. Rostocker Berichte aus dem Institut für Bauingenieurwesen, Heft 19, Universität Rostock, Institut für Bauingenieurwesen, 2008, ISSN 1438-7638, p. 83-107

[4] E. Stanoev. *A modified modal analysis method for damped multi-degree-of-freedom-systems in structural mechanics*. Zeitschrift für angewandte Mathematik und Mechanik (ZAMM), 2013, 1 – 23 (2013), http://onlinelibrary.wiley.com/doi/10.1002/zamm.201300061/abstract

[5] E. Stanoev. *A modal analysis method for structural models with non-modal damping.* MS "Multibody system dynamics and modal reduction" in the frame of 11th World Congress on Computational Mechanics (WCCM XI), 5th European Conference on Computational Mechanics (ECCM V), 20-25 Juli 2014, Barcelona, ISBN: 978-84-942844-7-2, Tomo IV, p. 3034-3045

[6] Peters, H. J.; Tiso, P.; Goosen, J.F.L.; van Keulen, F. *Modifying resonance modes of dissipative structures using magnitude and phase information,* 11th World Congress on Computational Mechanics (WCCM XI), 5th European Conference on Computational Mechanics (ECCM V), 20-25 Juli 2014, Barcelona, ISBN: 978-84-942844-7-2, Tomo II, p. 489-500

[7] Géradin, M; Rixen, D. *Mechanical vibrations – theory and applications to structural dynamics*, John Wiley & Sons Ltd, 1997

[8] Meskouris, K. *Baudynamik – Modelle, Methoden, Praxisbeispiele*, Ernst & Sohn, 1999

[9] Chorpa, A.K. *Dynamics of Structures. Theory and Applications to Earthquake Engineering*, Pearson Prentice Hall, New Jersey, 2007

[10] M.-C. Kim, L.-W. Lee. *Eigenproblems for large structures with non-proportional damping.* Earthquake Engng. Struct. Dyn, 28, 157-172, 1999

[11] D. C. Sorensen. *Implicitly Restarted Arnoldi/Lanczos Methods for Large Scale Eigenvalue Calculations*. Dep. Comp. Appl. Math., Rice University, Houston, 1995

[12] R. Lehoucq, D. C. Sorensen. *Implicitly Restarted Lanczos Method*. (Section 4.5,7.6) : Z. Bai, J. Demmel, J. Dongarra, A. Ruhe and H. van der Vorst, editors, Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, SIAM, Philadelphia, 2000, 67-80, 166-184

[13] T. Burton, T.; Jenkins, N.; Sharpe D.; Bossanyi, E. *Wind Energy Handbook*, John Wiley & Sons, 2011, chapter 5.7, 5.8

[14] Nan Li. *Berechnung der Querschnittssteifigkeiten des Rotorblatts einer WEA durch ein FE-Verfahren für dünnwandige mehrzellige Profile mit Einsatz von MATLAB*, Studienarbeit, University of Rostock, Chair of Wind Energy Technology, 2015

[15] E. Stanoev. *Eine alternative FE-Formulierung der kinetischen Effekte beim räumlich belasteten Stab*. Rostocker Berichte aus dem Institut für Bauingenieurwesen, Heft 17, Universität Rostock, Institut für Bauingenieurwesen, 2007, ISSN 1438-7638, p.143-161

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A FORCE-BASED FORMULATION FOR THE ANALYSIS OF 3-DIMENSIONAL INELASTIC STRUCTURAL FRAMES

Theodoros N. Patsios[1] **and Konstantinos V. Spiliopoulos**[2]

[1] Ph.D. Candidate, ISAAR, School of Civil Engineering, National Technical University of Athens
9 Heroon Polytechniou, Zografou Campus, 15780, Athens, Greece
e-mail: tpatsios@central.ntua.gr, tpatsios@gmail.com

[2] Professor, ISAAR, School of Civil Engineering, National Technical University of Athens
9 Heroon Polytechniou, Zografou Campus, 15780, Athens, Greece
e-mail: kvspilio@central.ntua.gr

Keywords: Force Method, 3D Frames, Plastic Hinges, Non-Holonomic Plasticity, Pushover Analysis.

Abstract. *A force-based formulation for the step-by-step non-linear (elastic-plastic) analysis of three-dimensional (3D) structural frames is presented. It uses the redundant force and moment components as primary unknowns, and approximates the non-linearity problem in an incremental pattern. Using a simple linear transformation, the equilibrium matrices are quickly formed via a partial multiplication of a subset of matrices with dimensions (3x3). The convex yield function that describes the static admissibility condition of each zero-length plastic hinge is approximated with a linear convex polyhedron (manifold), whose hyper plane equations are automatically defined with the help of De Bruijn sequences. In this way, a number of complex force/moment interaction criteria may easily be defined that take into account shear and torsion. Discontinuities (e.g. articulations) are also accounted for. Out of the partial derivatives of these yield functions with respect to the stresses, the corresponding plastic deformations are computed, with the help of Lagrange multipliers. The formulation may be solved using any non-linear optimization algorithm that solves for linear constraints. Results are compared to those of the equivalent direct stiffness method and to those of the existing literature, proving the efficiency of the proposed formulation.*

1 INTRODUCTION

In this paper, an existing step−by−step, force−based method [1] is further extended to the analysis of inelastic 3D structural frames. Techniques to define equilibrium matrices for incremental analysis have been already presented in [1], extending what was first presented in [2] for the case of optimal plastic design; remedies to these techniques in order to function for 3D structures were presented in [3].

Herein, an alternate linear transformation that requires fewer computer operations is introduced. Furthermore, a fully automated technique for generating any type of linearized yield function for up to six interacting stress components is developed, that is based on the assumption of lumped plasticity (*plastic hinge approach*). By definition, these generalized yield functions are able to also cater for shear and torsion, as well as for discontinuities (*e.g articulations*). The results are compared to those of widely accepted commercial packages and to those from the existing literature, proving the efficiency of the proposed formulation.

2 PROBLEM FORMULATION

2.1 Basic Equations

The force method is solely based on equilibrium arguments. For elastic analysis, an excellent review paper has been written by A. Kaveh [4]. The brief presentation that follows below incorporates plasticity.

The stress components of a structure may be represented by a vector "$\mathbf{Q_s}$" which is expressed as a linear combination of a vector "$\mathbf{Q_0}$" that contains a part of the stress components which is due to external loads and a vector "$\mathbf{Q_1}$" containing a part of the stress components which is due to internal redundant stresses; a load scaling factor "γ" may also be included:

$$\mathbf{Q_s} = \gamma \cdot \mathbf{Q_0} + \mathbf{Q_1} \tag{1}$$

The above equation may be further developed by expressing each of the two linearly independent vectors as the product of a matrix "$\mathbf{B_i}$" and a vector "$\mathbf{p_i}$", where $\mathbf{i}=\{0,1\}$. These matrices express the values of the stress components of the structure due to unit valued external loads "$\mathbf{p_0}$" and unit valued internal redundant stresses "$\mathbf{p_1}$", respectively:

$$\mathbf{Q_s} = \gamma \cdot \mathbf{B_0} \cdot \mathbf{p_0} + \mathbf{B_1} \cdot \mathbf{p_1} \tag{2}$$

Equation (2) may be satisfied for an infinite number of redundant stress vectors "$\mathbf{p_1}$", but the problem is narrowed down to a unique solution with the help of the compatibility condition, as was first proposed by James Clerk−Maxwell:

$$\mathbf{B_1^T} \cdot \mathbf{q_s} = \mathbf{0} \tag{3}$$

Where in (3), vector "$\mathbf{q_s}$" contains the generalized deformations of the actual structure at its points of reference (*nodes*), which may be expressed as the summation of their elastic "$\mathbf{q_{el}}$" and plastic "$\mathbf{q_{pl}}$" components:

$$\mathbf{q_s} = \mathbf{q_{el}} + \mathbf{q_{pl}} \tag{4}$$

The elastic components in (4) are computed with the help of the unassembled flexibility matrix "$\mathbf{F}$":

$$\mathbf{q_{el}} = \mathbf{F} \cdot \mathbf{Q_s} \tag{5}$$

The plastic components in (4) may be computed with the help of the axiom of maximization of plastic work [5], using the stress derivative of an adequately defined yield function "$g(\mathbf{Q_s})$":

$$\mathbf{q_{pl}} = \lambda \cdot \frac{\partial g(\mathbf{Q_s})}{\partial \mathbf{Q_s}} \tag{6}$$

Where in (6) "λ" denotes a Lagrange (*or plastic*) multiplier. This multiplier may be explicitly computed within the framework of mathematical programming [6], using a linear complementarity condition between the potential of the material to absorb stress up to its' conventional yield point (*plastic potential, or "Y_*"*), and the corresponding Lagrange (*plastic*) multiplier "λ":

$$\mathbf{Y}_*^T \cdot \boldsymbol{\lambda} = 0 \quad , \quad \mathbf{Y}_* \geq 0 \quad , \quad \boldsymbol{\lambda} \geq 0 \tag{7}$$

Schematically, the concept of the plastic potential "Y_*" and its' corresponding Lagrange (*plastic*) multiplier may be depicted in Figure 1 below:

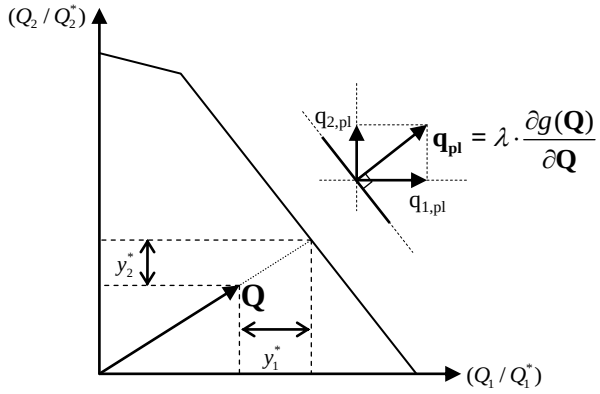


Figure 1: Plastic potential and corresponding Lagrange multiplier.

In algebraic terms, the plastic potential "Y_*" is a scalar, dimensionless quantity that expresses the numerical difference between the maximum and the actual value of the yield function, and is by definition nonnegative.

The displacements along the direction of the external loads may be computed with the help of the static−kinematic duality (SKD):

$$\mathbf{u} = \mathbf{B}_0^T \cdot \mathbf{q_s} \tag{8}$$

Historically, the SKD may be seen as another variant of the principle of virtual works (PVW), as was first proposed by Archimedes of Syracuse through the principle of leverage [7]; it expresses a "balance" between the work of external forces and the work of internal stresses that develop in the structure.

Another interesting remark is that the compatibility condition (3) may be seen as a special case of (8), where the displacements along the direction of the redundant components are zero.

2.2 Three-Dimensional Equilibrium

By considering the cross product in matrix form for a Cartesian space, the induced moments "**M**" at a particular point "*f*" due to a force vector "**F**" that is applied at another point "s" may be described by the following equation:

$$\mathbf{M} = \mathbf{d}_R \cdot \mathbf{F} = \begin{bmatrix} 0 & -\Delta z & +\Delta y \\ +\Delta z & 0 & -\Delta x \\ -\Delta y & +\Delta x & 0 \end{bmatrix} \cdot \begin{Bmatrix} F_x \\ F_y \\ F_z \end{Bmatrix} \tag{9}$$

Where

$$\Delta x = x_f - x_s \quad , \quad \Delta y = y_f - y_s \quad , \quad \Delta z = z_f - z_s \tag{10}$$

The equilibrium condition in global coordinates between the two points "*f*" and "*s*" may then be written in matrix form, as follows:

$$\begin{Bmatrix} \mathbf{F} \\ \mathbf{M} \end{Bmatrix}_f = -\begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{d}_R & \mathbf{I} \end{bmatrix} \cdot \begin{Bmatrix} \mathbf{F} \\ \mathbf{M} \end{Bmatrix}_s \tag{11}$$

2.3 Linear Projections from Global to Local Coordinates

A simple to implement, yet efficient linear transformation from global to local coordinates, is the following:

$$\mathbf{v}_L = \mathbf{T}_R \cdot \mathbf{v}_G \tag{12}$$

Where $\mathbf{v}_G = \begin{Bmatrix} v_x & v_y & v_z \end{Bmatrix}^T$ is a vector whose components are defined with respect to a global Cartesian coordinate system with basis {x,y,z}, and $\mathbf{v}_L = \begin{Bmatrix} v_1 & v_2 & v_3 \end{Bmatrix}^T$ is the same vector with respect to a local Cartesian coordinate system with basis {1,2,3}. The transformation matrix "$\mathbf{T}_R$" for the general case may then be expressed as follows:

$$\mathbf{T}_R = \begin{bmatrix} +\frac{v_x}{L} & +\frac{v_y}{L} & +\frac{v_z}{L} \\ -\frac{v_x}{L} \cdot \frac{v_z}{L_{xy}} \cdot \cos\alpha + \frac{v_y}{L} \cdot \sin\alpha & -\frac{v_y}{L} \cdot \frac{v_z}{L_{xy}} \cdot \cos\alpha - \frac{v_x}{L} \cdot \sin\alpha & +\frac{L_{xy}}{L} \cdot \cos\alpha \\ +\frac{v_x}{L} \cdot \frac{v_z}{L_{xy}} \cdot \sin\alpha + \frac{v_y}{L} \cdot \cos\alpha & +\frac{v_y}{L} \cdot \frac{v_z}{L_{xy}} \cdot \sin\alpha - \frac{v_x}{L} \cdot \cos\alpha & -\frac{L_{xy}}{L} \cdot \sin\alpha \end{bmatrix} \tag{13}$$

Where $L = \sqrt{{v_x}^2 + {v_y}^2 + {v_z}^2}$ and $L_{xy} = \sqrt{{v_x}^2 + {v_y}^2}$ and "α" is a rotation angle around the axis of the vector "$\mathbf{v}_G$", or "$\mathbf{v}_L$". This transformation is equivalent to the one presented in [3].

For each of the special cases where the global vector "$\mathbf{v}_G$" is perpendicular to a Cartesian plane {x,y} or {y,z} or {z,x}, the elements of the transformation matrix "$\mathbf{T}_R$" are efficiently defined using proper numerical exception handling workarounds with the help of the Levi−Civita cyclically interchanging index [8]. The rotation of angle "α" around the longitudinal axis is then applied separately, with the help of the corresponding (*rotation*) tensor "**R**":

$$\mathbf{R} = \begin{bmatrix} +1 & 0 & 0 \\ 0 & +\cos\alpha & +\sin\alpha \\ 0 & -\sin\alpha & +\cos\alpha \end{bmatrix} \tag{14}$$

2.4 Equilibrium Matrices

The computation of the equilibrium matrices depends on the automated selection of a statically determinate basis that carries the external loads (*matrix* "$\mathbf{B_0}$") and of a minimal (*or near minimal*) cycle basis that defines the redundant stresses (*matrix* "$\mathbf{B}_1$"). These automation techniques rely on graph theory and algorithms; analytical presentations may be found in [1], [2] and [3].

From a computational implementation perspective, it is convenient to define the equilibrium matrices "$\mathbf{B_0}$" and "$\mathbf{B}_1$" using a set of pre−solved mathematical formulae that simultaneously express equilibrium in 3D space as well as transformation of the stress components from the global coordinate system to the local of each member of the structure. By combining equations (9) − (14), we have:

$$\mathbf{B}_i = \left[\begin{array}{c|c} \mathbf{T}_R & \mathbf{0}_{3x3} \\ \hline \mathbf{T}_R \cdot \mathbf{d}_R & \mathbf{T}_R \end{array}\right] \quad , \quad i = \{0,1\} \tag{15}$$

An additional (±) sign is to be applied to each row of (15), according to whether the direction of the local axes of the surface of the corresponding reference section coincide with the direction of the local axes of the corresponding member, or not.

In this way, the stress and the elastic and plastic strain components of each structural member for every step of the incremental procedure may be quickly evaluated, using the incremental forms of equations (2), (5) and (6), respectively.

2.5 Linearized Yield Functions

For an analysis of a structural frame that takes into account material post−elastic behavior, a plastic hinge approach is traditionally followed. From the strength of materials' point of view, the conventional limits between elastic and plastic deformations may be defined using a convex hull, which, in the case of the proposed method, is further approximated by a linear convex polyhedron (*also referred to as "manifold"*):

$$g(\mathbf{Q}) = \sum_{i=1}^{n} s_i \cdot \frac{Q_i}{(\pm) Q_i^*} \le c \tag{16}$$

Where $i=\{1,2,\ldots,n\}$, "n" is the number of stress components, $\mathbf{Q} = \{ N \quad Q_2 \quad Q_3 \quad T \quad M_2 \quad M_3 \}^T$ is a vector that contains all potentially interacting stress components (see Figure 2 below), and "*c*" is a dimensionless constant (e.g. *c*=1).

Figure 2: Local force and moment components at the end of a linear finite beam/column element.

Due to the 3-dimensional nature of the problem, a series of stress interaction criteria may be defined, e.g. $\{N,M_2,M_3\}$ or $\{T,M_2,M_3\}$ or $\{T,Q_2,Q_3\}$ or $\{N,Q_2,M_3\}$ or $\{N,Q_3,M_2\}$ or $\{N,Q_2,Q_3,M_2,M_3\}$, etc. Applied implementations of such criteria may be found e.g. in structural norms.

For example, the AISC−LRFD [9] proposes a criterion which is a bi-segmented yield function for bending moment and axial force interaction; schematically, it is presented in Figure 3 below:

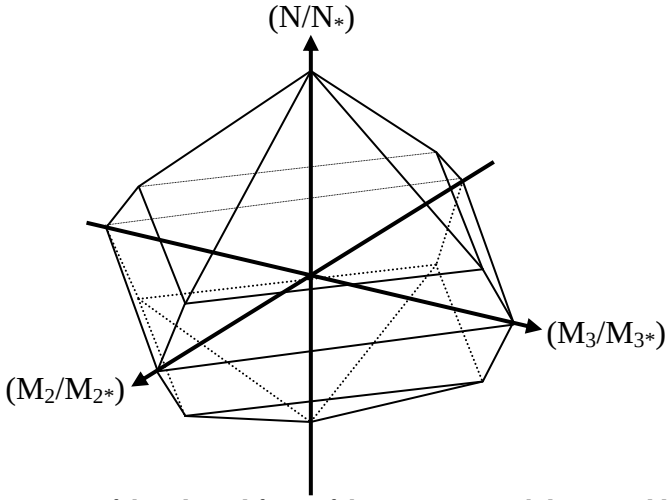


Figure 3: 3D illustration of the adopted form of the AISC-LRFD bilinear yield function.

Another example from engineering practice is the failure criterion according to DIN−18880, which is a multi−segmented yield function for bending moment, shear, and axial force interaction. A definition of this function for planar frames is analytically presented in [10]; herein, the corresponding coefficients for implementing this function according to (16) are summarized in Table 1 below:

S_i	N	Q	M
Eq. 1	−	−	1.00
Eq. 2	1.00	−	1.00/1.10
Eq. 3	−	0.45/1.15	1.00/1.15
Eq. 4	1.10/1.25	0.45/1.25	1.00/1.25
Eq. 5	−	1.00/0.90	−

Table 1: Coefficients "s_i" for the yield functions according to DIN-18800.

A practical problem that arises from the increase of the number of interacting components has to do with the (±) signs in (16), which are used to define the sector of the hyperspace wherein each one of the polyhedron's stress interaction hyper planes is located (*positive or negative plastic capacity for every stress component*); but this may be easily tackled with the help of De Bruijn sequences [11].

For example, consider the simplest case of two interacting stress components (n=2), where sets of two are formed from the alphabet with symbols {+,−} (q=2). Then, we have four possible sets ($q^n=2^2=4$) which form one distinct sequence ($B=(2!)^{2^{2-1}}/2^2=1$), shown in Figure 4:

$$\{+,+\},\{+,-\},\{-,+\},\{-,-\}$$

Figure 4: De Bruijn sequence for the signs of a yield function with two interacting components.

Next, consider the case of three interacting components (n=3) where sets of three are formed from the alphabet with symbols {+,−} (q=2). Then, we have eight possible sets ($q^n=2^3=8$) which form two distinct sequences ($B=(2!)^{2^{3-1}}/2^3=2$), shown in Figure 5:

$$\{+,+,+\},\{+,+,-\},\{+,-,+\},\{+,-,-\}$$
$$\{-,+,+\},\{-,+,-\},\{-,-,+\},\{-,-,-\}$$

Figure 5: De Bruijn sequences for the signs of a yield function with three interacting components.

It may be observed in Figure 5 above, that one sequence is a negated permutation of the other; in fact, the sequence consists of the sequence shown in Figure 4, embedded in a simple (*elementary*) set (of n=1 and q=2). This means, that, using the sequence of Figure 4 as a basis, and in combination with the basic principle of enumeration, all possible sequences may be constructed following the same pattern of successive embedding (*or concatenation*).

A last example follows below, where five interacting components (n=5) are forming sets of five from the alphabet with symbols {+,−} (q=2), yielding a total of thirty two sets ($q^n=2^5=32$) which form potentially two thousand and forty eight sequences ($B=(2!)^{2^{5-1}}/2^5=2048$); one of these has been formed using the aforementioned successive embedding (*concatenation*) technique, and is schematically presented in Figure 6:

$$\{+,+,+,+,+\},\{+,+,+,+,-\},\{+,+,+,-,+\},\{+,+,+,-,-\}$$
$$\{+,+,-,+,+\},\{+,+,-,+,-\},\{+,+,-,-,+\},\{+,+,-,-,-\}$$
$$\{+,-,+,+,+\},\{+,-,+,+,-\},\{+,-,+,-,+\},\{+,-,+,-,-\}$$
$$\{+,-,-,+,+\},\{+,-,-,+,-\},\{+,-,-,-,+\},\{+,-,-,-,-\}$$
$$\{-,+,+,+,+\},\{-,+,+,+,-\},\{-,+,+,-,+\},\{-,+,+,-,-\}$$
$$\{-,+,-,+,+\},\{-,+,-,+,-\},\{-,+,-,-,+\},\{-,+,-,-,-\}$$
$$\{-,-,+,+,+\},\{-,-,+,+,-\},\{-,-,+,-,+\},\{-,-,+,-,-\}$$
$$\{-,-,-,+,+\},\{-,-,-,+,-\},\{-,-,-,-,+\},\{-,-,-,-,-\}$$

Figure 6: De Bruijn sequences for the signs of a yield function with five interacting components.

In practice, only one of these 2048 sequences is required in order to properly describe all the hyper planes of the polyhedron (*manifold*), since, by definition, all other sequences are only permutations of the sets of five that form the initial sequence.

Sequences for the case of alphabets with two elements were first proved by Camille Flye Sainte−Marie [12], 24 years before Nicolaas Govert de Bruijn was born.

2.6 Implementation of Discontinuities

For the case of sections that include discontinuities (*e.g. articulations*), a disjunction of the interacting components of the corresponding plastic hinge may be imposed that allows plasticization only for the active stress component(s).

For example, consider an articulated beam element with plastic hinges according to the AISC−LRFD criterion, defined for $\{N,M_2,M_3\}$ interaction. Since the element has articulations at both of its' ends, the $\{M_2\}$ and $\{M_3\}$ components will be always zero, for both positive and negative rotations; this leaves only the {N} component as active candidate for plasticization. In practice, this means that a separate static admissibility condition for each stress component must be defined; thus, we will have a yield function consisting of three separate parts, one for each component, with the formulae for $\{M_2\}$ and $\{M_3\}$ being degenerate, in order to describe that the corresponding bearing capacity is zero. This disjunct yield function is presented in equation (17):

$$\begin{aligned} g_1(\mathbf{Q}) &= \left\{ \tfrac{1}{(\pm)N^*} \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \right\} \cdot \mathbf{Q} \le 1 \\ g_2(\mathbf{Q}) &= \left\{ 0 \quad 0 \quad 0 \quad 0 \quad \pm 1 \quad 0 \right\} \cdot \mathbf{Q} \le 0 \\ g_3(\mathbf{Q}) &= \left\{ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \pm 1 \right\} \cdot \mathbf{Q} \le 0 \end{aligned} \tag{17}$$

Where $\mathbf{Q}=\left\{ N \quad Q_2 \quad Q_3 \quad T \quad M_2 \quad M_3 \right\}^T$ is the stress vector of the corresponding critical section and the (±) sign is used to denote upper and lower bounds.

It should be noted that the proposed disjunction technique presented in (17) may be used to define all sorts of discontinuities; the corresponding Lagrange multiplier is added to its' elastic counterpart in order to yield the actual elastic generalized displacement of the corresponding discontinuity.

2.7 Problem Formulation

By combining equations (2), (15) and (16), the yield function "$g(\mathbf{Q})$" may be explicitly expressed as a function of the external loads and the redundant stresses. By packing the coefficients of this composite function into an incidence matrix "$\mathbf{N}$", a linear constraint inequality may be established:

$$\mathbf{Y}_* = \mathbf{c} - \left(\mathbf{N}^T \cdot \mathbf{B}_1\right) \cdot \mathbf{p}_1 - \gamma \cdot \left(\mathbf{N}^T \cdot \mathbf{B}_0\right) \cdot \mathbf{p}_0 \geq \mathbf{0} \tag{18}$$

By substituting equation (2) into (5), then evaluating (6) with the help of (16), then substituting both into (4) and then into (3), the first derivative of a Lagrange function is obtained:

$$\left(\mathbf{B}_1^T \cdot \mathbf{F} \cdot \mathbf{B}_1\right) \cdot \mathbf{p}_1 + \gamma \cdot \left(\mathbf{B}_1^T \cdot \mathbf{F} \cdot \mathbf{B}_0\right) \cdot \mathbf{p}_0 + \left(\mathbf{B}_1^T \cdot \mathbf{N}\right) \cdot \boldsymbol{\lambda} = \mathbf{0} \tag{19}$$

Equations (18) and (19) together with (7) may be seen as the Karush−Kuhn−Tucker (KKT) conditions of the following optimization problem:

$$\begin{aligned} \textit{Min.}: \quad & f(\mathbf{p}_1) = \frac{1}{2} \cdot \mathbf{p}_1^T \cdot \left(\mathbf{B}_1^T \cdot \mathbf{F} \cdot \mathbf{B}_1\right) \cdot \mathbf{p}_1 + \gamma \cdot \mathbf{p}_1^T \cdot \left(\mathbf{B}_1^T \cdot \mathbf{F} \cdot \mathbf{B}_0\right) \cdot \mathbf{p}_0 \\ \textit{s.t.} \quad & \left(\mathbf{N}^T \cdot \mathbf{B}_1\right) \cdot \mathbf{p}_1 \leq \mathbf{c} - \gamma \cdot \left(\mathbf{N}^T \cdot \mathbf{B}_0\right) \cdot \mathbf{p}_0 \end{aligned} \tag{20}$$

In order to efficiently trace all the plasticization events along any given loading path, the independent variables and the critical parameters stated in the problem of equation (20) are replaced by their corresponding increments:

$$\begin{aligned} \textit{Min.}: \quad & f(\Delta\mathbf{p}_{1,k}) = \frac{1}{2} \cdot \Delta\mathbf{p}_{1,k}^T \cdot \left(\mathbf{B}_1^T \cdot \mathbf{F} \cdot \mathbf{B}_1\right) \cdot \Delta\mathbf{p}_{1,k} + \Delta\gamma \cdot \Delta\mathbf{p}_{1,k}^T \cdot \left(\mathbf{B}_1^T \cdot \mathbf{F} \cdot \mathbf{B}_0\right) \cdot \Delta\mathbf{p}_{0,k} \\ \textit{s.t.} \quad & \left(\mathbf{N}^T \cdot \mathbf{B}_1\right) \cdot \Delta\mathbf{p}_{1,k} \leq \mathbf{c}_k - \mathbf{Q}_{k-1} - \Delta\gamma \cdot \left(\mathbf{N}^T \cdot \mathbf{B}_0\right) \cdot \Delta\mathbf{p}_{0,k} \end{aligned} \tag{21}$$

Where in (21) "k" is the step's counter. The problem is solved in an incremental pattern using any efficient algorithm (e.g. [13], [14]); a detailed description of the adopted numerical strategy may be found in [1].

3 EXAMPLES

In the following context, four indicative examples are presented in order to verify the functionality of the proposed force−based formulation. Where possible, results are quantitatively compared with those of a widely accepted commercial program that uses the equivalent direct stiffness method; where not, with results according to methods found in the literature.

3.1 1 Storey, 1 Bay, Eccentric Braced 3D Frame

A single-storey, single-bay, eccentric-braced frame is used as a first test example to prove the efficiency of the proposed formulation. The frame's height is H=3m, and distances between columns are L_x=L_y=6m. All columns are placed so that their strong bending axis is parallel to the y-axis of the global coordinate system, and are fully fixed at the basis. Beams are placed so that they bend along their strong axis due to external forces applied along the z-axis, and are subdivided into smaller elements of length L=0.6m each. The braces are placed inside the openings which are parallel to the y-z plane, are assumed articulated at both of their ends, and are load-free. Schematically, the frame is illustrated in Figure 7:

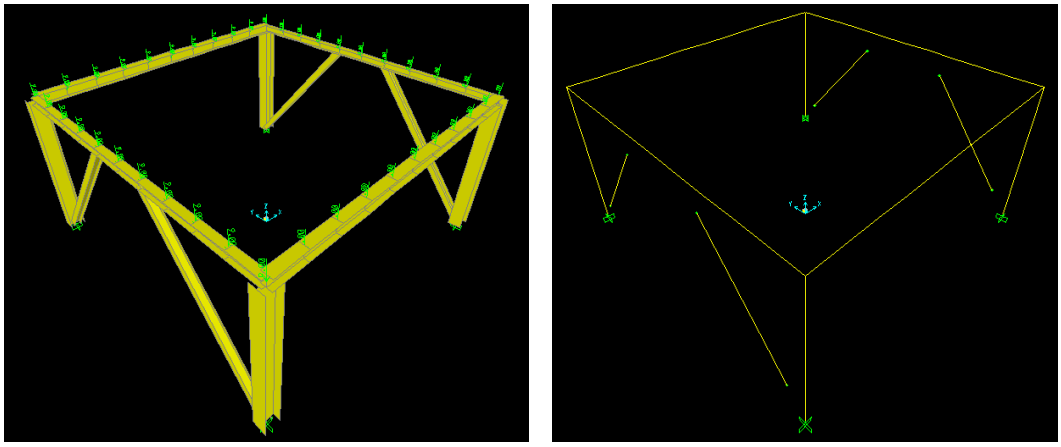


Figure 7: 3D illustrations of the 1−storey 1−bay, eccentric−braced frame (snapshots from SAP2000).

The frame's definitions (*nodal numbering & coordinates, member connectivity, section assignments and interacting stress components*) are summarized in Tables 2 and 3:

Node	x	y	z	Node	x	y	z
1	6.00	6.00	0.00	7	0.00	3.60	3.00
2	0.00	6.00	0.00	8	0.00	2.40	3.00
3	0.00	0.00	0.00	9	0.00	0.00	3.00
4	6.00	0.00	0.00	10	6.00	0.00	3.00
5	6.00	6.00	3.00	11	6.00	2.40	3.00
6	0.00	6.00	3.00	12	6.00	3.60	3.00

Table 2: Nodal coordinates of the 1−storey 1−bay, eccentric−braced frame.

Member	Node: Start → End	Section	Interacting Components	Member	Node: Start → End	Section	Interacting Components
1	1 → 5	HEM260	{N,M2,M3}	9	6 → 5	HEM180	{M2,M3}
2	2 → 6	HEM260	{N,M2,M3}	10	9 → 8	HEB140	{M2,M3}
3	3 → 9	HEM260	{N,M2,M3}	11	8 → 7	HEB140	{M2,M3}
4	4 → 10	HEM260	{N,M2,M3}	12	7 → 6	HEB140	{M2,M3}
5	1 → 12	HEB140	{N,M2,M3}	13	9 → 10	HEM180	{M2,M3}
6	2 → 7	HEB140	{N,M2,M3}	14	10 → 11	HEB140	{M2,M3}
7	3 → 8	HEB140	{N,M2,M3}	15	11 → 12	HEB140	{M2,M3}
8	4 → 11	HEB140	{N,M2,M3}	16	12 → 5	HEB140	{M2,M3}

Table 3: Connectivity, sections, and interacting components of the 1−storey 1−bay, eccentric−braced frame.

The material of the structure is S235, with a Young's Modulus E=2.0E+8kPa, a conventional yield stress of f_y=235MPa, and is considered to be elastic-perfectly plastic.

To simulate material non−linearity, the concentrated plasticity approach is followed. For the columns, the bilinear AISC−LRFD criterion is used [9], for $\{N,M_2,M_3\}$ interaction. For the beams, the coupled bending moment $\{M_2,M_3\}$ variant of the same criterion is applied, without the participation of the axial forces {N}; in order to be able to compare with SAP2000, the effect of torsion {T} on beams was purposefully left out, since no torsion-bending interaction is supported by the available version of the program [15]. A linear relation between $\{M_2\}$ and $\{M_3\}$ was adopted for both interaction criteria. Articulations at the ends of the braces were implemented according to equation (17).

All beams are subject to a uniform load of 15kN/m. Each beam is divided into smaller elements of 0.6m length, and uniform loads are converted into a finite set of equally sized point loads of 9kN. Lateral loads are applied to the four top-corner nodes of the frame.

Two pushover analyses were run, one for each horizontal direction {x,y}. The base shear vs. roof displacement curves of the structure are presented in Figures 8 and 9 below, where quantitative comparisons with SAP2000 [15] are also included; for SAP2000, default analysis parameters were used.

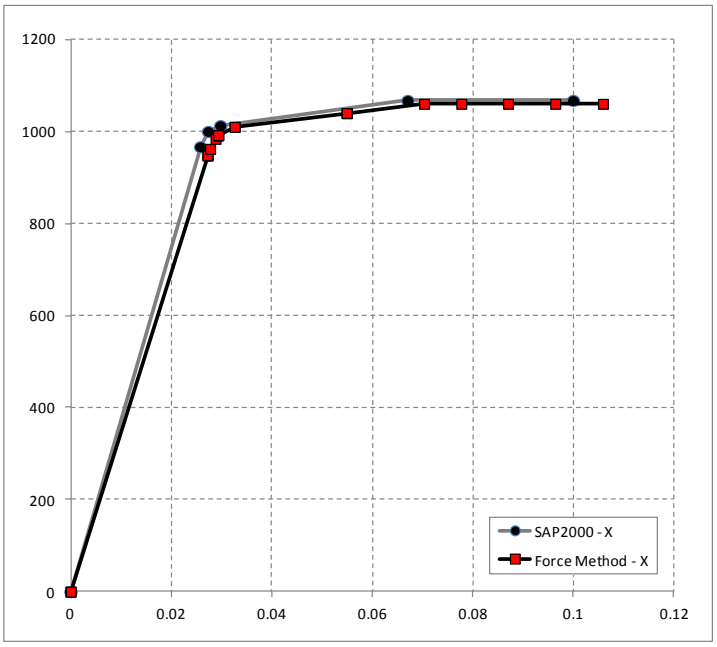


Figure 8: Pushover curve of the 1−storey, 1−bay, eccentric−braced frame; X−Direction; Units: {kN,m}.

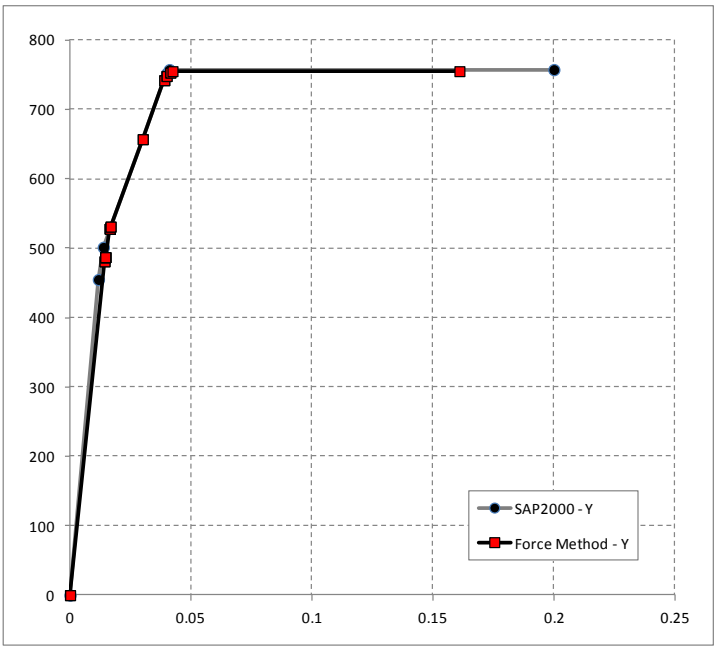


Figure 9: Pushover curve of the 1−storey, 1−bay, eccentric−braced frame; Y−Direction; Units: {kN,m}.

As it may be seen, results are in good accordance; the ultimate base shear is $V_{b,x}$~1067kN for the x-direction, and $V_{b,y}$~757kN for the y-direction.

3.2 6 Storey, 1 Bay, 3D Frame

A six-storey frame, with one-bay at each horizontal direction, is used as a second test example to prove the efficiency of the proposed formulation. Each storey has a height H=3.0m and each bay an opening L_x=L_y=6.0m. Beams are subdivided into smaller elements of length L=0.6m each; columns are placed so that their strong bending axis is parallel to the y-axis, and are fully fixed at the basis. The geometry of the frame is illustrated in Figure 10 below:

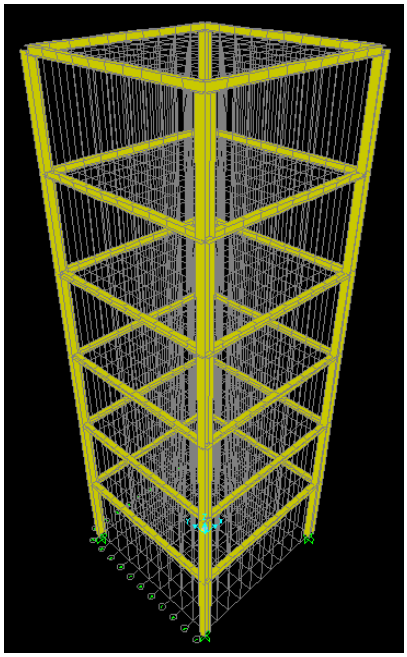


Figure 10: 3D illustration of the 6−storey frame (snapshot from SAP2000).

The sections of beams and columns were purposefully selected in order to form a strong column−weak beam sway mechanism, and are summarized in Table 4 below:

Members	Length	Section	Material
Columns	3.0m	HEM300	S235
Beams	6.0m	HEM180	S235

Table 4: Section properties of the six storey frame.

The material of the structure is S235, with a Young's Modulus E=2.0E+8kPa, a conventional yield stress of f_y=235MPa, and is considered to be elastic-perfectly plastic.

To simulate material non−linearity, the concentrated plasticity approach is followed, according to the exact assumptions that were made for the first example.

All beams are subject to a uniform vertical load of 15kN/m. Each beam is subdivided into ten smaller elements of length L=0.6m each, and the distributed loads are converted into a finite set of equally sized point loads of value 9kN each.

A lateral load pattern is applied to the beam−column junction nodes. The load distribution is defined as linearly varying with respect to the height of each storey (*1st eigenmode, with 100% participation*).

Two pushover analyses were run, one each horizontal direction (x,y). The base shear vs. roof displacement curves of the structure is presented in Figures 11 and 12 below, where a quantitative comparisons with SAP2000 [15] are also included; for SAP2000, default analysis parameters were used.

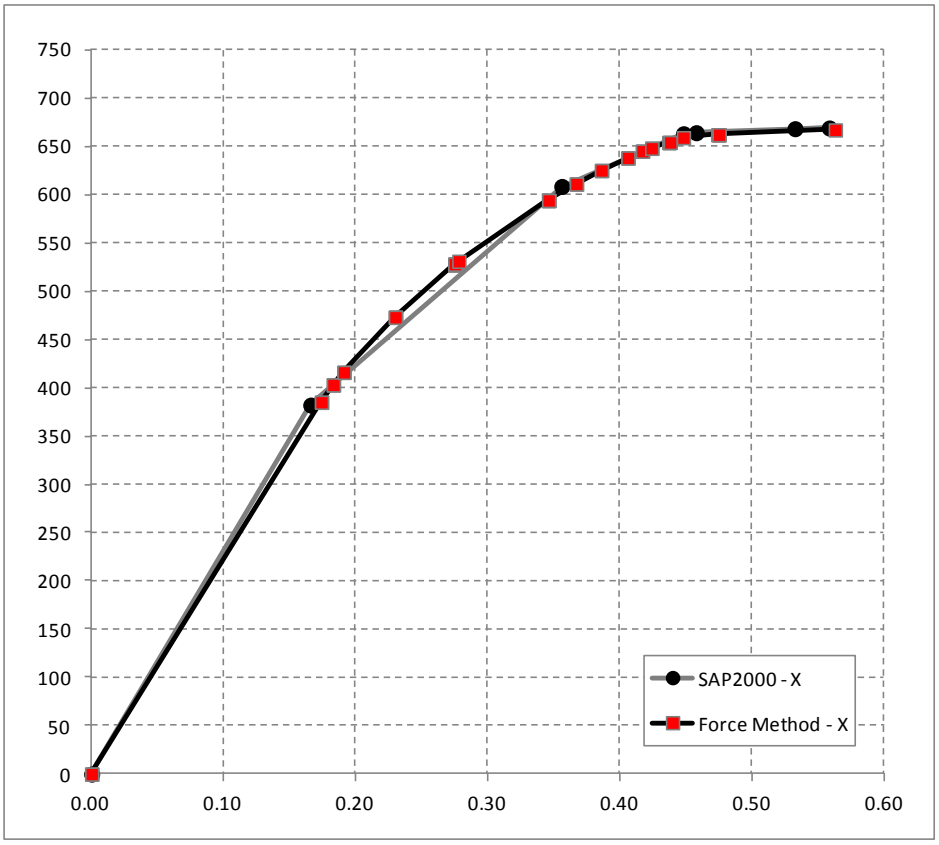


Figure 11: Pushover curves of the six−storey frame, x−direction; Units: {kN,m}.

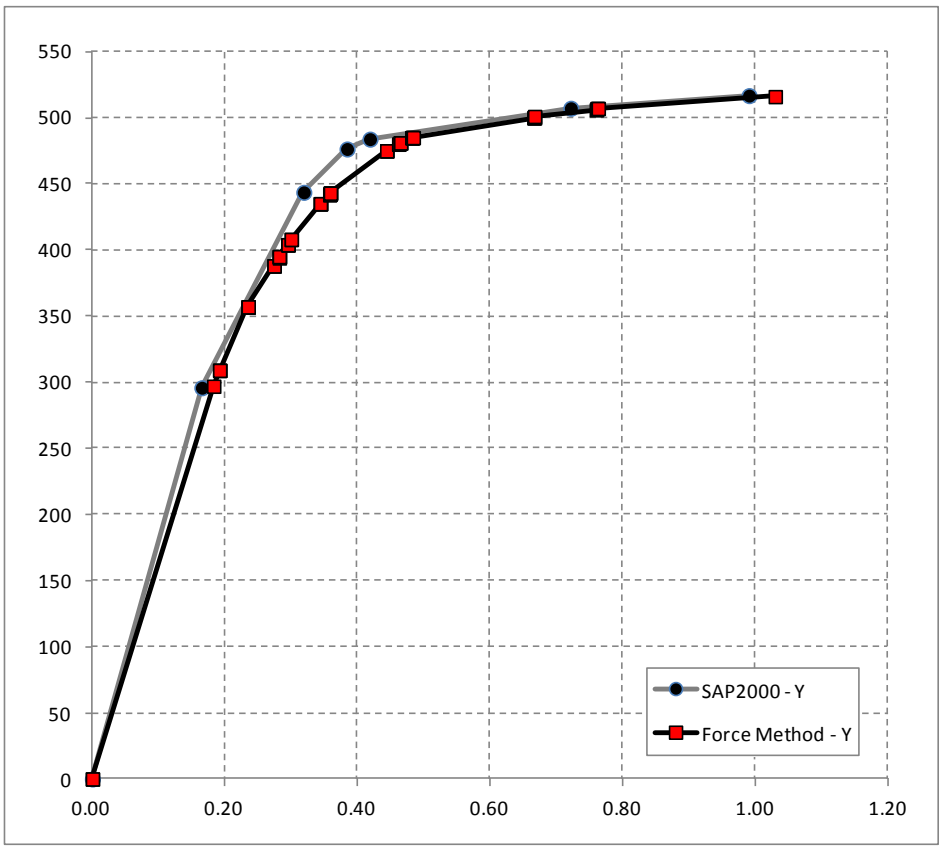


Figure 12: Pushover curves of the six−storey frame, y−direction; Units: {kN,m}.

As it may be seen, the results are in good accordance; the ultimate base shear along the x-direction is $V_{b,x}$~667kN, and along the y-direction is $V_{b,y}$~516kN.

3.3 1 Storey, 1 Bay, Portal 2D Frame

This example is used to test the functionality of the herein implemented yield function according to DIN−18800, via a comparison with AISC−LRFD; to this extend, I-beams are used.

The frame's height is H=3.0m and length L=6.0m. The beam's section is HEB160 and the columns' sections are HEM200. All structural elements are placed so as to bend around their strong axis. The beams are subdivided into smaller elements of length L=0.6m each. Columns are fully fixed at the basis. Schematically, the portal frame may be seen in Figure 13 below:

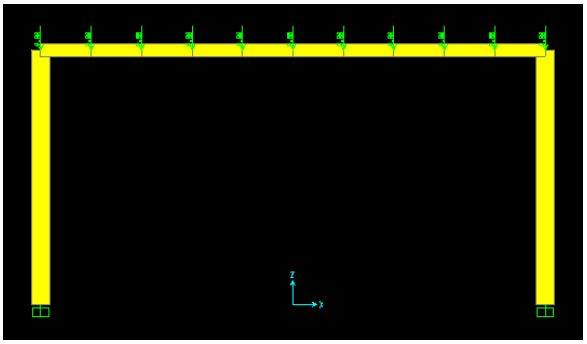


Figure 13: Geometry of the portal frame and vertical loads (snapshot from SAP2000).

The material of the structure is S235 with a Young's Modulus E=2.0E+8kPa, a conventional yield stress of f_y=235MPa, and is considered to be elastic-perfectly plastic.

A uniform vertical load of magnitude 15kN/m is applied to the beam, which is simulated by finite point loads of magnitude 9kN each, applied in equal distances of 0.6 m. A horizontal force is applied to the top left node.

Two pushover analyses were performed using the proposed formulation; one using the AISC−LRFD yield function and one using the DIN−18800. The results were compared to those of SAP2000 [15] and were found in good accordance. All base shear vs. roof displacement curves are plotted in Figure 14 below:

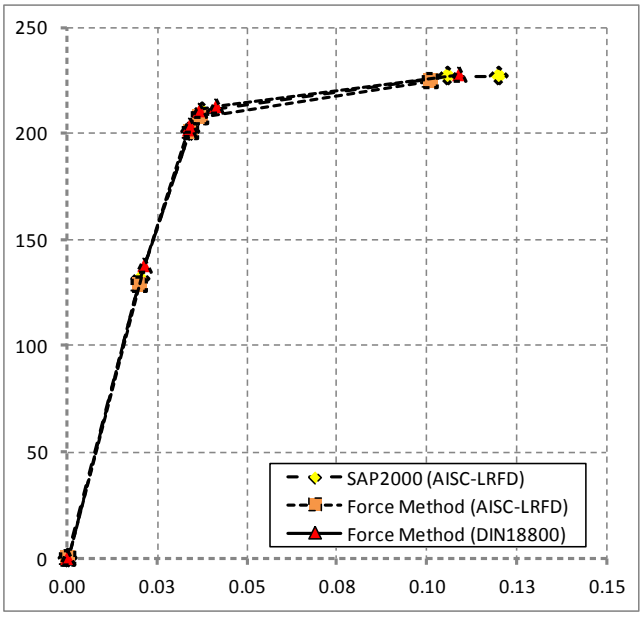


Figure 14: Pushover curves of the 2D portal frame; Units: {kN,m}.

3.4 A Simple Grillage

This simple example is used to demonstrate the functionality of the proposed formulation for structures where the contribution of torsion is important. Grillages are a typical case; a simple grillage found in [16] is used as reference. Schematically, it is presented in Figure 15:

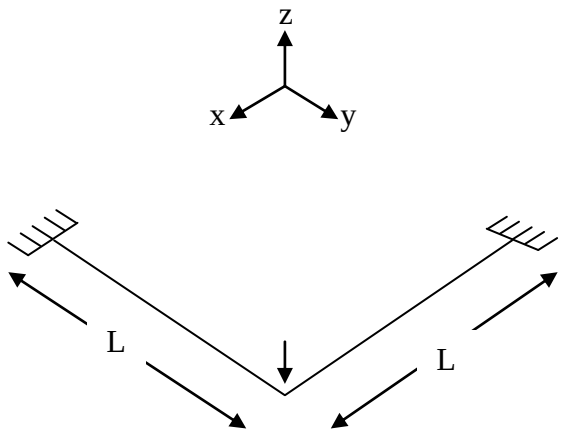


Figure 15: Geometry and loading of the grillage.

As it may be observed, the geometrical proportions and loading conditions are the same as in [16]; herein, elements have a length L=3m and a rectangular tube section with dimensions (bxh)=(160x160)mm and thickness t=10mm.

The material of the structure is S235 with a Young's Modulus E=2.0E+8kPa, a conventional yield stress of f_y=235MPa, and (*of course*) is considered to be elastic-perfectly plastic.

Additionally to the assumptions above, the yield locus was linearized so as to be in accordance with the proposed formulation. An analytical derivation of the collapse load according to the linear yield function adopted follows below:

$$g(T, M_3) = \frac{T}{T_p} + \frac{M_3}{M_{3,p}} - 1 \le 0 \tag{22}$$

The partial derivatives of the yield function with respect to each stress component are:

$$\frac{\partial g}{\partial T} = \frac{1}{T_p} \quad , \quad \frac{\partial g}{\partial M_3} = \frac{1}{M_{3,p}} \tag{23}$$

Thus, with the help of (23), the ratio of bending (θ) to torsion (γ) plastic rotations will be:

$$\frac{\theta}{\gamma} = \left(\frac{T_p}{M_{3,p}} \right) \cdot \frac{M_3}{T} \tag{24}$$

As also stated in [16], a plastic bending rotation in the proximity of a support node of one beam will result in an equal rotation due to torsion in the other beam, in the proximity of the connection node between the two elements. Thus, from (24) we infer the following linear proportion:

$$\frac{M_3}{M_{3,p}} = \frac{T}{T_p} \tag{25}$$

According to [16], the equation that gives the collapse load is the following:

$$\frac{1}{2} \cdot P \cdot L = M_3 + T \tag{26}$$

By combining (25) and (26), and by assuming for simplicity that $M_{3p}=T_p$, we have:

$$P_C = \frac{2 \cdot M_{3,p}}{L} \tag{26}$$

For the selected cross-section and material, $M_{3p}=79.4kNm$; thus, from (26), P_C=52.93kN.

An analysis using the proposed incremental formulation of (21) was run, and the resulting load vs. corresponding displacement curve is presented in Figure 16 below:

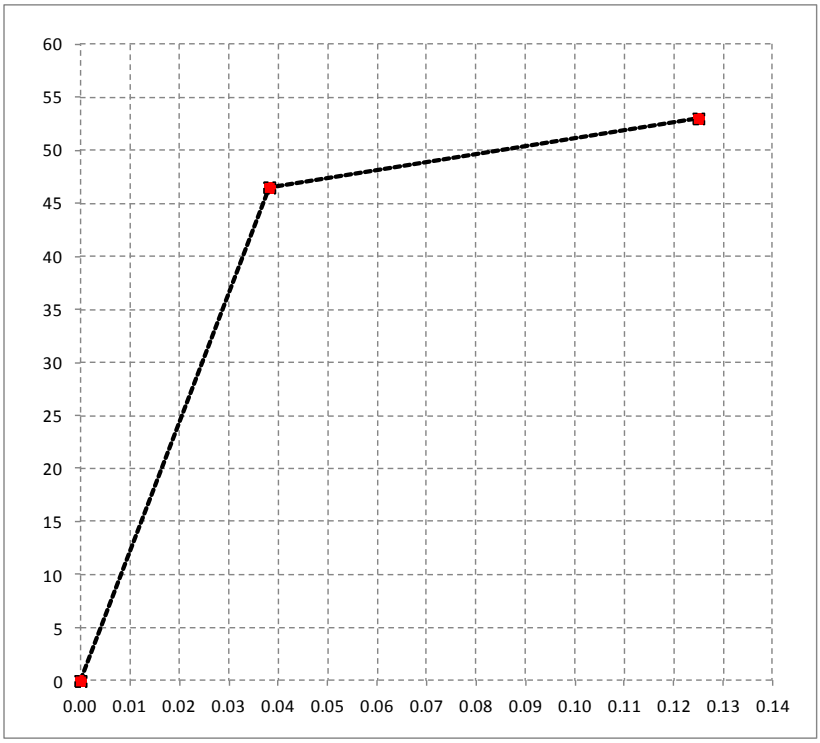


Figure 16: Load vs. corresponding displacement curve of the grillage; Units: {kN,m}.

As it may be seen, the results are in good accordance; the collapse load according to the proposed method is P_C~53kN.

4 CONCLUDING REMARKS

- The proposed formulation has good convergence properties and yields good results, even for relatively large scale problems. However, an implementation using a solver suitable for sparse matrices is required, in order to drastically reduce the CPU time.
- Non−holonomic plasticity is taken into account following a stepwise holonomic approach that is natively contained into the problem's formulation; no particular numerical remedies are required.
- Treatment of the effect of shear and torsion is also natively accommodated; however, it should be noted, that, when using the lumped plasticity approach, the warping stresses according to Vlasov [17] may be taken efficiently into account only in certain cases with the introduction of additional stress components, the bi−moments, which are defined as self−equilibrating systems on a section level.
- P−Δ effects and large displacements [18], material hardening [19] and softening, may also be included.

ACKNOWLEDGMENTS

The authors would like to thank Dimitra Voukia, M.Sc. Civil Engineer, for kindly providing us with a copy of [10].

REFERENCES

[1] K.V. Spiliopoulos, T.N. Patsios, An efficient mathematical programming method for the elastoplastic analysis of frames, *Engineering Structures*, **32** (5), 1199-1214, 2010.

[2] Spiliopoulos, K.V.. On the automation of the force method in the optimal plastic design of frames. *Computer Methods in Applied Mechanics and Engineering*, **141** (1–2), 141–156, 1997.

[3] K.V. Spiliopoulos, N.G. Dais, A powerful force-based approach for the limit analysis of three-dimensional frames. *Archive of Applied Mechanics*, **83** (5), 723–742, 2012.

[4] A.Kaveh, Recent Developments in the Force Method of Structural Analysis. *Applied Mechanics Reviews*, **45** (9), 401–418, 1992.

[5] R.O. Davis, A.P.S. Selvadurai, *Plasticity and Geomechanics*, Cambridge University Press, 2002.

[6] G. Maier, Quadratic programming and theory of elastic-perfectly plastic structures, *Meccanica*, **3** (4), 265–273, 1968.

[7] D. Capecchi, *History of Virtual Work Laws: A History of Mechanics Prospective*, Springer-Verlag, Italy, 2012.

[8] I. Vardoulakis, *Lecture Notes in Continuum Mechanics*, NTUA, 2008.

[9] *Load and resistance factor design specification for structural steel buildings, 2nd Edition*. Chicago: American Institute of Steel Construction; 1993.

[10] N. Gebbeken, *Eine Fliessgelenktheorie Höherer Ordnung für Raümliche Stabtragwerke*, Mitteilungen des Institus für Statik der Universität Hannover, Hannover, 1988.

[11] N.G. de Bruijn, A combinatorial problem, *Proc. Koninklijke Nederlandse Akademie v. Wetenschappen*, **49** (7), 758–764, 1946.

[12] Flye Sainte-Marie, C., Solution to question nr. 48, *L'intermédiaire des Mathématiciens*, **1**, 107–110, 1894.

[13] D. Goldfarb, A. Idnani, A numerically stable dual method for solving strictly convex quadratic programs, *Mathematical Programming*, **27**, 1–33, 1983.

[14] M.J.D. Powell, *TOLMIN: A Fortran Package for Linearly Constrained Optimization Calculation*, DAMTP, 1989/NA2, Cambridge, 1989.

[15] SAP2000, v14.0.0. User's manual. CSI Inc. 2009.

[16] J. Heyman, *Elements of the Theory of Structures*, Cambridge University Press, 1996.

[17] V.Z. Vlasov, *Thin-Walled Elastic Beams*, OTS 61−11400, National Science Foundation, Washington D.C., 1961.

[18] T.N. Patsios, K.V. Spiliopoulos, A force–based formulation for the 2nd order elastoplastic analysis of frames, *Proc. 2nd ECCOMAS Young Investigator's Conference*, 104–108, Bordeaux, France, September 2-6, 2013.

[19] T.N. Patsios, K.V. Spiliopoulos, A force–based formulation for the analysis of frames with non–holonomic hardening plastic hinges, *WCCM XI, ECCM V, ECFD VI*, Barcelona, Spain, July 20–25, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THE FEATURES OF ANALYTICAL SOLUTIONS OF BOUNDARY VALUE PROBLEMS OF THE ELASTICITY THEORY FOR FINITE DOMAINS WITH ANGULAR POINTS OF A BOUNDARY

Alexander P. Kerzhaev[1], Mikhail D. Kovalenko[2] and Irina V. Menshova[3]

[1] International Institute of Earthquake Prediction Theory and Mathematical Geophysics RAS (IIEPT RAS), Moscow, Russia, e-mail: alex_kerg@mail.ru

[2] IIEPT RAS, Moscow, Russia, e-mail: kov08@inbox.ru

[3] IIEPT RAS, Moscow, Russia, e-mail: menshovairina@yandex.ru

Keywords: biharmonic problem, rectangle, exact solution, residual stresses

Abstract. The paper is dedicated to the classical problem of elasticity theory – to the solution of the biharmonic equation in a semi-strip (rectangle) and to some conclusions that follow from the analysis of exact solutions of the biharmonic problem. The formulas describing the distribution of stresses and displacements in the semi-strip (rectangle) with free longitudinal sides and stresses set on the end face, as well as the mathematical apparatus used in this case can be found in the works of authors such as [1, 2]. The solutions are represented as the expansions in the Fadle–Papkovich functions. The coefficients of the expansions are determined as the simple Fourier integrals from the known boundary functions.

The exact solutions in the rectangle have some unusual properties that do not exist in any of the known exact solutions of elasticity theory. For example, they are not unique and, therefore, they may describe the residual stresses. Thus, solving the biharmonic problem for the rectangle, thus we constructed the theory of residual stresses – one of the fundamental problems of elasticity theory. The obtained solutions allow us to understand what it means to set boundary conditions on the rectilinear boundary of a domain; because she is already not rectilinear after application of the load. What is the angular point of an infinite wedge and how it differs from the angular point of the rectangle? Has got the angular point in the rectangle the singularity always, if the type of boundary conditions is changed at the angular point? And so on.

Based on the developed mathematical apparatus, it is possible to obtain the exact solutions of various mixed boundary value problems in the rectangle. For example, when there is a crack in the rectangle, when the part of the boundary of the rectangle is rigidly fixed, and the part is free, etc. The developed methods can be transferred to other coordinate systems: to polar coordinate system (cambered beam), to oblique (triangle, trapezium). They can be generalized to three-dimensional problems (rectangular parallelepiped).

Interest in solutions of boundary-value problems of elasticity theory in canonical domains with corner points of a boundary, in particular, in a rectangle, did not cease ever, having reached a peak years in 1940–1980, primarily thanks to the Soviet school of mathematics and mechanics. Several thousands of papers were published throughout these years. The last review (2003) V. V. Meleshko contains more than 700 references to the most important research in almost 200 years [3]. There are several schools that have been established in these years in the Soviet Union and were represented by the greatest mathematicians and mechanics of those years: the Leningrad school (P. F. Papkovich, A. I. Lurie, G. A. Grinberg, G. I. Dzhanelidze, V. K. Prokopov, B. M. Nuller and others), the Moscow school (M. I. Gusein-Zade, S. A. Lurie, V. V. Vasiliev, V. I. Maly, E. M. Zveryaev and many others), the Rostov-on-Don school under the leadership of I. I. Vorovich, very strong and numerous the Ukrainian school (V. T. Grinchenko, A. F. Ulitko, A. M. Gomilko, V. V. Meleshko). Strong and bright works have been published in the Proceedings of the Armenian, Azerbaijani, Georgian Academy of Sciences. After the 1980s there were no notable publications. Western papers were sketchy and much weaker than the Soviet.

The heart of the problem is simple. Let us explain it on the example of the first fundamental boundary-value problem for the semi-strip in the case of symmetric deformation. We can find the solution of the biharmonic equation in the semi-strip $\{x \ge 0, |\, y \,| \le 1\}$, that has free longitudinal sides, i.e.

$$\sigma_y(x,\pm 1) = \tau_{xy}(x,\pm 1) = 0, \tag{1}$$

and normal $\sigma_x(0,y) = \sigma(y)$ and tangential $\tau_{xy}(0,y) = \tau(y)$ stresses are set on the end face

$$\sigma_x(0,y) = \sigma(y),\ \tau_{xy}(0,y) = \tau(y). \tag{2}$$

Without loss of generality we can assume that the stresses $\sigma(y)$ are self-balanced. Solving the problem by the method of separation of variables, we come to the problem of determining the coefficients a_k of expansions of the two functions $\sigma(y), \tau(y)$ set on the semi-strip's end face in a series in the two systems of boundary value problem's eigenfunctions – the so-called Fadle–Papkovich functions:

$$\sigma(y) = \sum_{k=1}^{\infty} a_k s_x\left(\lambda_k, y\right) + \overline{a_k} s_x\left(\overline{\lambda_k}, y\right),\ \tau(y) = \sum_{k=1}^{\infty} a_k t_{xy}\left(\lambda_k, y\right) + \overline{a_k} t_{xy}\left(\overline{\lambda_k}, y\right). \tag{3}$$

The Fadle–Papkovich functions (in the case of symmetric deformation of the semi-strip) have the form:

$$\begin{aligned} s_x\left(\lambda_k, y\right) &= (1+\mu)\lambda_k \{\left(\sin\lambda_k - \lambda_k \cos\lambda_k\right)\cos\lambda_k y - \lambda_k y \sin\lambda_k \sin\lambda_k y\}, \\ t_{xy}\left(\lambda_k, y\right) &= (1+\mu)\lambda_k^2 \{\cos\lambda_k \sin\lambda_k y - y \sin\lambda_k \cos\lambda_k y\}. \end{aligned} \tag{4}$$

μ is a Poisson's ratio. The numbers λ_k are the set $\{\pm\lambda_k, \pm\overline{\lambda_k}\}_{k=1}^{\infty} = \Lambda$ of all complex zeros of the exponential type's entire function:

$$L\left(\lambda\right) = \lambda + \sin\lambda\cos\lambda. \tag{5}$$

If the expansion coefficients a_k are found, the final solution of the boundary value problem in the semi-strip will have the form (below only expressions for stresses are shown)

$$\sigma_x(x,y)=\sum_{k=1}^{\infty}a_k s_x(\lambda_k,y)e^{\lambda_k x}+\overline{a_k}s_x(\overline{\lambda_k},y)e^{\overline{\lambda_k}x},$$
$$\sigma_y(x,y)=\sum_{k=1}^{\infty}a_k s_y(\lambda_k,y)e^{\lambda_k x}+\overline{a_k}s_y(\overline{\lambda_k},y)e^{\overline{\lambda_k}x}, \tag{6}$$
$$\tau_{xy}(x,y)=\sum_{k=1}^{\infty}a_k t_{xy}(\lambda_k,y)e^{\lambda_k x}+\overline{a_k}t_{xy}(\overline{\lambda_k},y)e^{\overline{\lambda_k}x}\quad(\mathrm{Re}\,\lambda_k<0),$$

where

$$s_y(\lambda_k,y)=(1+\mu)\lambda_k\{(\sin\lambda_k+\lambda_k\cos\lambda_k)\cos\lambda_k y+\lambda_k y\sin\lambda_k\sin\lambda_k y\}.$$

The boundary conditions on the longitudinal sides of the semi-strip (1) will be satisfied exactly, because the Fadle–Papkovich functions are such, that

$$t_{xy}(\lambda_k,\pm1)=s_y(\lambda_k,\pm1)=0. \tag{7}$$

If the periodicity conditions were set on the longitudinal sides of the semi-strip, eigenfunctions would be the usual trigonometric set of functions, and then we would obtain instead of expansions (3), for example,

$$\sigma(y)=-\sum_{k=1}^{\infty}(A_k(1+\mu)k\pi+2B_k)\cos k\pi y,$$
$$\tau(y)=\sum_{k=1}^{\infty}(A_k(1+\mu)k\pi+B_k(3+\mu))\sin k\pi y. \tag{8}$$

The trigonometric sets of functions included in (8) are orthogonal and form a basis on the semi-strip's end face. Therefore, there aren't problems with determining of the coefficients A_k, B_k of expansions. Thus the well-known Filon–Ribiere solutions are obtained.

The Fadle–Papkovich functions are much more difficult than trigonometric series: they are complex-valued and non-orthogonal. But their main feature is that they do not form basis on the segment on which the expanded functions (end faces of the rectangle or the semi-strip) are set. Therefore, it is impossible to obtain explicit formulas for the coefficients of expansions in terms of the Fadle–Papkovich functions on the basis of the classical methods of the theory of the basis of functions. Numerous methods have been proposed for the determination of the coefficients of expansions in terms non-orthogonal sets of the Fadle–Papkovich functions. However, almost all of them, anyway, were reduced to an approximate determination of the unknown coefficients from the infinite, not reducible system of algebraic equations.

The basis of this method is a generalization of the classical notion of the basis of functions on the segment. If the classical basis on the segment can be considered as a basis in the complex plane, the Fadle–Papkovich functions form a basis on the Riemann surface of the logarithm. Therefore, it is possible to construct biorthogonal sets of functions, which are determined on the Riemann surface of the logarithm and, then, to find the coefficients of expansions. The Fadle–Papkovich functions are generalization of systems of exponential functions with complex exponents. A. F. Leontiev, and then Yu. F. Korobeinik were studied the basis properties of exponential functions' systems with complex exponents and were investigated that they do not form a basis in the classical sense, and the expansions on them are not unique. Korobeinik has named such function systems as the representing function systems. The base for study of basis properties of exponential functions' systems is the theory of entire functions of exponential type and Borel transform in this class of functions. The base

for study of basis properties of the Fadle–Papkovich functions is the theory of quasi-entire functions of exponential type and Borel transform in this class of functions.

The quasi-entire functions of exponential type were first introduced by A. Pfluger in [4] in 1936 and it is the only work of quasi-entire functions of exponential type. Based on the Pfluger's results, it is possible to construct systems of functions which biorthogonal to the Fadle–Papkovich functions, and then exactly determine the expansion coefficients, thereby to construct an exact solution of the problem. Such a solution was first published in [5].

The solutions that are obtained using of functions' biorthogonal sets are called exact because they are constructed on the same scheme as the Filon–Ribiere solutions, and the desired coefficients are explicitly determined (and not as the solution of the infinite system of algebraic equations). Moreover, if the complex eigenvalues in obtaining exact solutions of the boundary-value problem are aim to the numbers $k\pi$ $(k = 1,2,...)$, the Fadle–Papkovich functions turn into the usual trigonometric functions, and the solutions – into the Filon–Ribiere solutions.

Since the biorthogonal sets of functions are determined on a certain infinite curve lying on the Riemann surface of the logarithm, it is necessary to continue the expanded functions initially defined on the segment (end face of the semi-strip) outside the segment on this curve. Because such continuations are not unique, the solutions of the boundary-value problems in the semi-strip are not unique as well. In 1996 academician E.I. Shemyakin first drew attention to the non-uniqueness of boundary value problems' solutions of elasticity theory in finite domains with boundary's angular points and changing type's points of the boundary conditions. As the solutions are not unique, the nonzero solutions do exist despite the fact that the load on the sides of the semi-strip is equal to zero. Such solutions are called eigensolutions, and the corresponding stresses – residual or initial. Constructing exact solutions in the semi-strip or in a rectangle, we do obtained the theory describing the residual stresses (one of the most important problems of elasticity theory). The residual stresses satisfy the equilibrium equations and the boundary conditions, but do not satisfy the strain compatibility conditions and so the displacements are not uniquely determined. To understand the physics of this known statement, let us cut the rectangle $\{P : |x| \le a, |y| \le 1\}$ along the axis y. Replace the resulting right and left rectangles by almost rectangles that if a certain load is applied to them, they take their initial shape of rectangles. Let us apply the required load and then glue together the rectangles. As a result we will obtain the rectangle P, where there is a nontrivial field of residual stresses such that the relations $\sigma_x(0,y) = \sigma(y)$ and $\tau_{xy}(0,y) = \tau(y)$ are satisfied on the glue line. The fact that the displacements are not determined uniquely means the following: there are infinite numbers of profiles sides of almost rectangles, connected into the original rectangle, after gluing such profiles the normal and tangential stresses are the same on the glue line.

The eigensolutions have the property that the boundary conditions are strictly satisfied for them on rectilinear boundaries of the rectangle. Hereby they are fundamentally different from the solutions in the classical statement, in which "boundary conditions are conventionally moved to the nondeformed surface", i.e., either the missing material is added, or the extra material is removed. In 1940 D.I. Sherman proposed a more strict understanding of boundary conditions. Let us consider the content of understanding in the following example. Suppose that the semi-strip $\{|y| \le 1,\ x \ge 0\}$ is given, the longitudinal sides of which are free, and normal stresses $\sigma_x(0,y) = \sigma(y)$ are set on the end face (for simplicity, we assume that the tangential stresses are equal to zero, and the function $\sigma(y)$ is even). Following Sherman understanding, we can consider an infinite plane $\{|x,y| < \infty\}$, in which let us draw two

horizontal cuts $\{y=\pm 1,|\,x\,|<\infty\}$ and the vertical cut $\{x=0,|\,y\,|\le 1\}$. Let us make insertions along a cut of the same material as that of the plane material, which continuously glue with the sides of the cuts. Then the shapes of insertions can be chosen so that the normal and tangential stresses are equal to zero on axes of the horizontal insertion, and we have such relations on an axis of the vertical insertion: $\sigma_x(0,y)=\sigma(y),\ \tau_{xy}(0,y)=0$. At first sight this understanding of boundary conditions on the sides of the semi-strip seems more stricter, but actually nothing has changed since and in this case we have to "add" or "remove" material. As a result, the boundaries of the semi-strip are rectilinear before the deformation and remain rectilinear after the deformation. Due to this condition the strain compatibility conditions are satisfied, both in the classic sense of the boundary conditions, and acc. Sherman.

In solving the boundary value problems the need to continue the boundary functions set on the end faces of the rectangle P outside the segment, in which they were initially determined, is physically quite understandable. In fact, within the framework of classical concepts of mechanics of solids, not only values of boundary functions, but all their derivatives, both along the axis x, and along the axis y must be defined at each point of the rectilinear sides of the rectangle, including in its angular points (the same belongs to changing type's points of the boundary conditions). Therefore, for a correct formulation of the boundary-value problem in the rectangle it is necessary to specify: 1) how the boundary functions set on the end faces of the rectangle continues to the whole infinite straight line $|\,y\,|<\infty$ and 2) how the solution continues through the end faces of the rectangle. The necessity for the continuation of the boundary functions on the axis y is dictated by the fact that sets of functions biorthogonal to the Fadle–Papkovich functions are not finite. Let us explain the second point. Consider the same semi-strip Π^+ with free longitudinal sides, when on the end face of the semi-strip

$$\sigma_x(0,y)=\sigma(y),\ \tau_{xy}(0,y)=0,\ (|\,y\,|<1)\,. \tag{9}$$

If the solution from the right semi-strip continues to the same left semi-strip, i.e. in the semi-strip with free longitudinal sides, the stresses are finite at the angular points of the semi-strip. And if the solution continues to the left semi-strip, the longitudinal sides of which are rigidly fixed, but the conditions (9) are still satisfied, the stresses may have integrable singularities at the angular points $x=0,\ |\,y\,|=\pm 1$. The conditions, under which the singularity exists or not exist, are strictly established by the solution of the corresponding mixed boundary-value problem in a rectangle.

In conclusion, we give formulas for stresses in the semi-strip, when only the normal self-balanced stresses $\sigma_x(0,y)=\sigma(y)$ are set on its end face, and the tangential stresses are equal to zero

$$\begin{aligned}
\sigma_x(x,y)&=\sum_{k=1}^{\infty}2\,\mathrm{Re}\left\{\sigma_k\frac{s_x(\lambda_k,y)}{M_k}\frac{\mathrm{Im}(-\overline{\lambda_k}e^{\lambda_k x})}{\mathrm{Im}(\lambda_k)}\right\},\\
\sigma_y(x,y)&=\sum_{k=1}^{\infty}2\,\mathrm{Re}\left\{\sigma_k\frac{s_y(\lambda_k,y)\lambda_k\overline{\lambda_k}}{M_k\lambda_k^2}\frac{\mathrm{Im}(-\lambda_k e^{\lambda_k x})}{\mathrm{Im}(\lambda_k)}\right\},\\
\tau_{xy}(x,y)&=\sum_{k=1}^{\infty}2\,\mathrm{Re}\left\{\sigma_k\frac{t_{xy}(\lambda_k,y)}{\lambda_k M_k}\lambda_k\overline{\lambda_k}\frac{\mathrm{Im}(-e^{\lambda_k x})}{\mathrm{Im}(\lambda_k)}\right\},
\end{aligned} \tag{10}$$

where

$$\sigma_k = \int_{-1}^{1} \sigma(y) \frac{\cos \lambda_k y}{2(1+\mu)\lambda_k \sin \lambda_k} dy,\ M_k = \frac{L'(\lambda_k)}{2} = \cos^2 \lambda_k .$$

This research was supported by the Russian Foundation for Basic Research, research project nos. 16-31-60028 mol_a_dk and 15-41-02644 r_povolzhe_a.

REFERENCES

[1] M.D. Kovalenko, T.D. Shulyakovskaya, Expansion in Fadle-Papkovich Functions in a Strip. Theory Foundations. *Mechanics of Solids*, **46(5)**, 721-738, 2011.

[2] M.D. Kovalenko, I.V. Menshova, T.D. Shulyakovskaya, Expansions in Fadle-Papkovich functions: Examples of solutions in a half-strip. *Mechanics of Solids*, **48(5)**, 584-602, 2013.

[3] V.V. Meleshko, Selected topics in the history of two-dimensional biharmonic problem, *Appl. Mech. Rev.*, **56(1)**, 33-85, 2003.

[4] A. Pfluger, Über eine Interpretation gewisser Konvergenz- und Fortsetzungseigenschaften Dirichletscher Reihen. *Comment. Math. Helv.*, **8**, 89-129, 1935/36.

[5] M.D. Kovalenko, S.V. Shibirin, A half-strip under the action of concentrated force: an exact solution to the problem. *Doklady Physics*, **42(10)**, 289-294, 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DISCRETE VELOCITY MODELS: A STUDY OF THE HYDRODYNAMIC LIMIT

Hans Babovsky

Technische Universität Ilmenau
D-98684 Ilmenau, Germany
e-mail: hans.babovsky@tu-ilmeau.de

Keywords: Discrete Velocity Models, diffusive scaling, hydrodynamic limit, closure relations.

Abstract. *We investigate Discrete Velocity Models as a numerical tool for the simulation of rarefied flows in the hydrodynamic limit. The study is based on the analysis of the bifurcation structure of a related linearized transport operator. In combination with a meaningful scaling this produces explicit formulas for the transport coefficients and closure relations depending on details of the model collision operator. The theory is applicable even in cases when the classical asymptotic analysis approach fails or leads to physically irrelevant results [7, 8].*

1 INTRODUCTION

Discrete Velocity Models (DVM) are a promising alternative to the commonly used DSMC methods for the numerical simulation of rarefied gas flows. While the latter should keep for the next time their prominent position as long as real gas effects are concerned, the first ones are able to provide new insights into questions e.g. concerning the hydrodynamic limit. This is on one hand due to the impressing development of powerful yet achievable hard ware tools (GPU, GPU). On the other hand have recent contributions taught how to interpret DVMs as numerical integration tools for the Boltzmann collision operator [1, 2] which allows to tune such models with respect to order of consistency and to numerical efficiency [3].

In [4] a formalism was introduced for DVM's based on two-particle collisions which provides an analytic expression for the pseudo inverse of the linearized collision operator and allows to calculate Navier-Stokes corrections to the Euler equations in the hydrodynamic limit. In this context it was argued why at least small velocity grids can reveal an appropriate macroscopic description only in the case of small Mach numbers. In the present paper we study the behaviour of DVM's at small bulk velocities. We identify a bifurcation phenomenon of the linearized steady transport operator when perturbing the Mach zero Maxwellian. This feature yields an appropriate macroscopic description when applying the correct scaling limit. The approach is constructive in the sense that it provides explicit formulas for hydrodynamic flow parameters and closure relations for the moment hierarchy. In terms of an evaporation condensation problem for a binary gas mixture we shortly discuss the difference between the present approach and some standard asymptotic analysis techniques.

Though the following framework is applicale to three-dimensional velocity grids, we restrict for sake of brevity to the two-dimensional case.

2 THE COLLISION MODEL

We consider a DVM on a finite integer grid $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\} \subset \mathbb{R}^2$, $\mathbf{v}_i = (v_x^{(i)}, v_y^{(i)})$. The collision operator is composed of single collision events described a quadruple

$$\alpha = (i, j, k, l) \doteq (\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k, \mathbf{v}_l) \in \mathcal{V}^4$$

and the single-collision operator J_α for a density $\mathbf{f} = (f_r)_{r=1}^N \doteq (f(\mathbf{v}), \mathbf{v} \in \mathcal{V})$ given by

$$J_\alpha[\mathbf{f}, \mathbf{f}] = (f_k f_l - f_i f_j) \cdot \mathbf{s}_\alpha, \quad \mathbf{s}_\alpha = (\mathbf{e}_i + \mathbf{e}_j - \mathbf{e}_k - \mathbf{e}_l) \tag{1}$$

with $\mathbf{e}_r$ = r-th canonical unit vector. (For a more detailed introduction into this formalism see [4].) J_α satisfies the physical conservation laws for mass, momenta and energy, iff $\overline{\mathbf{v}_i \mathbf{v}_j}$ and $\overline{\mathbf{v}_k \mathbf{v}_l}$ are the diagonals of a rectangle in $\mathbb{R}^2$. The set of quadruples α with this property is denoted by $\mathcal{R}$. The full collision operator reads

$$J[\mathbf{f}, \mathbf{f}] = \sum_{\alpha \in \mathcal{R}} \gamma_\alpha \cdot J_\alpha[\mathbf{f}, \mathbf{f}] \tag{2}$$

with collision frequencies $\gamma_\alpha \geq 0$. We assume the set of α with $\gamma_\alpha > 0$ large enough such that the besides mass, momenta and energy there are no other invariants. It is well-known that

in this case the classical H-Theorem is satisfies, and as a consequence the set of equilibrium functions of J is equal to the set of Maxwellians

$$\mathbf{f} = \mathbf{f}[\rho, \overline{\mathbf{v}}, \Theta] = \left(\rho \exp\left(-\frac{|\mathbf{v} - \overline{\mathbf{v}}|^2}{2\Theta} \right), \mathbf{v} \in \mathcal{V} \right) \tag{3}$$

The linearization of J around a Maxwellian $\mathbf{f}$ is given by the operator $J[\mathbf{f} + \phi, \mathbf{f} + \phi]$ neglecting the terms quadratic in ϕ. Taking into account the identity $f_i f_j = f_k f_l =: \phi_\alpha$ for $\alpha = (i, j, k, l) \in \mathcal{R}$ and defining the matrices

$$P_\alpha = (\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k, \mathbf{e}_l), \quad \Gamma = \mathbb{1}_\pm \cdot \mathbb{1}_\pm^T, \quad \mathbb{1}_\pm = (1, 1, -1, -1)^T$$

the linearization takes the form

$$J[\mathbf{f}, \phi] + J[\phi, \mathbf{f}] = CF^{-1} \tag{4}$$

with $F = \text{diag}(f_i, i = 1 \dots N)$ and

$$C = -\sum_{\alpha \in \mathcal{R}} \pi_\alpha P_\alpha \Gamma P_\alpha^T, \quad \pi_\alpha = \gamma_\alpha \phi_\alpha \tag{5}$$

The main properties of C are easy to prove [4] and given as follows.

Lemma: *C is symmetric and negative semidefinite. The geometric zero eigenspace has dimension 4 and is given by*

$$\mathcal{M} = \textit{span}\left(\mathbb{1} = (1, \dots, 1)^T, \mathbf{v}_x = (v_x^{(i)})_{i=1}^N, \mathbf{v}_y = (v_y^{(i)})_{i=1}^N, |\mathbf{v}| = (|\mathbf{v}_i|^2)_{i=1}^N \right) \tag{6}$$

Much of the following is based on symmetry arguments. We assume the grid $\mathcal{V}$ and the collision model on $\mathcal{V}$ to be invariant with respect to reflections at $v_x = 0$ resp. $v_y = 0$. Precisely this means the following. Given $\mathbf{v} = (v_x, v_y) \in \mathcal{V}$, define $T_x\mathbf{v} = (-v_x, v_y)$ and $T_y = (v_x, -v_y)$. T_x- and T_y-*invariance* of $\mathcal{V}$ means that

$$\mathbf{v} \in \mathcal{V} \quad \Leftrightarrow \quad T_x\mathbf{v} \in \mathcal{V} \quad \Leftrightarrow \quad T_y\mathbf{v} \in \mathcal{V}$$

Given $\mathbf{f} \in \mathbb{R}^N$ (as a function on $\mathcal{V}$), define $T_x\mathbf{f} \in \mathbb{R}^N$ by $T_x\mathbf{f}(\mathbf{v}) = \mathbf{f}(T_x\mathbf{v})$, and similarly $T_y\mathbf{f}$. T_x- and T_y-*invariance* of the collision operator $J[\mathbf{f}, \mathbf{f}]$ means that $J[T_x\mathbf{f}, T_x\mathbf{f}] = T_x J[\mathbf{f}, \mathbf{f}]$ and $J[T_y\mathbf{f}, T_y\mathbf{f}] = T_y J[\mathbf{f}, \mathbf{f}]$.

$\mathbf{f}$ is called T_x-*even*, if $\mathbf{f} = T_x\mathbf{f}$, and T_x-*odd*, if $\mathbf{f} = -T_x\mathbf{f}$. Similarly, T_y-*even* and T_y-*odd* are defined.

3 STEADY PROBLEMS AND BIFURCATION

Consider the steady one-dimensional problem

$$v_x \partial_x f = J[f, f] \tag{7}$$

on $[-1, 1]$ with boundary conditions at $x = \pm 1$. In the case of totally reflecting boundaries, the flux into wall direction is zero. Consider a small perturbation $\mathbf{f}_0 + \phi$, $\phi \perp F_0\mathcal{M}$, around a centered Maxwellian

$$\mathbf{f}_0(v) = \mathbf{f}[\rho, 0, \Theta](v)$$

Neglecting the terms quadratic in ϕ yields the linearized system

$$v_x \partial_x \phi = J[\mathbf{f}_0, \phi] + J[\phi, \mathbf{f}_0] = C F_0^{-1} \phi \tag{8}$$

We assume that $\mathcal{V}$ contains no velocities with $v_x = 0$. In this case, because of T_x-invariance, $N = 2n$ has to be even. We choose a numbering of the velocities $\mathbf{v}_i = (v_x^{(i)}, v_y^{(i)})$ such that $v_x^{(i)} > 0$, and $\mathbf{v}_{n+i} = T_x \mathbf{v}_i$ for $i = 1, \ldots, n$. Under this restriction, the linearized equation represents an ODE system

$$\partial_x \phi = L_0 \phi \tag{9}$$

Under reasonable conditions on the collision model (see [5]), the linearized collision operator

$$L_0 = V^{-1} C F_0^{-1} \tag{10}$$

has a Jordan normal form given by

$$N = \mathrm{diag}(0, 0, N_0, N_0, \Lambda, -\Lambda)$$

with a positive diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_p)$ and the Jordan block

$$N_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

The columns $\mathbf{t}_i^{(0)}$ of the corresponding transformation matrix $T^{(0)}$ are characterized as follows.

- $\mathbf{t}_1^{(0)} = F_0 v_x$ is T_x-odd and zero-eigenvector of L.
- $\mathrm{span}\{\mathbf{t}_2^{(0)}, \mathbf{t}_3^{(0)}\} = F_0 \cdot \mathrm{span}\{\mathbb{1}, |v|^2\}$; in particular, $\mathbf{t}_2^{(0)}$ and $\mathbf{t}_3^{(0)}$ are T_x-even zero-eigenvectors of L_0.
- $\mathbf{t}_4^{(0)}$ is T_x-odd and solution of $L_0 \mathbf{t}_4^{(0)} = \mathbf{t}_3^{(0)}$.
- $\mathbf{t}_5^{(0)} = F_0 v_y$ is T_x-even and T_y-odd and zero-eigenvector of L.
- $\mathbf{t}_6^{(0)}$ is T_x- and T_y-odd and solution of $L_0 \mathbf{t}_6^{(0)} = \mathbf{t}_5^{(0)}$.
- $L_0 \mathbf{t}_{6+i}^{(0)} = \lambda_i \mathbf{t}_{6+i}^{(0)}$, $i = 1, \ldots, p$.
- $L_0 \mathbf{t}_{6+p+i}^{(0)} = -\lambda_i \mathbf{t}_{6+p+i}^{(0)}$, $i = 1, \ldots, p$.

Thus the general solution ϕ of the ODE system is

$$\begin{aligned} \phi(x) \;=\; & c_4(\mathbf{t}_4^{(0)} + x \cdot \mathbf{t}_3^{(0)}) + c_5(\mathbf{t}_6^{(0)} + x \cdot \mathbf{t}_5^{(0)}) \\ & + \exp\left((1+x)\Lambda\right) \phi_+(-1) + \exp\left((1-x)\Lambda\right) \phi_-(+1) \end{aligned} \tag{11}$$

The sum $F = c_4 \mathbf{t}_4^{(0)} + c_6 \mathbf{t}_6^{(0)}$ is the socalled *fluctuation part* which is mapped to the zero eigenspace of L. The constant gradient part $x \cdot (c_4 \mathbf{t}_3^{(0)} + c_6 \mathbf{t}_5^{(0)})$ is called the *macroscopic part*, while the last two terms represent the *boundary layers*. From now on we restrict to T_y-even solutions, putting $c_5 = 0$.

The solvability condition for $L_0\mathbf{t}_4^{(0)} = \mathbf{t}_3^{(0)}$ is that $\mathbf{t}_3^{(0)}$ is orthogonal to the kernel of L_0^T, in particular to $\mathbf{v}_x^2$. This means that we may choose $\mathbf{t}_3^{(0)} = F_0(\mathbb{1} + \xi \cdot |\mathbf{v}|^2)$ with

$$\xi = \frac{\langle \mathbf{v}_x^2, f_0 \rangle}{\langle \mathbf{v}_x^2, f_0 |\mathbf{v}|^2 \rangle}$$

We now introduce a small shift by changing the Maxwellian into

$$f_\delta(\mathbf{v}) = \exp(-|\mathbf{v} - \delta \cdot \mathbf{e}_x|^2/2\Theta) \tag{12}$$

and the corresponding matrix F_0 into

$$F_\delta(\mathbf{v}) = \mathrm{diag}(f_\delta(\mathbf{v}), \mathbf{v} \in \mathcal{G}) = F_0 \cdot \left(I + \frac{\delta}{\Theta} V_x\right) + \mathcal{O}(\delta^2) \tag{13}$$

In the following we investigate the bifurcation structure of the zero eigenspace. Thus we are interested only in infinitesimal shifts and restrict an $\mathcal{O}(\delta)$ theory. The linearized collision operator L_0 is affected twofold by this change and takes the form

$$L_\delta \doteq V_x^{-1}(C + \delta \cdot C')(F_0 + \delta \cdot F')^{-1} \doteq L_0 + \delta \cdot V_x^{-1} C' F_0 - \delta \cdot V_x^{-1} C F' \tag{14}$$

The first part of the perturbation, $\delta \cdot V_x^{-1} C' F_0$ stems from the change of the factor ϕ_α in (5). This part does not contribute to the bifurcation structure and is neglected here. Thus we obtain as the new linearized collision operator

$$L_\delta \doteq L_0 \cdot \left(I - \frac{\delta}{\Theta} V_x\right) \tag{15}$$

The perturbation changes the nullspace into

$$F_\delta \mathcal{M} \doteq F_0 \left(I + \frac{\delta}{\Theta} V_x\right) \mathcal{M}$$

The eigenspace data corresponding to nonzero eigenvectors suffer an analytic change and do not influence the qualitative structure of the Jordan normal form. The main change results from the non-solvability of the equation $L_\delta \mathbf{t}_4^\delta = \mathbf{t}_3^\delta$. It has to be replaced with the equation

$$L_\delta \mathbf{t}_4^\delta \;\; = \;\; \mathbf{t}_3^\delta + \frac{\delta \cdot \lambda}{\Theta} \cdot \mathbf{t}_4^\delta \tag{16}$$

The choice

$$\mathbf{t}_3^\delta = \left(I + \frac{\delta}{\Theta} V_x\right) \mathbf{t}_3^0 + \frac{\delta \cdot \mu}{\Theta} V_x \mathbf{f}_0 \tag{17}$$

and the ansatz $\mathbf{t}_4^\delta = \mathbf{t}_4^0 + \delta \Theta^{-1} \mathbf{t}_4'$ leads to the equation

$$L_0 \mathbf{t}_4' = V_x \mathbf{t}_3^0 + L_0 V_x \mathbf{t}_4^0 + \lambda \mathbf{t}_4^0 + \mu V_x \mathbf{f}_0 \tag{18}$$

for the T_x-even function $\mathbf{t}_4'$ with the solvability condition

$$\begin{pmatrix} \langle \mathbf{v}_x, \mathbf{t}_4^0 \rangle & \langle \mathbf{v}_x^2, \mathbf{f}_0 \rangle \\ \langle \mathbf{v}_x |\mathbf{v}|^2, \mathbf{t}_4^0 \rangle & \langle \mathbf{v}_x^2 |\mathbf{v}|^2, \mathbf{f}_0 \rangle \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = - \begin{pmatrix} \langle \mathbf{v}_x^2, \mathbf{t}_3^0 \rangle + \langle \mathbf{v}_x, L_0 V_x \mathbf{t}_4^0 \rangle \\ \langle \mathbf{v}_x^2 |\mathbf{v}|^2, \mathbf{t}_3^0 \rangle + \langle \mathbf{v}_x |\mathbf{v}|^2, L_0 V_x \mathbf{t}_4^0 \rangle \end{pmatrix} \tag{19}$$

for (λ, μ).

As a consequence, choosing $\overline{v} := \delta \Theta^{-1}$, the corresponding Jordan block N_0 changes into

$$N_\delta = \begin{pmatrix} 0 & 1 \\ 0 & \lambda \overline{v} \end{pmatrix} \tag{20}$$

which is similar to the diagonal matrix $\mathrm{diag}(0, \lambda \overline{v})$. The following result is an immediate consequence of this.

Theorem: *(a) The solution of the perturbed ordinary differential system takes the form*

$$\begin{aligned} \phi(x) &= c_4 \cdot \left(\exp(\lambda \overline{v} x) \mathbf{t}_4 + \frac{\exp(\lambda \overline{v} x) - 1}{\lambda \overline{v}} \cdot \mathbf{t}_3 \right) \\ &\quad + \exp\left((1+x) \Lambda_\delta^+ \right) \phi_+(-1) + \exp\left((1-x) \Lambda_\delta^- \right) \phi_-(+1) \end{aligned}$$

(b) The fluxes of moment vectors $\mathbf{m} \in \mathcal{M}$ are constant and given by

$$\left\langle \mathbf{v}_x \mathbf{m}, \phi \right\rangle = c_4 \langle \mathbf{v}_x \mathbf{m}, \mathbf{t}_4^\delta \rangle$$

4 DIFFUSIVE SCALING AND CLOSURE RELATIONS

In order to derive a meaningful macroscopic limit we introduce the diffusive scaling (see, e.g. [6]) for the equation

$$(\partial_t + v_x \partial_x) \phi = C F^{-1} \phi \tag{21}$$

It consists in replacing the macroscopic variables t and x with $\epsilon^{-2} t$ and $\epsilon^{-1} x$ and leads to the rescaled equation

$$(\partial_t + \epsilon^{-1} v_x \partial_x) \phi = \epsilon^{-2} C F^{-1} \phi \tag{22}$$

Formally this is equivalent to replacing the space $\mathcal{V}$ of microscopic velocities with $\epsilon^{-1} \mathcal{V}$ and scaling up the collision frequency by a factor ϵ^{-2}. This is the approach which we take here.

Replacing $\mathbf{v}_i$ with $\mathbf{w}_i = \epsilon^{-1} \mathbf{v}_i$ requires to change the Maxwellians
$\mathbf{f}_\delta = (\exp(-|\mathbf{v}_i - \delta \mathbf{e}_x|^2 / 2\Theta), i = 1 \dots N)^T$ to
$\mathbf{f}_\delta^{(\epsilon)} = (\exp(-|\mathbf{w}_i - \delta \mathbf{e}_x|^2 / 2T), i = 1 \dots N)^T = (\exp(-|\mathbf{v} - \epsilon \delta \mathbf{e}_x|^2 / 2\epsilon^2 T), i = 1 \dots N)^T$
(leaving the macroscopic bulk velocity $\overline{\mathbf{v}}$ unchanged) which itself makes only sense if we rescale the temperature as $T = \epsilon^{-2} \Theta$. From now on we define

$$\mathbf{f}_\delta^{(\epsilon)} = \mathbf{f}_{\epsilon\delta}^{(1)} = (\exp(-|\mathbf{v} - \epsilon \delta \mathbf{e}_x|^2 / 2\Theta), \mathbf{v} \in \mathcal{V})^T \tag{23}$$

with $\Theta > 0$ constant. We write $\delta =: \Theta \overline{w}$.

Associated to $\mathbf{f}_{\Theta \overline{w}}^{(\epsilon)}$ are the moments

$$\begin{array}{ll} \textit{density} & \rho^{(\epsilon)}_{\Theta\overline{w}} = \langle \mathbb{1}\,, \mathbf{f}^{(1)}_0\rangle + \mathcal{O}(\epsilon^2) \doteq \langle \mathbb{1}\,, \mathbf{f}^{(1)}_0\rangle \\ \textit{flux} & \phi^{(\epsilon)}_{\Theta\overline{w}} = \langle \mathbf{w}_x, \mathbf{f}^{(\epsilon)}_{\Theta\overline{w}}\rangle = \overline{w}\cdot\langle \mathbf{v}_x^2, \mathbf{f}^{(1)}_0\rangle + \mathcal{O}(\epsilon^2) \doteq \overline{w}\cdot\langle \mathbf{v}_x^2, \mathbf{f}^{(1)}_0\rangle \end{array}$$

while temperature T given by $\rho T \sim \langle |\mathbf{w}|^2 \mathbf{f}^{(\epsilon)}_{\Theta\overline{w}}\rangle \nearrow \infty$

The rescaled steady problem reads

$$\partial_x \phi = \epsilon^{-1} L^{(\epsilon)}_{\Theta\overline{w}}\phi = \epsilon^{-1} L^{(1)}_{\epsilon\Theta\overline{w}}\phi \tag{24}$$

Repeating the above calculations for the rescaled problem leads to the following result which only concerns the T_y-even solutions.

Theorem: *(a) For $\epsilon \searrow 0$, the general T_y-even solution of the ODE system takes the form*

$$\begin{aligned} \phi(x) &= c_4 \cdot \left(\epsilon \cdot \exp(\overline{w}\lambda x)\mathbf{t}^0_4 + \mathcal{O}(\epsilon^2) + \frac{\exp(\overline{w}\lambda x) - 1}{\overline{w}\lambda} \cdot \mathbf{t}^{\epsilon\Theta\overline{w}}_3 \right) \qquad (25) \\ &\quad + \exp\left((1+x)\epsilon^{-1}\Lambda^+_{\overline{w}}\right)\phi_+(-1) + \exp\left((1-x)\epsilon^{-1}\Lambda^-_{\overline{w}}\right)\phi_-(+1) \\ &\approx c_4 \cdot \left(\epsilon(1+\overline{w}\lambda x)\mathbf{t}^0_4 + x\cdot[\mathbf{t}^0_3 + \epsilon\overline{w}V_x(\mathbf{t}^0_3 + \mu\mathbf{f}_0)]\right), \qquad \textit{for } |x| \ll 1 \qquad (26) \end{aligned}$$

$\mathbf{t}^0_i$, λ, μ are the quantities of the preceding section.
(b) Let $\mathbf{n} = (n_x, n_y)$ be a multiindex and $\mathbf{v}^{\mathbf{n}} = \mathbf{v}_x^{n_x}\mathbf{v}_y^{n_y}$ the $\mathbf{n}$-th moment vector. In the limit $\epsilon = 0$, the macroscopic moment is given in first order of $\overline{w}$ as

$$\langle \mathbf{w}^{\mathbf{n}}, \phi\rangle = c_4\epsilon^{|\mathbf{n}|}\cdot \begin{cases} x\cdot\langle \mathbf{v}^{\mathbf{n}}, \mathbf{t}^0_3\rangle & n_x \textit{ even} \\ \epsilon\langle \mathbf{v}^{\mathbf{n}}, \mathbf{t}^0_4\rangle + \epsilon\overline{w}\lambda x\langle \mathbf{v}^{\mathbf{n}}, \mathbf{t}^0_4 + V_x\mathbf{t}^0_3 + \mu V_x\mathbf{f}_0\rangle & n_x \textit{ odd} \end{cases} \tag{27}$$

From this follow easily closure relations betwen moments and their fluxes. For example, the heat coefficient (as the ratio between heat flux and temperature gradient) is given as

$$q = \langle \mathbf{v}_x|\mathbf{v}|^2, \mathbf{t}^{\mathbf{0}}_{\mathbf{4}}\rangle / \langle |\mathbf{v}|^{\mathbf{2}}, \mathbf{t}^{\mathbf{0}}_{\mathbf{3}}\rangle \tag{28}$$

The above theory is easily extendable to related situations like the following example.

Example: *The system*

$$v_x\partial_x\mathbf{g} = J[\mathbf{f},\mathbf{g}], \quad v_x\partial_x\mathbf{h} = J[\mathbf{f},\mathbf{h}], \quad \mathbf{f} = \mathbf{g} + \mathbf{h} \tag{29}$$

describes a binary gas mixture of two mechanically identical species A and B. Suppose A represents vapor which can evaporate or condensate at the walls, while B is totally reflected thus producing zero flux between the wall. Introducing a pressure difference between the walls induces a flow of A from one wall to the other. Classical asymptotic analysis produces an anomaly ("ghost effect") in the form of infinitesimally thin boundary layers of species B completely stopping the flow of A [7, 8]. Instead, the above scaling procedure leads to a boundary layer of finite thickness slowing down the flow of A depending on the concentration of B. The reason for the differences lies in the fact that the formal asymptotic expansion procedure in [7, 8] produces the wrong equilibrium state since it is not capable of the arising bifurcation mode. For details, see [9].

5 CONCLUSIONS

We have investigated symmetric DVMs at small Mach numbers. A typical feature is a bifurcation phenomenon of the steady linearized transport operator at zero bulk velocity. A detailed analysis allows to derive explicit formulas for the closure relations of the moment system in the hydrodynamic limit.

REFERENCES

[1] F. Rogier, J. Schneider. A direct method for solving the Boltzmann equation. *Transp. Theory Stat. Phys.*, 23:313–338, 1994.

[2] A. Palczewski, J. Schneider and A. V. Bobylev. A consistency result for a discrete velocity model of the Boltzmann equation. *SIAM J. Numer. Anal.*, 34:1865–1883, 1997.

[3] Stefan Brechtken. Classification of Lattice Group Models, High Order Discretizations of Boltzmann's Collision Operator and Parallelization. Dissertation, Technische Universität Ilmenau, 2015.

[4] H. Babovsky. Discrete kinetic models in the fluid dynamic limit. *Computers and Mathematics with Applications*, 67:256–271, 2014.

[5] H. Babovsky and T. Płatkowski, Kinetic boundary layers for the Boltzmann equation on discrete velocity lattices. *Arch. Mech.*, 60:87–116, 2003

[6] H. Babovsky and P. Kowalczyk. Diffusion limits for discrete velocity models in a thin gap. *Multiscale Modelling and Simulation*, 6:631–655, 2007.

[7] S. Takata and K. Aoki. The ghost effect in the continuum limit for a vapor-gas mixture around condensed phases: Asymptotic analysis of the Boltzmann equation. *Transport Theor. Statist. Phys.*, 30:205–237, 2001.

[8] K. Aoki, S. Takata, and S. Taguchi. Vapor flows with evaporation and condensation in the continuum limit: effect of a trace of noncondensable gas. *Eur. Journ. Mech. B/Fluids*, 22:51–71, 2003.

[9] Hans Babovsky. Macroscopic limit for an evaporation-condensation prolem. Preprint, Technische Universität Ilmenau, 2015, submitted.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MESO - MACRO MODELS FOR A HARD SPHERE GAS

S.V. Bogomolov[1], **N.B. Esikova**[1], **and A.E. Kuvshinnikov**[1]

[1]M.V. Lomonosov Moscow State University
e-mail: bogomo@cs.msu.su
e-mail: esikova.nata@yandex.ru
e-mail: kuvsh90@yandex.ru

Keywords: Boltzmann equation, Kolmogorov – Fokker – Planck equation, Navier – Stokes equation; random processes, stochastic differential equations with respect to Poisson and Wiener measures, discontinuous particle method.

Abstract. *Macroscopic system of gas dynamic equations, differing from Navier – Stokes and quasi gas dynamic ones, is derived from a stochastic microscopic model of a hard sphere gas in a phase space. The model is diffusive in velocity space and valid for moderate Knudsen numbers. The main peculiarity of our derivation is more accurate velocity averaging due to analytical solving stochastic differential equations with respect to Wiener measure which describe our original meso model. It is shown at an example of a shock wave front structure that our approach leads to larger than Navier – Stokes front widening that corresponds to reality. The numerical solution is performed by a (well suited to high performance computer applications) special "discontinuous" particle method.*

1 INTRODUCTION

A description of gas dynamic phenomena at the base of hierarchies of micro – macro models has become a long time ago a classical part of theoretical physics as well as a foundation for high performance industrial calculations [1]. In the last years more attention was drawn to "meso" models in the phase space. That models are often called Kolmogorov – Fokker – Planck equations. They are used by theoreticians [2] and applied mathematicians [3, 4]. The last ones mostly implement the models of so-called maxwellian molecules, not hard sphere ones, that lead to considerably different results [5].

We consider a model valid, on our opinion, at moderate Knudsen numbers (Kn), transient between a molecular description and an imagination of a gas as continuous medium. Kn is a parameter of nondimensionalization depending on a space subdomain. Its physical meaning is a ratio of an average mean free pass to a character dimension of the subdomain. Our model [6, 7, 8, 9] is a system of stochastic differential equations (SDE) with respect to Wiener measure $dw(t)$ describing a movement of a particle ($x(t)$ is its position and $v(t)$ is its velocity) in the phase space at moderately small Kn:

$$dx(t) = v(t)dt, \qquad (1)$$
$$dv(t) = -\frac{1}{Kn}a(c)(v(t) - V)dt + \frac{1}{\sqrt{Kn}}\sigma(c)dw(t),$$

where c is an absolute value of the dimensionless heat velocity $\mathbf{c} \equiv v(t) - V$, $V(x,t)$ is a macroscopic velocity, the coefficients in the second equation (vector $\mathbf{a}(c) = a(c)\mathbf{c}$ and matrix $\sigma(c)$ will be determined later.

The realisations of that random process (the set of trajectories) generate a measure with a density $F(x,v,t)$ which satisfies an equation of Kolmogorov – Fokker – Planck's type:

$$\frac{\partial F}{\partial t} + \sum_{i=1}^{3}\frac{\partial v_i F}{\partial x_i} - \frac{1}{Kn}\sum_{i=1}^{3}\frac{\partial(a_i(F)(v_i - V_i)F)}{\partial v_i} = \frac{1}{Kn}\frac{1}{2}\sum_{i,j=1}^{3}\frac{\partial^2(\sigma_{ij}^2(F)F)}{\partial v_i \partial v_j}. \qquad (2)$$

We study that, computationally more efficient than Boltzmann equation, diffusive in velocity space gas model which is a link in a chain of multiscale algorithms based on micro – macro models depending on different subdomains (distinguished by their Knudsen numbers) of a whole problem under consideration. That model, from one side, is connected to microscopic model and, from the other side, leads to more accurate macroscopic equations [6, 7, 8, 9].

That equation is well known as an heuristic model Boltzmann equation with Fokker – Planck collision integral [10, 11, 12]. But before now its coefficients a, σ^2 and the limits of its applicability were not specified. In [8] it was shown that Boltzmann equation can be approximated by Kolmogorov – Fokker – Planck one (2) at moderate Knudsen numbers. Its coefficients are integrals in phase space representing moments of a jump random process describing molecules collisions formulized by means of stochastic integration with respect to Poisson measures [13]:

$$\mathbf{a}(x_1(t), v_1(t), t) = \int\int\int \mathbf{f}(\theta, x_1(s), v_1(s), \tilde{x}, \tilde{v})m(d\theta)Fd\tilde{x}d\tilde{v}, \mathbf{a} \equiv a(c)\mathbf{c}$$

$$\sigma^2(x_1(t), v_1(t), t) = \int\int\int f^2(\theta, x_1(s), v_1(s), \tilde{x}, \tilde{v})m(d\theta)Fd\tilde{x}d\tilde{v}.$$

Here $\mathbf{f}$ is a jump function (a molecule velocity increment as a result of a collision with another one), θ is a goal parameter, m is an intensity of random collision process, F is a distribution

function, v, v_1 are velocity vectors, x, x_1 are position vectors. For a gas of hard spheres and with assumption of jump function locality in x the integrals take the following form:

$$\mathbf{a}(x_1(t), v_1(t), t) = \frac{\pi}{2} \int_{\mathbf{R^3}} (v_1 - v) \mid v_1 - v \mid \delta(x - x_1) F dx dv,$$

$$\sigma_{ij}^2(x_1(t), v_1(t), t) = \frac{1}{4\sqrt{\pi}} \int_{\mathbf{R^3}} (\frac{1}{3} \mid v_1 - v \mid^3 + (v_i - v_{1i})(v_j - v_{1j}) \mid v_1 - v \mid) \delta(x - x_1) F dx dv,$$

$\delta(x)$ is is Dirac delta - function.

At small Kn for hard spheres gas (at an assumption of distribution function F local maxwellity and isotropy by thermal velocity c inside eight – foled integrals calculations of "drift" vector $\mathbf{a}$ and "diffusion" matrix σ^2 in velocity space), these coefficients in equations (1), (2) are obtained as follows:

$$\mathbf{a}(\mathbf{c}) = \frac{\mathbf{c}}{c} T^{1/2} \frac{\sqrt{\pi}}{4} [\sqrt{\pi} \mathrm{erf}(c)(2c^2 + 2 - \frac{1}{2c^2}) + e^{-c^2}(2c + \frac{1}{c})],$$

$$\begin{aligned} \sigma_{ij}^2(c) &= T^{3/2} \frac{\sqrt{\pi}}{4} [\delta_{ij} P(c) + c_i c_j S(c)], \\ P(c) &= \sqrt{\pi} \mathrm{erf}(c)(\frac{c^3}{3} + \frac{3}{2}c + \frac{3}{4c} - \frac{1}{8c^3}) + e^{-c^2}(\frac{c^2}{3} + \frac{4}{3} + \frac{1}{4c^2}), \\ S(c) &= \sqrt{\pi} \mathrm{erf}(c)(c + \frac{3}{2c} - \frac{3}{4c^3} + \frac{3}{8c^5}) + e^{-c^2}(1 + \frac{1}{c^2} - \frac{3}{4c^4}), \end{aligned} \tag{3}$$

T is dimensionless temperature, δ_{ij} is Kronecker symbol.

A square root σ of matrix σ^2 is found by the standard way with the implementation of orthonormal basis of eigenvectors:

$$\begin{aligned} \sigma_{ij}(c) &= T^{3/4} \pi^{1/4} / 2 [\delta_{ij} \sqrt{\lambda_2} + c_i c_j / c^2 (\sqrt{\lambda_1} - \sqrt{\lambda_2})], \\ \lambda_1 &= c^2 S(c) + P(c), \lambda_2 = \lambda_3 = P(c). \end{aligned}$$

The computations [14] have shown that these coefficients are quite adequate for gas description at moderate Knudsen numbers. So we are continuing to develop that model.

In the present paper we'll try to get more accurate, than in [8] and than Navier – Stokes system, of macroscopic gas dynamics equations in case of hard sphere gas. The higher in [8] accuracy is based on analytical solving system (1), more precisely, its version, simplified to a handwriting level.

Note also that our approach with the help of SDE technique differs from other approaches for obtaining gas dynamics equations connected to application of deterministic equations for distribution function in phase space [15, 16, 17, 18] as well as other hierarchical models [19]. The models like ours with the coefficients $\mathbf{a}$ and σ depending on velocity are in use by the physics theoreticians for the study of phenomena in turbulent flows (f. e. [20]).

2 SDE SYSTEM

We'll rearrange the system (1) in a way that makes it possible to get macroscopic equations keeping at the same time maximum of microscopic information. The system (1) is a system

of equations for unknown functions $x(t)$ and $v(t)$. Express $v(t)$ through $x(t)$ from the second equation and substitute it in the first one having got an equation only for $x(t)$.

That equation will give us macroscopic equations. To do it we need to make some simplifications. Analitical solving SDE is more difficult than solving ordinary differential equations. Only some successful examples are known. Let lead our equation to one of it.

The coefficients (3) at large c behave themselves as $\mathbf{a}(\mathbf{c}) \sim a_1 c\mathbf{c}, a_1 \equiv \pi/2$ and $\sigma \sim \sigma_1 c^{3/2}$ (we denote $\sigma_1 \equiv (2\sqrt{\pi}/(3\sqrt{3}))T^{1/4}$). To derive them to a form enabling to get an exact solution let us put $\sigma = k\sigma_1 c$, introducing a parameter k which can be taken, for example, so that one of the terms in our macro – model coincides with the thermodynamical equation of state. We'll do it later. Underline that the introduction of the parameter k was done to get an analytical solution, we do not need the thermodynamical equation of state – all the coefficients in our equations are obtained from the model of hard spheres. Moreover, let us restrict ourselves to the matrix σ in a diagonal form not taking into consideration the elements out of the diagonal because of their smallness in the assumption of the isotropy in velocity space of the distribution function and further averaging. Then the system (1) takes the form:

$$\begin{aligned} &dx_i = v_i dt, \\ &dv_i = -\frac{1}{Kn} a_{1i} c(v_i - V_i)dt + \frac{1}{\sqrt{Kn}} k\sigma_{1ij} c dw_j, \\ &x_i|_{t=0} = x_{i0}, \quad v_i|_{t=0} = v_{i0}, \quad i = 1, 2, 3, \end{aligned}$$

Here $c = \sqrt{c_1^2 + c_2^2 + c_3^2}$ is an absolute value of dimensionless thermal velocity. Because of its presence the second equation can still not be solved exactly, so let us continue our simplifications.

Let $c_i > 0$, then:

$$\begin{aligned} &dc_i = -\frac{1}{Kn} a_0 c_i^2 dt + \frac{\sigma_0}{\sqrt{Kn}} c_i dw(t), \\ &c_i|_{t=0} = c_{i0}, \quad i = 1, 2, 3, \end{aligned}$$

we've replaced c by $3\sqrt{c_i}$, consequently: $a_0 = \sqrt{3}a_{1i}$, $\sigma_0 = \sqrt{3}k\sigma_{1ij}$. For reducing transformations assume that $dc_i = dv_i$, instead of $dc_i = dv_i - \partial V/\partial t dt$ because of slowness of V changing in comparison to c. To take into account $\partial V/\partial t$ technically is not difficult but we'll not do it here.

The exact solution of this equation (see Appendix 1) has the form:

$$c_i = \frac{\exp\left(\frac{\sigma_0}{\sqrt{Kn}} w_t - \frac{\sigma_0^2}{2Kn} t\right)}{c_{0i}^{-1} + \frac{a_0}{Kn} \int_0^t \exp\left(\frac{\sigma_0}{\sqrt{Kn}} w_s - \frac{\sigma_0^2}{2Kn} s\right) ds},$$

$$v_i = V_i + \frac{\exp\left(\frac{\sigma_0}{\sqrt{Kn}} w_t - \frac{\sigma_0^2}{2Kn} t\right)(v_{0i} - V_{0i})}{1 + \frac{a_0(v_{0i} - V_{0i})}{Kn} \int_0^t \exp\left(\frac{\sigma_0}{\sqrt{Kn}} w_s - \frac{\sigma_0^2}{2Kn} s\right) ds}, \quad i = 1, 2, 3.$$

where $w_t = w(t)$, $w_s = w(s)$.

Transform obtained v_i expanding the exponent in Taylor – Ito series near $t = 0$:

$$e^{-\frac{\alpha^2}{2}t + \alpha w_t} = 1 + \alpha w_t + \frac{1}{2!}\alpha^2\left(w_t^2 - t\right) + \frac{1}{3!}\alpha^3\left(w_t^3 - 3t w_t\right) + \ldots.$$

Taking it into account we get

$$v_i = V_i + \frac{\left(1 + \frac{\sigma_0}{\sqrt{Kn}} w_t + \frac{1}{2!}\frac{\sigma_0^2}{Kn}\left(w_t^2 - t\right) + \frac{1}{3!}\frac{\sigma_0^3}{Kn\sqrt{Kn}}\left(w_t^3 - 3tw_t\right) + \ldots\right)(v_{0i} - V_{0i})}{1 + \frac{a_0(v_{0i}-V_{0i})}{Kn}\int_0^t \left(1 + \frac{\sigma_0}{\sqrt{Kn}} w_s + \frac{1}{2}\frac{\sigma_0^2}{Kn}\left(w_s^2 - s\right) + \ldots\right) ds},$$

$$v_i = V_i + \frac{\left(Kn + \sigma_0\sqrt{Kn} w_t + \frac{1}{2}\sigma_0^2\left(w_t^2 - t\right) + \frac{\sigma_0^3}{6}\frac{1}{\sqrt{Kn}}\left(w_t^3 - 3tw_t\right) + \ldots\right)(v_{0i} - V_{0i})}{Kn + a_0\left(v_{0i} - V_{0i}\right)\left(t + \int_0^t \frac{\sigma_0}{\sqrt{Kn}} w_s ds + \ldots\right)}.$$

Rejecting the terms of order Kn:

$$v_i = V_i + \frac{\left(\sigma_0\sqrt{Kn} w_t + \frac{1}{2}\sigma_0^2\left(w_t^2 - t\right) + \frac{\sigma_0^3}{6}\frac{1}{\sqrt{Kn}}\left(w_t^3 - 3tw_t\right) + \ldots\right)}{a_0\left(t + \frac{\sigma_0}{\sqrt{Kn}}\eta\frac{t^{\frac{3}{2}}}{\sqrt{3}} + \ldots\right)},$$

$\eta \sim \mathcal{N}(0,1)$ is a standard normally distributed random value.

At $t \to 0$ we get finally:

$$\begin{aligned} v_i &= V_i + \frac{\sigma_0\sqrt{Kn}}{a_0}\frac{w_t}{t} + \frac{\sigma_0^2}{2a_0}\left(\frac{w_t^2}{t} - 1\right), \\ dx_i &= v_i dt. \end{aligned}$$

In the same manner, for the case $c_i < 0$:

$$\begin{aligned} v_i &= V_i + \frac{\sigma_0\sqrt{Kn}}{a_0}\frac{w_t}{t} - \frac{\sigma_0^2}{2a_0}\left(\frac{w_t^2}{t} - 1\right), \\ dx_i &= v_i dt. \end{aligned}$$

So:

$$\begin{aligned} v_i &= V_i + \frac{\sigma_0\sqrt{Kn}}{a_0}\frac{w_t}{t} + \mathtt{sign}(c)\frac{\sigma_0^2}{2a_0}\left(\frac{w_t^2}{t} - 1\right), \\ dx_i &= v_i dt. \end{aligned}$$

$$x_i = \int v_i dt = x_0 + V_i t + \frac{\sigma\sqrt{Kn}}{a_0}\int_0^t \frac{w_s}{s} ds + \mathtt{sign}(c)\frac{\sigma_0^2}{2a_0}\int_0^t \left(\frac{w_t^2}{t} - 1\right) ds.$$

The calculation of integral $\int_0^t w_s/s ds$ is shown in Appendix 2:

$$\int_0^t w_s/s ds = \sqrt{t}(\varepsilon + \tilde{\varepsilon}),$$

where $\varepsilon, \tilde{\varepsilon} \sim \mathcal{N}(0,1)$ are independent random values normally distributed with zero mean and unit dispersion.

$$x_i = x_{0i} + V_i t + \frac{\sigma_0\sqrt{Kn}}{a_0}(\varepsilon + \tilde{\varepsilon})\sqrt{t} + \mathtt{sign}(c)\frac{\sigma_0^2}{2a_0}\int_0^t \left(\frac{w_t^2}{t} - 1\right) ds.$$

Our goal is an obtaining macroscopic equations for non – random macro – parameters, that's why we'll not take into account in the last expression the fast changing sign, blinking term with the mean equal to zero. Then:

$$dx_i = V_i dt + \frac{\sigma_0\sqrt{Kn}}{a_0} dw_t + \frac{\sigma_0\sqrt{Kn}}{a_0} d\tilde{w}_t,$$

$$dw_t dw_t = dt, \quad d\tilde{w}_t d\tilde{w}_t = dt, \quad dw_t d\tilde{w}_t = 0.$$

So we're coming to the system:

$$\begin{aligned} dx(t) &= V dt + \sqrt{Kn}\tilde{\sigma}\left(dw + d\tilde{w}\right), \\ dv(t) &= -\frac{1}{Kn} a(v(t) - V)dt + \frac{1}{\sqrt{Kn}}\sigma dw, \end{aligned} \tag{4}$$

where:

$$\tilde{\sigma}_{ij} \equiv \sigma_{0ij}/a_{0i} = 0,43\ kT^{1/4}, \tag{5}$$

and V is vector with the coordinates V_i, $i = 1, 2, 3$.

In the second equation instead of $a(c)$ and $\sigma(c)$ from (3) we take their computed values, averaged in velocity space with respect to the local maxwellian:

$$a \approx 2,979\ T^{1/2},\ \sigma \approx 1,73\ T^{3/4}. \tag{6}$$

The choice of coefficients depending only on x and t is frequently used, for instance, in the context of model collision integral in Fokker – Planck form [12]. Note, that from (6) with our simplifications we get Einstein's fluctuation – dissipation relation (in dimensional form for explicity): $\sigma^2/a = 2RT$.

The presence of the increment of stochastic term in the first equation in the form of two independent processes, which has appeared at the calculation of $\int_0^t w_s/sds$, is not trivial, at our glance. That is a significant step forward comparing to our previous results.

Let us derive the equations of stochastic gas dynamics for that set of coefficients. It means that we need to construct the equations for measures in 3D space which are generated by the random processes $x(t)$ and $v(t)$ belonging to the phase space. A physical meaning of that measures is the evolution of mass, momentum and energy distributions.

3 CONTINUITY EQUATION WITH SELF DIFFUSION

An amount of gas in a domain D is, from one side, a whole mass of molecules and, from the other side, an integral with respect to a measure: $\sum_{l:x_l\in D} m_l = \int_D \mu_t(dx)$, or, if all the particles possess the equal masses $1/N$: $\frac{1}{N}\sum_{l=1}^N \chi(x_l(t) \in D) = \int_D \mu_t(dx)$, where χ is a characteristic function. N can be considered as a number of realizations of the random process $x(t)$ which is a solution of the system (4).

We define a stochastic empirical measure $\mu_t(dx)$ by an expression: for any function $\psi \in C_b^{(2)}(\mathbf{R}^3)$ (a space of continuously differentiable finite functions)

$$\int \psi(x)\mu_t(dx) = \frac{1}{N}\sum_{l=1}^N \psi(x_l(t)), \tag{7}$$

more precisely: $\forall\psi \in C_b^{(2)}(\mathbf{R}^3)$ and $\forall D \in \mathbf{R}^3$

$$\int_D \psi(x)\mu_t(dx) = \frac{1}{N}\sum_{l=1}^{N}\psi(x_l(t))\chi(x_l(t) \in D).$$

That expression, connecting the measure distribution to realizations of particle positions at time moment t, is a Chebyshev quadrature formula (the weights are known and the nodes are parameters) if to read if from left to right.

For obtaining an equation for measure $\mu_t(dx)$, let us take a stochastic differentials from both of two sides of (7). We'll use Ito's formula for complex function differentiation

$$d\psi(x) = \sum_{i=1}^{3}\frac{\partial\psi}{\partial x_i}dx_i + \frac{1}{2}\sum_{i,j=1}^{3}\frac{\partial^2\psi}{\partial x_i\partial x_j}dx_i dx_j,$$

where stochastic differentials dx_i are taken from the system (4):

$$dx_i = V_i dt + \sqrt{Kn}\sum_{j=1}^{3}\widetilde{\sigma}_{ij}\left(dw_j + d\tilde{w}_j\right),$$

and because of definition of a standard three – dimensional Wiener process increment, the smallness of which is $\sqrt{dt}$ [21]:

$$dw_i dw_j = \delta_{ij}dt, \qquad dw_i dt = 0, \qquad dt^2 = 0, \tag{8}$$

that leads to

$$dx_i dx_j = 2Kn\sum_{m,n=1}^{3}\left(\widetilde{\sigma}_{im}\widetilde{\sigma}_{jn}\right)\delta_{mn}dt = 2Kn\sum_{m=1}^{3}\left(\widetilde{\sigma}_{im}\widetilde{\sigma}_{jm}\right)dt \equiv 2Kn\widetilde{\sigma}_{ij}^2 dt,$$

which means that Ito's formula in our case turns out to be:

$$\begin{aligned} d\psi(x) = &\left(\sum_{i=1}^{3}V_i\frac{\partial\psi}{\partial x_i} + Kn\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}^2\frac{\partial^2\psi}{\partial x_i\partial x_j}\right)dt \\ &+\sqrt{Kn}\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}\frac{\partial\psi}{\partial x_i}\left(dw_j + d\tilde{w}_j\right). \end{aligned} \tag{9}$$

Then we get the stochastic differentials from both of two sides of (7):

$$\begin{aligned} d\int\psi(x)\mu_t(dx) = &\frac{1}{N}\sum_{l=1}^{N}[(\sum_{i=1}^{3}V_i\frac{\partial\psi}{\partial x_i} + Kn\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}^2\frac{\partial^2\psi}{\partial x_i\partial x_j})(x_l(t))]dt \\ &+\frac{1}{N}\sum_{l=1}^{N}\left\{\sqrt{Kn}\left[\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}\frac{\partial\psi}{\partial x_i}\right](x_l(t))\left(dw_j + d\tilde{w}_j\right)\right\}, \end{aligned}$$

or, applying the formula (7) from right to left for the right hand side of the last expression:

$$\begin{aligned} d\int\psi(x)\mu_t(dx) = &\int[(\sum_{i=1}^{3}V_i(x,t)\frac{\partial\psi}{\partial x_i}(x) \\ &+Kn(x,t)\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}^2(x,t)\frac{\partial^2\psi}{\partial x_i\partial x_j}(x))dt]\mu_t(dx) \\ +\int\left\{\sqrt{Kn(x,t)}\left[\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}(x,t)\frac{\partial\psi}{\partial x_i}(x)\right]\left(dw_j + d\tilde{w}_j\right)\right\}&\mu_t(dx). \end{aligned}$$

Assuming existence of a density $\rho(x,t)$ of stochastic empirical measure $\mu_t(dx)$ (taking the usual steps while deriving from a generalized equation an equation in partial derivatives: having integrated by parts one or two times in appropriate places) we get a stochastic continuity equation in the form:

$$d\rho = \left[-\sum_{i=1}^{3} \frac{\partial}{\partial x_i} (V_i \rho) + \sum_{i,j=1}^{3} \frac{\partial^2}{\partial x_i \partial x_j} \left(Kn\, \widetilde{\sigma}_{ij}^2 \rho \right) \right] dt$$
$$- \sum_{i,j=1}^{3} \frac{\partial}{\partial x_i} \left(\sqrt{Kn}\, \widetilde{\sigma}_{ij} \rho \right) (dw_j + d\tilde{w}_j),$$

and having averaged over the time (taking mathematical expectation and having in mind that the mathematical expectations of the terms with $dw_j, d\tilde{w}_j$ are equal to zero ([16], theorem 3.2.1)) we get a deterministic continuity equation for time averaged deterministic mass density $\overline{\rho}(x,t)$:

$$\frac{\partial \overline{\rho}}{\partial t} + \sum_{i=1}^{3} \frac{\partial}{\partial x_i} \left(\overline{V_i \rho} \right) = \sum_{i,j=1}^{3} \frac{\partial^2}{\partial x_i \partial x_j} \left(Kn\, \overline{\widetilde{\sigma}_{ij}^2 \rho} \right),$$

which is valid for small Knudsen numbers. The right hand side reflects the trace left by the thermal motion of molecules, or self – diffusion. It does not destroy conservation because it describes diffusion and has a divergent form. So the doubts on the absence of the conservation law in our macro – model, marked in [22], seems to be not correct.

It is natural to regard the random values ρ, V_i and $\widetilde{\sigma}_{ij}^2$ (which depends on thermal velocity c) independent, that gives the product of averaged values after the averaging procedure. If we assume that the time averaging leads to the values using by traditional gas dynamics, then we get a continuity equation taking into account the self – diffusion (we omit the lines denoting time averaging above the macroparameters):

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{3} \frac{\partial}{\partial x_i} (V_i \rho) = \sum_{i,j=1}^{3} \frac{\partial^2}{\partial x_i \partial x_j} \left(Kn\, \overline{\widetilde{\sigma}_{ij}^2} \rho \right).$$

If to perform our derivations for a gas of maxwellian molecules, the self – diffusion corrector in the right hand side will vanish.

4 MOMENTUM EVOLUTION

Let us get equations for momentum and its density. We connect an amount of movement of a gas in D with a vector measure $\nu_t(dx)$ by $\sum_{l:x_l \in D} v_l m_l = \int_D \nu_t(dx)$, or, in a generalized form:

$$\forall \psi \in C_b^{(2)}(\mathbf{R}^3): \quad \int \psi(x) \nu_{t,i}(dx) = \frac{1}{N} \sum_{l=1}^{N} v_i(x_l(t)) \psi(x_l(t)) \qquad (i = 1,2,3), \tag{10}$$

considering the process $v(t)$ (the solution of (6)) as a function of $x(t)$.

Let us take stochastic differentials from both of two sides of that equality. We'll need a stochastic formula for product differentiation [21]:

$$d(v_i \psi) = \psi dv_i + v_i d\psi + dv_i d\psi.$$

The stochastic differentials dv_i are the equations of the system (4):

$$dv_i = -\frac{a_i}{Kn}(v_i - V_i)dt + \frac{1}{\sqrt{Kn}} \sum_{j=1}^{3} \sigma_{ij} dw_j$$

Applying Ito's formula (9) and the rules (8), calculate:

$$dv_i d\psi = \left(-\frac{a_i}{Kn}(v_i - V_i)dt + \frac{1}{\sqrt{Kn}}\sum_{j=1}^{3}\sigma_{ij}dw_j\right)$$
$$\times[\left(\sum_{i=1}^{3}V_i\frac{\partial\psi}{\partial x_i} + Kn\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}^2\frac{\partial^2\psi}{\partial x_i\partial x_j}\right)dt + \sqrt{Kn}\sum_{i,j=1}^{3}\widetilde{\sigma}_{ij}\frac{\partial\psi}{\partial x_i}\left(dw_j + d\tilde{w}_j\right)]$$
$$= \sum_{j=1}^{3}\sigma_{ij}dw_j\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}dw_n = \sum_{j=1}^{3}\sum_{m,n=1}^{3}\sigma_{ij}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\delta_{jn}dt$$
$$= \sum_{j=1}^{3}\sum_{m=1}^{3}\sigma_{ij}\widetilde{\sigma}_{mj}\frac{\partial\psi}{\partial x_m}dt \equiv \sum_{m=1}^{3}\sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}dt. \tag{11}$$

Rewrite as well:

$$\psi dv_i = \left(-\frac{a_i v_i}{Kn} + \frac{a_i V_i}{Kn}\right)\psi dt + \frac{1}{\sqrt{Kn}}\sum_{n=1}^{3}\sigma_{in}\psi dw_n,$$
$$v_i d\psi = v_i\left(\sum_{m=1}^{3}V_m\frac{\partial\psi}{\partial x_m} + Kn\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}^2\frac{\partial^2\psi}{\partial x_m\partial x_n}\right)dt$$
$$+v_i\sqrt{Kn}\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\left(dw_n + d\tilde{w}_n\right).$$

Then:

$$d(v_i\psi) = \left(-\frac{a_i v_i}{Kn} + \frac{a_i V_i}{Kn}\right)\psi dt + [v_i\left(\sum_{m=1}^{3}V_m\frac{\partial\psi}{\partial x_m} + Kn\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}^2\frac{\partial^2\psi}{\partial x_m\partial x_n}\right)$$
$$+\sum_{m=1}^{3}\sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}]dt + \frac{1}{\sqrt{Kn}}\sum_{n=1}^{3}\sigma_{in}\psi dw_n + v_i\sqrt{Kn}\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\left(dw_n + d\tilde{w}_n\right).$$

Therefore, the stochastic differentials from both of two sides of that equality (10) is:

$$d\int\psi(x)\nu_{t,i}(dx) = \frac{1}{N}\sum_{l=1}^{N}\left\{\left[-\frac{a_i v_i}{Kn} + \frac{a_i V_i}{Kn}\right](x_l(t))\right\}dt$$
$$+\frac{1}{N}\sum_{l=1}^{N}\left\{\left[v_i\sum_{m=1}^{3}V_m\frac{\partial\psi}{\partial x_m} + \sum_{m=1}^{3}\sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}\right](x_l(t))\right\}dt$$
$$+\frac{1}{N}\sum_{l=1}^{N}\left\{\left[v_i Kn\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}^2\frac{\partial^2\psi}{\partial x_m\partial x_n}\right](x_l(t))\right\}dt$$
$$+\frac{1}{N}\sum_{l=1}^{N}[(\frac{1}{\sqrt{Kn}}\sum_{n=1}^{3}\sigma_{in}\psi dw_n + v_i\sqrt{Kn}\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\left(dw_n + d\tilde{w}_n\right))(x_l(t))],$$

Remembering the definitions of the measures $\mu_t(dx)$ and $\nu_t(dx)$, we get the equation for the measure $\nu_{t,i}(dx)$:

$$d\int\psi(x)\nu_{t,i}(dx) = \left[-\int\frac{a_i}{Kn}\psi\nu_{t,i}(dx) + \int\frac{a_i}{Kn}\psi V_i(x,t)\mu_t(dx)\right]dt$$

$$+\left[\sum_{m=1}^{3}\int V_m\frac{\partial\psi}{\partial x_m}\nu_{t,i}(dx)+\sum_{m=1}^{3}\int\sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}\mu_t(dx)\right]dt$$

$$+\sum_{m,n=1}^{3}\int Kn\;\widetilde{\sigma}^2_{mn}\frac{\partial^2\psi}{\partial x_m\partial x_n}\nu_{t,i}(dx)dt$$

$$+\sum_{n=1}^{3}\int\frac{1}{\sqrt{Kn}}\sigma_{in}\psi\mu_t(dx)dw_n+\sum_{m,n=1}^{3}\int\sqrt{Kn}\;\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\nu_{t,i}(dx)\,(dw_n+d\tilde{w}_n)\,.$$

Denoting as $\rho V_i(x,t)$ a density of the measure $\nu_{t,i}(dx)$ and integrating by parts one or two times in appropriate places, we get a stochastic differential equation, which is a stochastic analogue of Navier – Stokes equation:

$$d(\rho V_i)=-\sum_{m=1}^{3}\frac{\partial}{\partial x_m}\left(V_m\rho V_i\right)dt-\sum_{m=1}^{3}\frac{\partial}{\partial x_m}\left(\sigma_{im}\widetilde{\sigma}_{im}\rho\right)dt$$

$$+\sum_{m,n=1}^{3}\frac{\partial^2}{\partial x_m\partial x_n}\left(Kn\;\widetilde{\sigma}^2_{mn}\rho V_i\right)dt$$

$$+\sum_{n=1}^{3}\frac{1}{\sqrt{Kn}}\sigma_{in}\rho dw_n+\sum_{m,n=1}^{3}\frac{\partial}{\partial x_m}\left(\sqrt{Kn}\;\widetilde{\sigma}_{mn}\rho V_i\right)(dw_n+d\tilde{w}_n)\,,$$

if Knudsen numbers are small.

Taking a mathematical expectation, using the same arguments as at derivation of deterministic continuity equation with self – diffusion, we get deterministic quasi gas dynamics equations for the momentum density (omitting the lines above gas dynamics values):

$$\frac{\partial}{\partial t}(\rho V_i)+\sum_{m=1}^{3}\frac{\partial}{\partial x_m}\left(V_m\rho V_i\right)=-\sum_{m=1}^{3}\frac{\partial}{\partial x_m}\left(\overline{\sigma_{im}\widetilde{\sigma}_{im}}\rho\right)$$

$$+\sum_{m,n=1}^{3}\frac{\partial^2}{\partial x_m\partial x_n}\left(Kn\;\overline{\widetilde{\sigma}^2_{mn}}\rho V_i\right),\qquad(i=1,2,3).$$

5 ENERGY DISTRIBUTION

Let us get equations for energy density. Define a measure $\epsilon_t(dx)$:

$$\forall\psi\in C_b^{(2)}(\mathbf{R}^3):\quad\int\psi(x)\epsilon_t(dx)=\frac{1}{N}\sum_{l=1}^{N}\psi(x_l(t))\sum_{i=1}^{3}\frac{v_i^2}{2}(x_l(t)).\tag{12}$$

Take the stochastic formula for product differentiation

$$d(\frac{v_i^2}{2}\psi)=\psi d(\frac{v_i^2}{2})+\frac{v_i^2}{2}d\psi+d(\frac{v_i^2}{2})d\psi,$$

$$d(\frac{v_i^2}{2})=v_i dv_i+\frac{1}{2}(dv_i)^2,$$

the system (4), Ito's formula (9), expressions (8):

$$\frac{1}{2}(dv_i)^2=\frac{1}{2Kn}\sum_{n=1}^{3}\sigma_{in}dw_n\sum_{m=1}^{3}\sigma_{im}dw_m=\frac{1}{2}\frac{\sigma_{ii}^2}{Kn}dt,$$

the formula (11):

$$dv_i d\psi = \sum_{m=1}^{3} \sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}dt,$$

$$d(\frac{v_i^2}{2}\psi) = \psi v_i dv_i + \frac{v_i^2}{2}d\psi + v_i dv_i d\psi + \psi\frac{1}{2}\frac{\sigma_{ii}^2}{Kn}dt$$
$$= \psi\left[\left(-\frac{a_i v_i^2}{Kn} + \frac{a_i v_i V_i}{Kn} + \frac{1}{2}\frac{\sigma_{ii}^2}{Kn}\right)dt + \frac{1}{\sqrt{Kn}}\sum_{m=1}^{3}\sigma_{im}dw_m\right]$$
$$+\frac{v_i^2}{2}\left(\sum_{m=1}^{3}V_m\frac{\partial\psi}{\partial x_m} + Kn\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}^2\frac{\partial^2\psi}{\partial x_m\partial x_n}\right)dt$$
$$+\frac{v_i^2}{2}\sqrt{Kn}\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\left(dw_n + d\tilde{w}_n\right) + v_i\sum_{m=1}^{3}\sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}dt.$$

The stochastic differentiation of the formula (12) with account to the just obtained expressions leads to an equation for evolution of the measure $\epsilon_t(dx)$ (regarding to assumed isotropy in velocity space we can put $a_i = a$):

$$d\int\psi(x)\epsilon_t(dx) = [-2\int\psi\frac{a}{Kn}\epsilon_t(dx) + \int\psi\frac{1}{2}\sum_{i=1}^{3}\frac{\sigma_{ii}^2}{Kn}\mu_t(dx) + \int\psi\sum_{i=1}^{3}\frac{a}{Kn}V_i(x,t)\nu_{t,i}(dx)]dt$$
$$+\int\left(\sum_{m=1}^{3}V_m\frac{\partial\psi}{\partial x_m} + Kn\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}^2\frac{\partial^2\psi}{\partial x_m\partial x_n}\right)\epsilon_t(dx)dt + \sum_{i,m=1}^{3}\sigma_{im}\widetilde{\sigma}_{im}\frac{\partial\psi}{\partial x_m}\nu_{t,i}(dx)dt$$
$$+\int\psi\frac{1}{\sqrt{Kn}}\sum_{i,m=1}^{3}\sigma_{im}\mu_t(dx)dw_m + \int\sqrt{Kn}\sum_{m,n=1}^{3}\widetilde{\sigma}_{mn}\frac{\partial\psi}{\partial x_m}\epsilon_t(dx)\left(dw_n + d\tilde{w}_n\right).$$

Assuming existence of densities $\rho V_i(x,t)$ and $\rho E(x,t)$ of measures $\nu_{t,i}(dx)$ and $\epsilon_t(dx)$ we get for the term inside the first square brackets:

$$[-\int\psi\frac{2a}{Kn}\left(\rho E - \sum_{i=1}^{3}(\frac{\rho V_i^2}{2} + \frac{\sigma_{ii}^2}{4a})\right)dx].$$

For ideal gas, because of the fluctuation – dissipation theorem [12, 23], we can write: $\sigma^2/2a = RT$ (in dimensional form for clearness). That, together with the definition of temperature and total energy density

$$\rho E = \frac{\rho V^2}{2} + \frac{3}{2}RT\rho$$

brings zero to this expression after averaging.

Then a stochastic differential equation for energy density looks like

$$d(\rho E) = \left[-\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j\rho E) - \sum_{i,j=1}^{3}\frac{\partial}{\partial x_j}\left(\sigma_{ij}\widetilde{\sigma}_{ij}\rho V_i\right)\right]dt + \left[\sum_{i,j=1}^{3}\frac{\partial^2}{\partial x_i\partial x_j}\left(Kn\,\widetilde{\sigma}_{ij}^2\rho E\right)\right]dt$$
$$+\frac{1}{\sqrt{Kn}}\sum_{i,m=1}^{3}\sigma_{im}\rho dw_m + \sqrt{Kn}\sum_{m,n=1}^{3}\frac{\partial}{\partial x_m}\left(\widetilde{\sigma}_{mn}\rho E\right)\left(dw_n + d\tilde{w}_n\right),$$

and its deterministic part

$$\frac{\partial(\rho E)}{\partial t} + \sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j\rho E) = -\sum_{i,j=1}^{3}\frac{\partial}{\partial x_j}\left(\overline{\sigma_{ij}\widetilde{\sigma}_{ij}}\rho V_i\right) + \sum_{i,j=1}^{3}\frac{\partial^2}{\partial x_i\partial x_j}\left(Kn\,\overline{\widetilde{\sigma}_{ij}^2\rho E}\right),$$

6 STOCHASTIC GAS DYNAMICS SYSTEM

Let us write the obtained system in Cartesian coordinates:

$$\frac{\partial}{\partial t}\rho+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(\rho V_j)=\sum_{i,j=1}^{3}\frac{\partial^2}{\partial x_i\partial x_j}\left(Kn\,\overline{\widetilde{\sigma}_{ij}^2}\rho\right),$$
$$\frac{\partial}{\partial t}(\rho V_i)+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j\rho V_i)=-\sum_{j=1}^{3}\frac{\partial}{\partial x_j}\left(\overline{\sigma_{ij}\widetilde{\sigma}_{ij}}\rho\right)+\sum_{k,j=1}^{3}\frac{\partial^2}{\partial x_k\partial x_j}\left(Kn\,\overline{\widetilde{\sigma}_{kj}^2}\rho V_i\right),$$
$$\frac{\partial}{\partial t}(\rho E)+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j\rho E)=-\sum_{i,j=1}^{3}\frac{\partial}{\partial x_j}\left(\overline{\sigma_{ij}\widetilde{\sigma}_{ij}}\rho V_j\right)+\sum_{i,j=1}^{3}\frac{\partial^2}{\partial x_i\partial x_j}\left(Kn\,\overline{\widetilde{\sigma}_{ij}^2\rho E}\right).$$

Denote and compute taking into account (5), (6): $A\equiv\overline{\widetilde{\sigma}_{ij}^2}=0.085\,T^{1/2},\ B\equiv\overline{\sigma_{ij}\widetilde{\sigma}_{ij}}=0.5\,T.$ The latter means our choice of the parameter $k=0.675$ for denoting a combination (we'll find it in the second equation of the system below) $B\rho$ as p calling it "pressure". Then the equality $p=B\rho$ or $p=\rho RT$ (in dimensional form) can be called an equation of state. So:

$$\frac{\partial\rho}{\partial t}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(\rho V_j)=\sum_{j=1}^{3}\frac{\partial^2}{\partial x_j^2}(KnA\rho)$$
$$\frac{\partial(\rho V_i)}{\partial t}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j(\rho V_i))=-\frac{\partial}{\partial x_i}(B\rho)+\sum_{j=1}^{3}\frac{\partial^2}{\partial x_j^2}(KnA(\rho V_i)),$$
$$\frac{\partial(\rho E)}{\partial t}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j(\rho E))=-\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(B(\rho V_j))+\sum_{j=1}^{3}\frac{\partial^2}{\partial x_j^2}(KnA(\rho E)).$$

Introducing the notions $p\equiv B\rho$, $\nu\equiv KnA$ and calling p by pressure and ν by coefficient of kinematic viscosity, we'll have the macroscopic system in traditional form:

$$\frac{\partial\rho}{\partial t}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(\rho V_j)=\sum_{j=1}^{3}\frac{\partial^2}{\partial x_j^2}(\nu\rho)$$
$$\frac{\partial(\rho V_i)}{\partial t}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j(\rho V_i))=-\frac{\partial p}{\partial x_i}+\sum_{j=1}^{3}\frac{\partial^2}{\partial x_j^2}(\nu(\rho V_i)),\ i=1,2,3$$
$$\frac{\partial(\rho E)}{\partial t}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(V_j(\rho E))=-\sum_{j=1}^{3}\frac{\partial(pV_j)}{\partial x_j}+\sum_{j=1}^{3}\frac{\partial^2}{\partial x_j^2}(\nu(\rho E)).$$

7 A NUMERICAL EXAMPLE USING DISCONTINUOUS PARTICLE METHOD

We take a well known problem of shock wave structure as a test. Strictly speaking, it is not a test for macroscopic models because of quite high value of Knudsen number.

The original model (1) is of interest not only because of the possibility to construct new stochastic and deterministic macro – models but also is a base for direct modeling by the help of stochastic particle method [8], attractable for high performance simulation. In the present paper we use a deterministic particle method, alternative to stochastic one, free from parasitic fluctuations.

The applied explicit particle method [24] has a minimal dissipation that makes it possible to get a solution with high accuracy for correct comparison of different models. Here we'll describe our method shortly, paying attention at algorithmic peculiarities.

Divide a 1D computational domain in equal intervals. The heights of particles of three types (representing mass, momentum and energy densities) are taken equal to values of initial $\rho(x,0)$, $\rho v(x,0)$, $\rho E(x,0)$ in points $x_i(0)$. The heights of particles with number i at j time step are denoted by ρ_i^j, ρv_i^j, ρE_i^j; p_i^j and T_i^j are the values of pressure and temperature in a particle with number i at j time step.

1. Predictor. A system for positions of mass, momentum and energy particles is solved

$$\frac{dx_i}{dt} = v_i;\ x_i(0) = x_{i0}.$$

by explicit Euler scheme (its accuracy is quite enough because the main source of error is transport nonlinearity):

$$x_i^{j+1} = x_i^j + \tau * v_i^j; \quad v_i^j = \rho v_i^j / \rho_i^j$$

2. Corrector (the main step to struggle with nonlinearity and the main peculiarity of our "discontinuous" method). As a result of particles movements "overlaps" and "gaps" occurs. It destroys the approximation of a function by discrete set of particles. To avoid it let us use the particles reconstruction from [24].

3. The pressure calculation. The difference of pressures leads to the particles volumes change (see more in [24]).

4. The account of viscosity. To do it we take a standard difference approximation for the second derivative at nonuniform grid:

$$\frac{\partial^2(\nu u)}{\partial x^2} \approx (\nu u)_{\bar{x}\hat{x}} \equiv \frac{\nu}{\hbar_i}\left(\frac{u_{i+1}-u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}\right),$$

where $\hbar_i = 0.5(h_i + h_{i+1})$.

We do not use limiters as in [24] because of sufficient natural viscosity. Our system turned out to be less demanding to the values of time steps than Navier – Stokes one.

We consider a one – atom gas ($\gamma = 5/3$, $Pr = 2/3$) at inlet flow Mach number $M = 1.55$. At Fig.1 the profiles of the normed density for the stochastic gas dynamics (SDE) and Navier – Stokes systems are given.

$$\rho' = \frac{\rho - \rho_1}{\rho_2 - \rho_1}.$$

where ρ_1 and ρ_2 are the values at infinity.

Note that our system gives a result close to the experiment [25].

8 CONCLUSIONS

Regardless to our quite severe simplifications we obtained a gas dynamics system which has clear microscopic origin and gives more adequate than usual results at the well known test. Moreover, we've got the hierarchy of micro – macro stochastic and deterministic models each of which has its own place in a row of unified solvers.

9 APPENDIX 1. EQUATION $dx_t = \beta x_t^\gamma dt + \alpha x_t dw_t$

Consider a stochastic differential equation:

$$dx_t = \beta x_t^\gamma dt + \alpha x_t dw_t, \quad \text{where } \alpha, \beta, \gamma - const. \tag{13}$$

Introduce an integrating multiplyer ([21], chapter 5)

$$F_t = \exp\left(-\alpha w_t + \frac{\alpha^2}{2}t\right).$$

Multiply by it both sides of the equation:

$$F_t dx_t = \beta x_t^\gamma F_t dt + \alpha x_t F_t dw_t, \quad F_t dx_t - \alpha x_t F_t dw_t = \beta x_t^\gamma F_t dt,$$

and show that

$$F_t dx_t - \alpha x_t F_t dw_t = d(F_t x_t).$$

By the formula of product stochastic differentiation

$$d(F_t x_t) = F_t dx_t + x_t dF_t + dx_t dF_t.$$

Apply Ito formula for a function $g(t,x) = \exp\left(-\alpha x + \frac{\alpha^2}{2}t\right)$

$$dF_t = \frac{\partial g}{\partial t}dt + \frac{\partial g}{\partial x}dw_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(dw_t)^2,$$

$$dtdt = 0, \quad dtdw_t = 0, \quad dw_t dt = 0, \quad (dw_t)^2 = dt,$$

$$\frac{\partial g}{\partial t} = F_t\frac{\alpha^2}{2}, \quad \frac{\partial g}{\partial x} = (-\alpha)F_t, \quad \frac{\partial^2 g}{\partial x^2} = \alpha^2 F_t.$$

Then:

$$dF_t = \frac{\alpha^2}{2}F_t dt + (-\alpha)F_t dw_t + \frac{\alpha^2}{2}F_t dt = \alpha^2 F_t dt + (-\alpha)F_t dw_t.$$

Replace dx_t using equation (13):

$$dx_t dF_t = -\alpha^2 F_t x_t dt,$$

$$d(F_t x_t) = F_t dx_t + \alpha^2 F_t x_t dt - \alpha F_t x_t dw_t - \alpha^2 F_t x_t dt = F_t dx_t - \alpha F_t x_t dw_t.$$

So, equation (13) is equivalent to equation:

$$d(F_t x_t) = \beta x_t^\gamma F_t dt. \tag{14}$$

Denote $F_t(\xi)x_t(\xi) = y_t(\xi)$ considering ξ as a parameter, then $x_t = F_t^{-1}y_t$, and equation (14) can be written as a deterministic differential equation with respect to a function $t \to y_t(\xi)$ for each ξ [21]:

$$\begin{aligned} &\frac{dy_t(\xi)}{dt} = F_t(\xi)h(t, F_t^{-1}(\xi)y_t(\xi)) \\ &y|_{t=0} = x_0(\xi), \end{aligned} \tag{15}$$

where $h(t,x_t)=\beta x_t^{\gamma}$. Solve equation (15) for $y_t(\xi)$.

$$x_t = F_t^{-1} y_t = \exp\left(\alpha w_t - \frac{\alpha^2}{2}t\right) y_t,$$

$$dy_t/dt = \exp\left(-\alpha w_t + \frac{\alpha^2}{2}t\right)\beta y_t^{\gamma} / \left(\exp\left(-\alpha w_t + \frac{\alpha^2}{2}t\right)\right)^{\gamma},$$

$$\frac{dy_t}{y_t^{\gamma}} = \beta\left(\exp\left(-\alpha w_t + \frac{\alpha^2}{2}t\right)\right)^{1-\gamma} dt,$$

$$\frac{y_t^{-\gamma+1}}{-\gamma+1} = \beta\int_0^t \exp\left(\left(-\alpha w_s + \frac{\alpha^2}{2}s\right)(1-\gamma)\right) ds,$$

$$y_t = \left(y_0^{1-\gamma} + (1-\gamma)\,\beta\int_0^t \exp\left(\left(-\alpha w_s + \frac{\alpha^2}{2}s\right)(1-\gamma)\right) ds\right)^{\frac{1}{1-\gamma}}.$$

Finally, for x_t we get:

$$x_t = \exp\left(\alpha w_t - \frac{\alpha^2}{2}t\right)\left(x_0^{1-\gamma} + (1-\gamma)\,\beta\int_0^t \exp\left(\left(-\alpha w_s + \frac{\alpha^2}{2}s\right)(1-\gamma)\right) ds\right)^{\frac{1}{1-\gamma}}.$$

We are interested in values $\gamma = 2$ and $\alpha = \frac{\sigma_0}{\sqrt{Kn}}$, $\beta = -\frac{a_0}{Kn}$:

$$x_t = \exp\left(\alpha w_t - \frac{\alpha^2}{2}t\right)\left(x_0^{-1} - \beta\int_0^t \exp\left(\alpha w_s - \frac{\alpha^2}{2}s\right) ds\right)^{-1}.$$

10 APPENDIX 2. CALCULATION OF INTEGRAL $\int_0^t w_s/s ds$.

Show that

$$\int_0^t f(s) w_s ds = \sigma(t)\eta,$$

where $f(t)$ is an arbitrary function, $\sigma^2(t) = \int_0^t \left(\int_s^t f(\tau)d\tau\right)^2 ds$ is a dispersion [26],

$$\sigma^2(t) = \langle \int_0^t f(\tau) w_\tau d\tau, \int_0^t f(\tau) w_\tau d\tau \rangle,$$

a random value η is not independent of Wiener wandering w_t, where $w_t = \varepsilon\sqrt{t}$, η can be represented as $\eta = \rho\varepsilon + \sqrt{1-\rho}\tilde{\varepsilon}$ where ε and $\tilde{\varepsilon}$ are independent random values, ρ is correlation cofficient:

$$\rho = <\varepsilon, \eta> = \left(1/\left(\sigma(t)\sqrt{t}\right)\right)\int_0^t (\int_s^t f(\tau)d\tau) ds$$

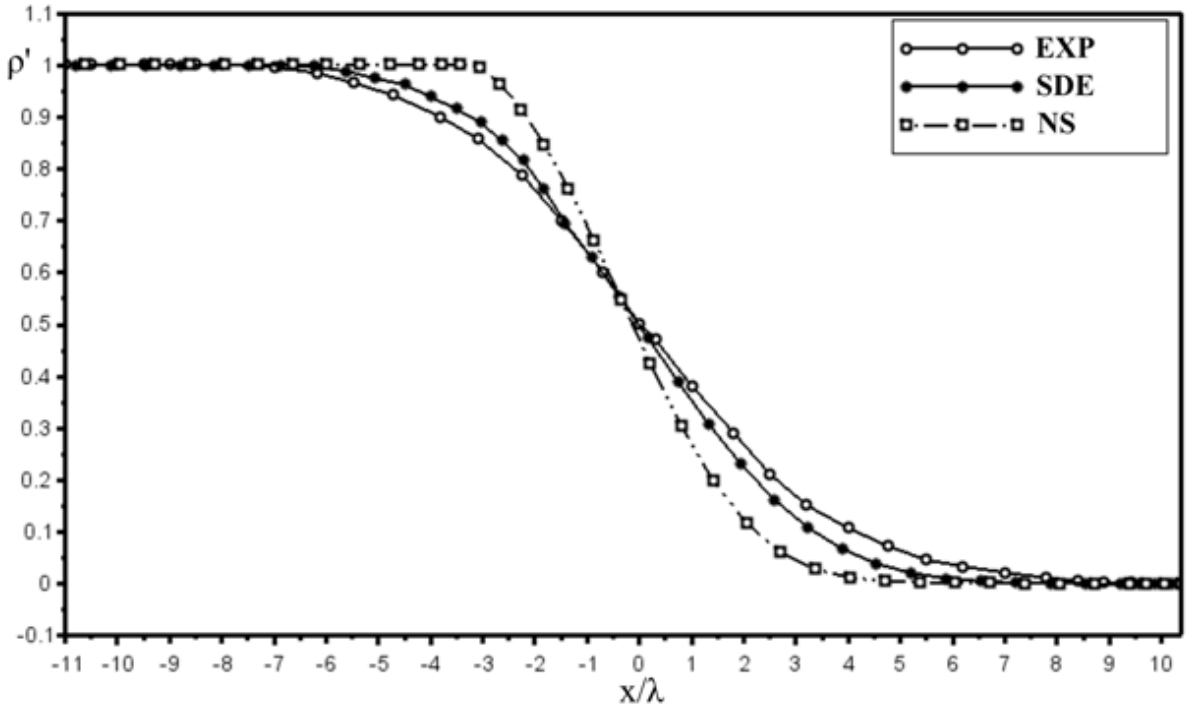


Figure 1: The profiles of the normed density in a shock wave for the stochastic gas dynamics (SDE) and Navier – Stokes (NS) systems, compared to EXPeriment.

Determine the integral by an integral sum S_n.

Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random values, having normal distribution with zero mean and unit dispersion: $\varepsilon_i \sim \mathcal{N}(0,1)$. Divide a segment $[0;t]$ in n segments of length $\Delta t = t/n$. The values of Wiener process at the end k-th segment is equal to a sum of k independent random Gaussian increments at each segment $k = 1, 2, \ldots, n$. Denote $f_i \equiv f((i-1)\Delta t)$, then

$$S_n = (f_1\varepsilon_1 + f_2(\varepsilon_1+\varepsilon_2) + \ldots + f_{n-1}(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_{n-1}))\sqrt{\Delta t}\Delta t,$$

and its dispersion:

$$\begin{aligned}\sigma_n^2 = \langle S_n, S_n\rangle = \langle((f_1\varepsilon_1 + f_2(\varepsilon_1+\varepsilon_2) + \ldots + f_{n-1}(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_{n-1}))\sqrt{\Delta t}\Delta t),\\ ((f_1\varepsilon_1 + f_2(\varepsilon_1+\varepsilon_2) + \ldots + f_{n-1}(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_{n-1}))\sqrt{\Delta t}\Delta t)\rangle\\ = \langle\Delta t^{\frac{3}{2}}((f_1+f_2+\ldots+f_{n-1})\varepsilon_1 + (f_2+f_3+\ldots+f_{n-1})\varepsilon_2 + \ldots + f_{n-1}\varepsilon_{n-1}),\\ \Delta t^{\frac{3}{2}}((f_1+f_2+\ldots+f_{n-1})\varepsilon_1 + (f_2+f_3+\ldots+f_{n-1})\varepsilon_2 + \ldots + f_{n-1}\varepsilon_{n-1})\rangle\\ = \Delta t((f_1+f_2+\ldots+f_{n-1})^2\Delta t^2 + (f_2+f_3+\ldots+f_{n-1})^2\Delta t^2 + \ldots + f_{n-1}^2\Delta t^2).\end{aligned}$$

At $\Delta t \to 0$

$$\lim_{\Delta t\to 0}(f_k + f_{k+1} + \ldots + f_{n-1})^2\Delta t^2 = (\int_{t_k}^{t} f(\tau)d\tau)^2,$$

$$\begin{aligned}\sigma^2(t) = \lim_{\Delta t\to 0}\Delta t\left(\left(\int_{t_0}^{t} f(\tau)d\tau\right)^2 + \left(\int_{t_1}^{t} f(\tau)d\tau\right)^2 + \ldots + \left(\int_{t_k}^{t} f(\tau)d\tau\right)^2\right)\\ = \int_0^t\left(\int_s^t f(\tau)d\tau\right)^2 ds,\end{aligned}$$

Let's calculate $\sigma^2(t)$ and $\rho(t)$ for $f(t) = 1/t$.

$$\sigma^2 = \int_0^t \left(\int_s^t f(\tau) d\tau \right)^2 ds = \int_0^t \left(\int_s^t 1/\tau d\tau \right)^2 ds = \int_0^t \left((\ln|\tau|) \,|_s^t \right)^2 ds$$
$$= \int_0^t (\ln t - \ln s)^2 ds = t \int_0^t \ln^2(s/t) d(s/t)$$
$$= t(s/t \ln^2(s/t) - 2s/t ln(s/t) + 2s/t)|_0^t = 2t,$$

An integral $I = \int_0^t (\int_s^t 1/\tau d\tau) ds$:

$$I = \int_0^t (\ln|\tau|)|_s^t ds = \int_0^t (\ln t - \ln s) ds = -t \int_0^t \ln(s/t) d(s/t)$$
$$= -t(s/t \ln(s/t) - s/t)|_0^t = t,$$

then $\rho = 1/\sqrt{2}$, $\eta = (\varepsilon + \tilde{\varepsilon})/\sqrt{2}$.

Finally, we have $\int_0^t w_s/s ds = \sqrt{t}(\varepsilon + \tilde{\varepsilon})$.

REFERENCES

[1] S.A. Zabelok, R.R. Arslanbekov, V. I. Kolobov, Adaptive kinetic-fluid solvers for heterogeneous computing architectures. *Journal of Computational Physics*, **303**, 455–469, 2015.

[2] Ya.B. Zel'dovich, S.A. Molchanov, A.A. Ruzmaikin, D.D. Sokolov, Intermittency in random media. *Sov. Phys. Uspekhi*, **30**, 353–369, 1987.

[3] M.H. Gorji, P. Jenny, Fokker-Planck-DSMC algorithm for simulations of rarefied gas flows. *Journal of Computational Physics*, **287**, 110–129, 2015.

[4] J. Zhang, D. Zeng, J. Fan, Analysis of transport properties determined by Langevin dynamics using Green-Kubo formulae. *Physica A: Statistical Mechanics and its Applications*, **411**, 104–112, 2014.

[5] V.K. Gupta, M. Torrilhon, Comparison of relaxation phenomena in binary gas-mixtures of Maxwell molecules and hard spheres. *Computers & Mathematics with Applications*, **70**, 73–88, 2015.

[6] S.V. Bogomolov, An approach to deriving stochastic gas dynamics models. *Doklady Mathematics*, **78**, 929–931, 2008.

[7] A.A. Arsen'yev, On the approximation of the solution of the Boltzmann equation by solutions of the ito stochastic differential equations. *USSR Computational Mathematics and Mathematical Physics*, **27**, 51–59, 1987.

[8] S.V. Bogomolov, L.V. Dorodnitsyn, Equations of stochastic quasi-gas dynamics: Viscous gas case. *Mathematical Models and Computer Simulations*, **3**, 457–467, 2011.

[9] S.V. Bogomolov, Stochastic quasi gas dynamics equations as a base for particle methods. *Proc. V European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, Lisbon, Portugal, 2010.

[10] S. Chandrasekhar, Stochastic problems in physics and astronomy. *Rev. Modern. Phys.*, **15**, 1–89, 1943.

[11] J.G. Kirkwood, The statistical mechanical theory of transport processes. *J. Chem. Phys.*, **14**, 180–201, 1946.

[12] C. Cercignani, *Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations*. Cambridge University Press, 2000.

[13] A. V. Skorokhod, *Stochastic Equations for Complex Systems*. Dordrecht : Kluwer Academic, 1987.

[14] S.V. Bogomolov, I.G. Gudich, Verification of a stochastic diffusion gas model. *Mathematical Models and Computer Simulations*, **6**, 305–316, 2014.

[15] B.N. Chetverushkin, Resolution limits of continuous media mode and their mathematical formulations. *Mathematical Models and Computer Simulations*, **5**, 266-279, 2013.

[16] T.G. Elizarova, *Quasi-Gas Dynamic Equations*. Springer, 2009.

[17] H. Brenner, Bi-velocity hydrodynamics. *Physica A: Statistical Mechanics and its Applications*, **388**, 3391–3398, 2009.

[18] J. Mathiaud, L. Mieussens, A Fokker-Planck model of the Boltzmann equation with correct Prandtl number. *Journal of Statistical Physics*, **162**, 397–414, 2016.

[19] K. Morinishi, A continuum/kinetic hybrid approach for multi-scale flow. *Proc. ECCOMAS CFD 2006*, Egmond aan Zee, Netherlands, 2006.

[20] A.K. Aringazin, M.I. Mazhintov, Stochastic Models of Lagrangian Acceleration of Fluid Particle in Developed Turbulence. *Int. J. Mod. Phys. B*, **18**, 3095–3168, 2004

[21] B. Oksendal, *Stochastic Differental Equations. 6th edition*. Berlin : Springer, 2000.

[22] S.K. Dadzie, J.M. Reese, On the thermodynamics of volume/mass diffusion in fluids, arXiv : math-ph/1202.3169, 2011.

[23] G. Röpke, *Statistische Mechanik für das Nichtgleichgewicht*. Berlin : Deutscher Verlag der Wissenschaften, 1987.

[24] S.V. Bogomolov, An Entropy Consistent Particle Method for Navier-Stokes Equations. *Proc. IV European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS 2004* , Jyvaskyla, Finland, 2004.

[25] H. Alsmeyer, Density profiles in argon and nitrogen shock waves measured by the absorption of an electronic beam. *J. Fluid Mech.*, **74**, 497–513, 1976.

[26] S.S. Stepanov, *Stochastic world*. Springer International Publishing, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ADVANCES IN THE BOUNDARY ELEMENT METHOD IN GEOMECHANICS

Gernot Beer[1,2]

[1]emeritus Professor
Institute for structural analysis
University of Technology, Graz, Austria
e-mail: gernot.beer@tugraz.at

[2] conjoint Professor
ARC Centre of Excellence for Geotechnical Science and Engineering
The University of Newcastle, Australia

Keywords: Boundary Element Method, geomechanics.

Abstract. *In this work a novel approach is presented for the Boundary Element analysis of problems in geomechanics. Firstly, Non-Uniform Rational B-Spines (NURBS) are used for the description of the geometry and for the approximation of the unknowns. This results in a significant decrease in the number of parameters used for an accurate description of the geometry as well as a decrease in the number of degrees of freedom required for good quality results. Secondly, NURBS are also used for the description of the geometry of geological inclusions, which can have properties different to the rock mass and can experience inelastic behavior.*

After a short introduction to the theory, some details of implementation are shown. On test examples, involving elastic homogeneous domains, it is first shown that the method delivers accurate results with fewer parameters and number of unknowns as compared with conventional analysis. Solutions are compared to either known solutions or with conventional BEM analyses. Geological inclusions are introduced next and results of test examples are compared with Finite Element analyses. Finally a practical example is shown.

1 INTRODUCTION

The Boundary Element Method (BEM) is ideally suited for the analysis of problems in geomechanics as it can easily consider infinite and semi-infinite domains because the radiation condition is implicitly fulfilled. In the case of elastic, homogeneous domains only boundary integrals appear, and the solution involves a discretization of the boundary, thereby reducing the analysis effort by an order of magnitude.

However, to analyze real problems in geomechanics the consideration of heterogeneous and inelastic ground conditions is essential. The BEM can be extended to analyze these problems, but additional volume integrals appear. The numerical solution of the integral equations requires the discretization of a volume, therefore partially destroying the attractiveness of the method. However, the volume integrals only cover the part of the domain that has different material properties or behaves in an inelastic way. Currently the most popular method is to use internal cells for the volume discretization. Cells are like Finite Elements but the main difference is that their only purpose is the evaluation of the volume integral. This means that no additional degrees of freedom are introduced. The requirement for an additional volume discretization seems to have severely restricted the application of the BEM in geomechanics, with the Finite Element or similar domain methods dominating. In this paper it will be shown how piecewise heterogeneous, inelastic domains can be analyzed without a volume discretization, thereby enhancing the applicability of the method for problems in geomechanics.

Isogeometric analysis [1] has gained significant popularity in the last decade because of the fact that geometry data can be taken directly from Computer Aided Design (CAD) programs, potentially eliminating the need for mesh generation. NURBS basis functions, that are used for the definition of the geometry, are able to describe certain geometries such as arcs exactly. Therefore, as will be shown, the number of parameters, required to accurately define geometry, can be reduced significantly. NURBS are also used to define geological inclusions, eliminating the need for a volume discretization.

2 THE BEM WITH VOLUME EFFECTS

To apply the BEM to heterogeneous and inelastic problems, so called body force effects have to be included. Using the theorem of Betti as explained in [2], the boundary integral equation with body forces acting in a sub-volume V_0 can be written in incremental form and in matrix notation as (see Figure 1):

$$\begin{aligned} \mathbf{c}\,\dot{\mathbf{u}}(\mathbf{y}) = &\int_S \mathbf{U}(\mathbf{y},\mathbf{x})\dot{\mathbf{t}}(\mathbf{x})\,\mathrm{d}S + \int_{S_0} \mathbf{U}(\mathbf{y},\bar{\mathbf{x}})\dot{\mathbf{t}}_0(\bar{\mathbf{x}})\,\mathrm{d}S_0 \\ &- \int_S \mathbf{T}(\mathbf{y},\mathbf{x})\dot{\mathbf{u}}(\mathbf{x})\,\mathrm{d}S + \int_{V_0} \mathbf{U}(\mathbf{y},\bar{\mathbf{x}})\dot{\mathbf{b}}_0(\mathbf{x})\,\mathrm{d}V_0 \end{aligned} \tag{1}$$

where $\mathbf{c}$ is a free term, $\mathbf{U}(\mathbf{y},\mathbf{x})$ and $\mathbf{T}(\mathbf{y},\mathbf{x})$ are matrices containing fundamental solutions (Kernels) for the displacements and tractions at a point $\mathbf{x}$ due to a unit force at a point $\mathbf{y}$ [3], $\dot{\mathbf{u}}(\mathbf{x})$ and $\dot{\mathbf{t}}(\mathbf{x})$ are increments of the displacement and traction vectors on the surface S, defining the problem domain. $\dot{\mathbf{b}}_0(\bar{\mathbf{x}})$ are increments of body force inside the inclusion and $\dot{\mathbf{t}}_0(\bar{\mathbf{x}})$ are increments of tractions related to the body force acting on surface S_0 bounding V_0.

The integral equations can be solved for the unknowns $\mathbf{u}$ or $\mathbf{t}$ by discretization. As in the majority of previous work on the isogeometric BEM [4, 5, 6, 7, 8, 9, 10] we use the collocation method, i.e. we write the integral equations for a finite number, N, of source or collocation

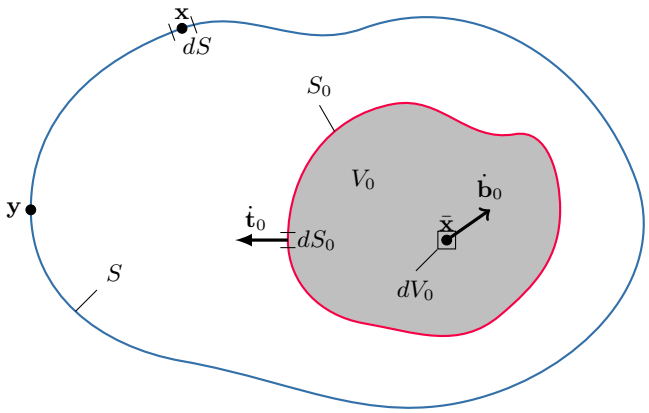


Figure 1: Explanation of the derivation of the integral equation with volume effects

points $\mathbf{y}_n$:

$$\mathbf{c}\,\dot{\mathbf{u}}\,(\mathbf{y}_n) = \int_S \mathbf{U}\,(\mathbf{y}_n,\mathbf{x})\,\dot{\mathbf{t}}\,(\mathbf{x})\,\mathrm{d}S + \int_{S_0} \mathbf{U}\,(\mathbf{y}_n,\bar{\mathbf{x}})\,\dot{\mathbf{t}}_0\,(\bar{\mathbf{x}})\,\mathrm{d}S_0 \\ - \int_S \mathbf{T}\,(\mathbf{y}_n,\mathbf{x})\,\dot{\mathbf{u}}\,(\mathbf{x})\,\mathrm{d}S + \int_{V_0} \mathbf{U}\,(\mathbf{y}_n,\bar{\mathbf{x}})\,\dot{\mathbf{b}}_0\,(\bar{\mathbf{x}})\,\mathrm{d}V_0 \tag{2}$$

with $n = \{1,\ldots,N\}$.

For the numerical evaluation of the surface integrals over S we divide the boundary into patches and use a geometry independent field approximation approach for each patch, i.e. we use different basis functions for the description of the geometry and for the field values:

$$\mathbf{x}^e = \sum_{k=1}^{K} N_k(\mathrm{u}) \cdot \mathbf{x}_k^e \tag{3}$$

$$\mathbf{u}^e = \sum_{k=1}^{K^d} N_k^d(\mathrm{u}) \cdot \mathbf{u}_k^e \tag{4}$$

$$\mathbf{t}^e = \sum_{k=1}^{K^t} N_k^t(\mathrm{u}) \cdot \mathbf{t}_k^e \tag{5}$$

In the above the superscript e refers to the number of the patch, N_k,N_k^d, N_k^t are NURBS basis functions of the local coordinate u for describing the geometry $\mathbf{x}^e$, displacements $\mathbf{u}^e$ and tractions $\mathbf{t}^e$, $\mathbf{x}_k^e$ specify the location of control points, $\mathbf{u}_k^e$,$\mathbf{t}_k^e$ are parameter values and K, K^d, K^t specify the number of parameters for each patch.

For an excavation problem for example the following system of equations can be assembled:

$$[\mathbf{T}]\,\{\mathbf{u}\} = \{\mathbf{F}\} + \{\mathbf{F}\}_0 \tag{6}$$

where $[\mathbf{T}]$ is an assembled matrix with coefficients related to Kernel $\mathbf{T}$ and $\{\mathbf{u}\}$ is a vector that collects all displacement components on points $\mathbf{y}_n$. $\{\mathbf{F}\}$ is a vector related to the the applied excavation tractions and $\{\mathbf{F}\}_0 = \{\mathbf{F}\}_0^{S_0} + \{\mathbf{F}\}_0^{V_0}$ is the right hand side related to the body force effects, i.e. related to the integrals over S_0 and V_0 in Equation (2). Details of the implementation of the isogeometric BEM for elastic homogeneous domains can be found in [3, 11].

3 NURBS basis functions

An detailed treatise on NURBS basis functions is presented in [3], here only a short explanation is given. NURBS or Non-uniform rational B-splines are an extension of classical B-splines. To define B-splines we start with a *knot vector*. This is a vector containing a series of non-decreasing values of the local coordinate:

$$\Xi = \begin{pmatrix} \mathrm{u}_0 & \mathrm{u}_1 & \cdots & \mathrm{u}_N \end{pmatrix} \tag{7}$$

We define the entries in the vector as *knots*. With the knot vector a recursive formula is applied. First we compute the functions for order $p = 0$ (constant) and for $i = 0, \cdots, N$.

$$N_{i,0}(\mathrm{u}) = \begin{cases} 1 & \text{if} \quad \mathrm{u}_i \leqslant \mathrm{u} < \mathrm{u}_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Higher order basis functions are defined by referencing lower order functions:

$$N_{i,p}(\mathrm{u}) = \frac{\mathrm{u} - \mathrm{u}_i}{\mathrm{u}_{i+p} - \mathrm{u}_i} N_{i,p-1}(\mathrm{u}) + \frac{\mathrm{u}_{i+p+1} - \mathrm{u}}{\mathrm{u}_{i+p+1} - \mathrm{u}_{i+1}} N_{i+1,p-1}(\mathrm{u}) \tag{9}$$

NURBS basis functions are obtained by including weights, w_i :

$$R_{i,p}(\mathrm{u}) = \frac{N_{i,p}(\mathrm{u})\,\mathrm{w}_i}{\sum_{j=0}^{I} N_{j,p}(\mathrm{u})\,\mathrm{w}_j} \tag{10}$$

4 Geometry description with NURBS

NURBS are ideally suited for the description of geometry (for example they are able to describe circular arcs exactly) and this is one of the main reasons they are used by the CAD community. The main difference to commonly used Lagrange polynomials, is that the concept of nodal points is replaced by a concept of control points, which do not always lie on the curve.

As an example we show the description of the geometry of an NATM tunnel, where the design shape is specified by arcs (center, radius and extent) as shown in the tables in Figure 2. One half of the tunnel can be described with 1 NURBS patch of order 2 and only 7 control points.

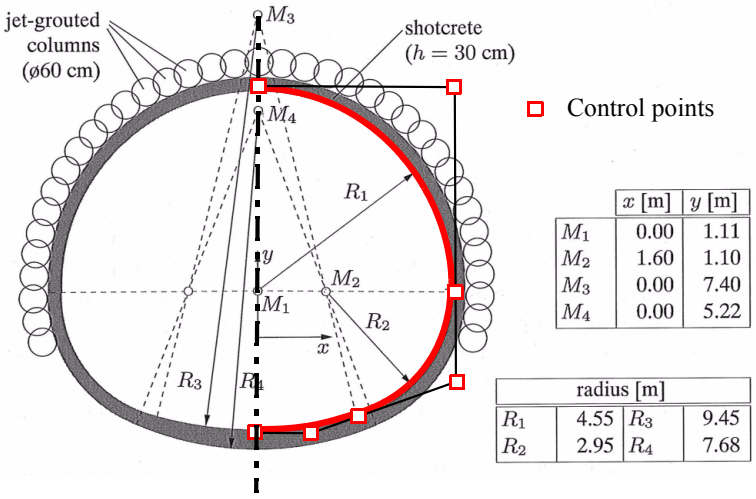
This describes exactly the design geometry. It is noted that some control points do not lie on the curve. The control polygon that connects the control points indicates that there is a smooth transition (unique tangent) between the arcs.

5 Approximation of unknowns with NURBS or B-splines

NURBS or B-spline basis functions are also ideally suited for the approximation of the unknown. In contrast to the commonly used Lagrange polynomials they offer more advanced refinement features. The main difference is that basis functions are not linked to nodal points, but to anchors.

The continuity of B-splines can be controlled by the knot vector. Knots that are repeated at the beginning and the end, control the order of the function. For example knot vectors $\Xi = (0,0,0,1,1,1)$ and $\Xi = (0,0,0,0,1,1,1,1)$ result in basis functions of order 2 and 3 respectively as shown in Figure 3.

In Figure 4 we show how the extent and continuity of the basis functions can also be controlled by the knot vector. In the left we see that the first and last basis function has a limited



	x [m]	y [m]
M_1	0.00	1.11
M_2	1.60	1.10
M_3	0.00	7.40
M_4	0.00	5.22

radius [m]			
R_1	4.55	R_3	9.45
R_2	2.95	R_4	7.68

Figure 2: Example of geometry description with NURBS: NATM Tunnel definition, description of right half with NURBS basis functions, showing control points and control polygon.

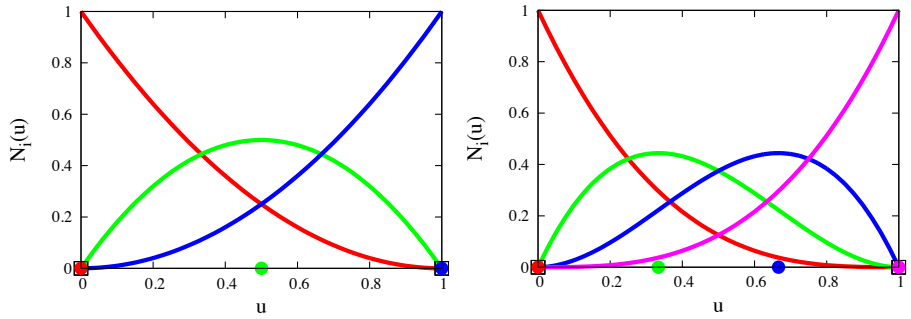


Figure 3: Example of order elevation: Left, B-splines of order 2 and right of order 3. Knot locations are depicted by squares, anchors by filled circles

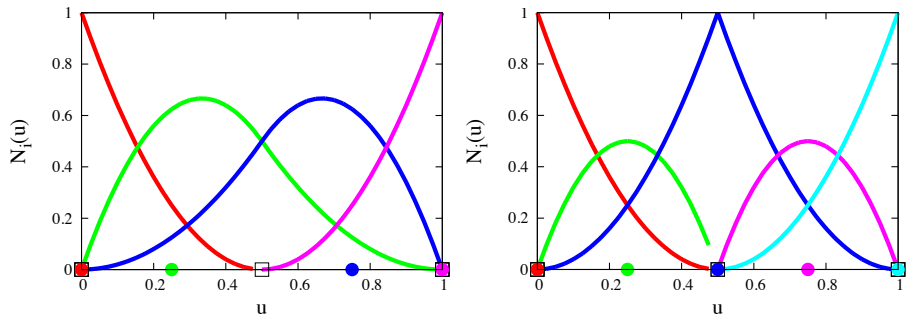


Figure 4: Example of knot insertion: B-splines of order 2 with (left) $\Xi = (0,0,0,0.5,1,1,1)$ and (right) $\Xi = (0,0,0,0.5,0.5,1,1,1)$.

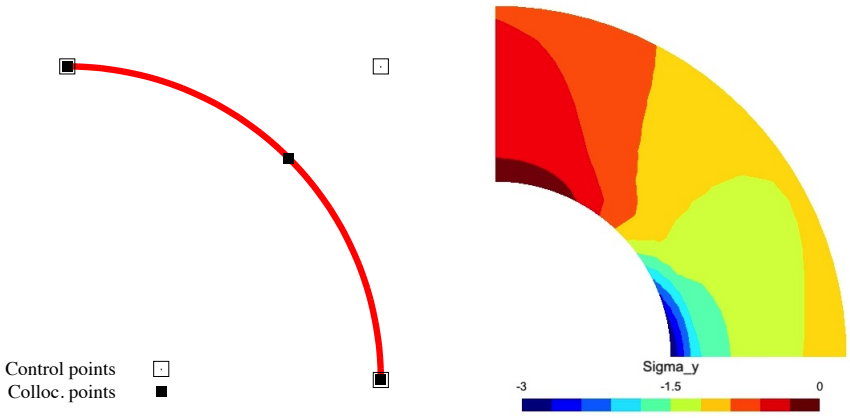


Figure 5: Test example circular excavation: Left definition of a quarter of a circle with one NURBS curve and right resulting stress distribution

span, i.e. span only half the parameter space, whereas in the right we can see that the basis functions are discontinuous at u=0.5.

This means that refinement strategies are simply implemented by changing the knot vector and this is in contrast to isoparametric elements, that involves a change of element type (to higher order) or an increase in the number of elements. The classical p-refinement can be achieved by increasing the number of repetitions at the beginning and end of the knot vector, whereas a refinement similar to the classical h-refinement can be achieved by inserting values into the knot vector. For basis functions of oder 2, for example, inserting one value means that the basis functions have limited span but are still continuous. Repetition of the inserted value means that the basis functions are discontinuous. This refinement strategy would be identical to a classical h-refinement.

In the following we show first examples involving a homogeneous, elastic domain and then consider geological inclusions.

6 Test example

This example is designed to demonstrate how very accurate (or in this case exact) results can be obtained with a very coarse discretization. It relates to the simulation of a circular excavation in an elastic prestressed domain. The following data describes the problem:

E=10; Poissons ratio=0.3, virgin stress $\sigma_{x0} = 0$, $\sigma_{y0} = -1$.

The exact solution of this problem is known. Only one quarter of the problem was analyzed with two axes of symmetry assumed. The geometry is described by one NURBS patch defined with a basis function of order 2 and 3 control points. This exactly describes a quarter circle. The same basis function as for the description of the geometry was used (isogeometric approach) resulting in the collocation points shown in the left of Figure 5. This means that the mesh has only 6 degrees of freedom. The exact solution for the stress concentration is 3.0. As can be seen on the right of Figure 5 the simulation yields the exact solution in this case.

7 Practical example

Here we present an example of the analysis of an NATM tunnel. The geometry of the tunnel is defined by arcs as shown in Figure 2. The tunnel is excavated in a homogeneous elastic domain with the following properties: $E = 10000\,MPa, \nu = 0.3$. The virgin stress field was

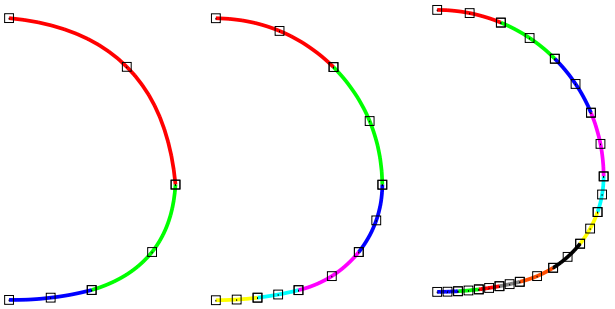


Figure 6: Approximation of geometry of half the tunnel with conventional BEM and a classical h-refinement (Elements are color coded). Nodal points are depicted by hollow squares. Note that the approximation of the geometry changes (i.e. improves) with refinement.

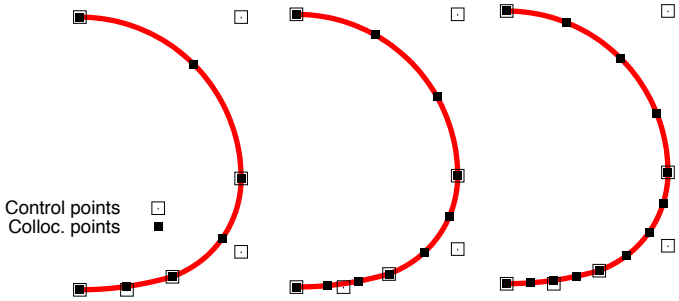


Figure 7: Definition of geometry of half the tunnel using one NURBS patch of order 2. Refinement of solution by order elevation showing collocation points (from left to right oder 2 to 4). Collocation points are depicted by filled squares. Note that in contrast to the above the description of the geometry (already exact) does not change.

assumed to be $\sigma_{x0} = -0.4, \sigma_{y0} = -1.0\,MPa$. We show the salient differences between a conventional and a NURBS based analysis by comparing the two.

7.1 Conventional BEM analysis

In the conventional BEM analysis Lagrange polynomials are used for the approximation of the geometry and the unknown (= isoparametric boundary elements). A refinement of the solution can be obtained either by increasing the order of the element (p-refinement), or by increasing the number of elements (h-refinement). In most cases h-refinement is used. Figure 6 shows the approximation of the geometry of half of the tunnel and of the unknowns by increasing the number of elements. We see that for this refinement strategy both a better approximation of the geometry as well of the unknown is achieved.

7.2 NURBS based approach

Here half of the tunnel is discretized with 1 NURBS patch of order 2. The simulation starts with assuming the same basis functions for the approximation of the displacements as for the geometry. Refinement of the results can be achieved by order elevation or knot insertion. In the isogeometric BEM the collocation points are chosen to be the anchors of the basis functions used for the approximation of the unknown. Figure 7 shows the location of collocation points for the original and the first two stages of refinement by order elevation.

7.3 Results and Comparison

First we compare the variation of the tangential stress[1] along the tunnel surface, as a function of the number of degrees of freedom in Figure 8.

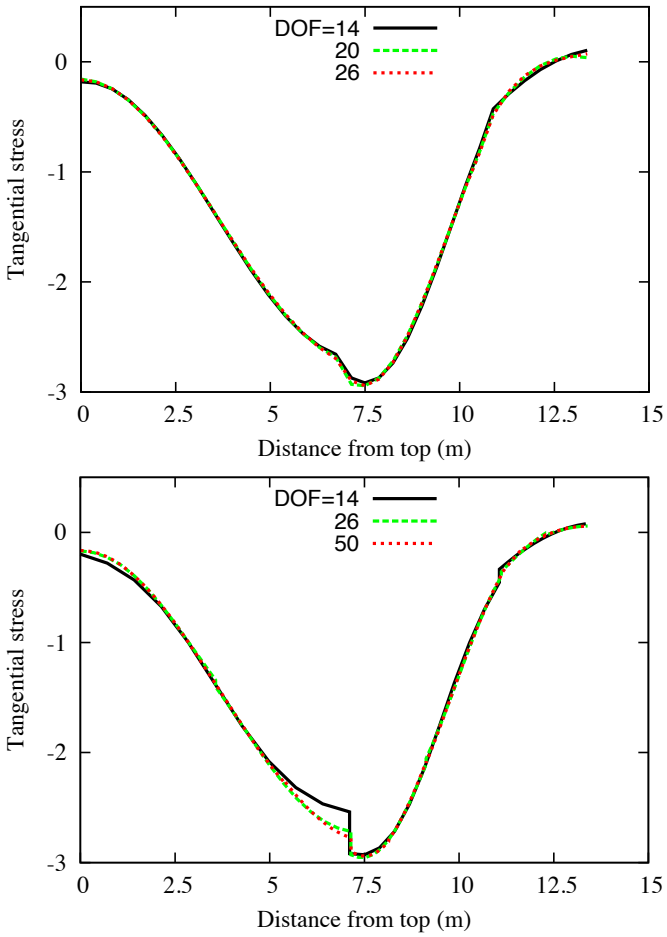


Figure 8: Distribution of tangential stress along tunnel wall for different degrees of freedom. Top: analysis with NURBS, bottom: conventional BEM analysis

It can be seen that even without refinement the NURBS based analysis gives good results, with a smooth stress distribution and very little difference between the solutions with increasing degrees of freedom. The conventional BEM analysis shows a jump in stresses between elements, which decreases as the number of degrees of freedom is increased. This is because the approximation of the geometry with isoparametric boundary elements does not have C^I continuity between elements and a small kink appears which is reduced, as the number of elements is increased. In Figure 9 we show the results of the unrefined NURBS based simulation. It can be seen that good results can be obtained with very few degrees of freedom.

8 Extension of method to include geological features

Being restricted to elastic homogeneous domains severely restricts the practical applicability of the method in geomechanics.

[1]The computation of the tangential stress using the stress recovery method is explained in [3]

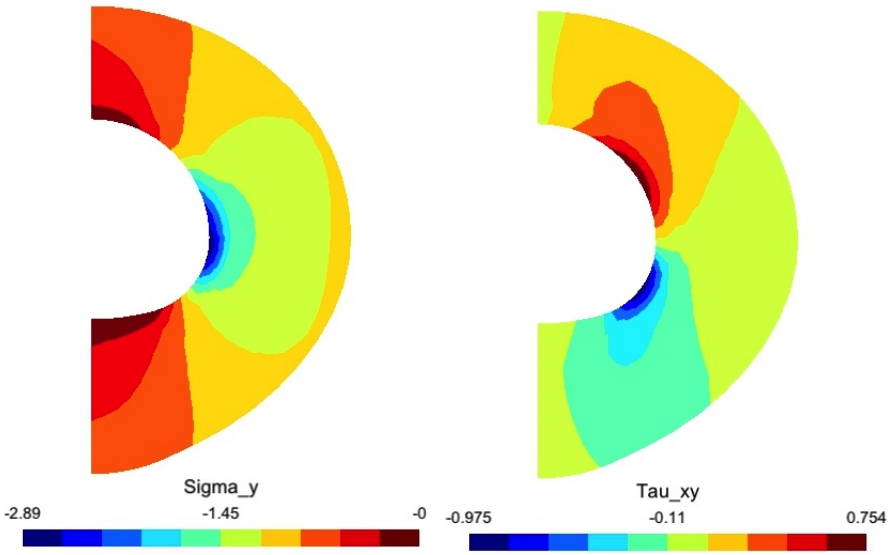


Figure 9: NATM Tunnel: Resulting stress fields

Here we extend the capabilities by considering inclusions that have different material properties and may exhibit non-linear material behavior. As mentioned earlier volume effects will appear, that have to be dealt with. The aim in the implementation will be to avoid additional discretization effort because of this.

The basic approach is to use an iterative solution method. First the problem is solved considering an elastic homogeneous domain. Then the solution is modified to account for the presence of inclusions and inelastic behavior.

The procedure is similar to the initial stress method used in Finite Element work and can be summarized as follows:

1. Solve the elastic, homogeneous problem and determine the increment of stress $\dot{\boldsymbol{\sigma}}$ inside the inclusion V_0.
2. Determine an increment in initial stress $\dot{\boldsymbol{\sigma}}_0$ due to the fact that the elastic material properties of the inclusion are different from the ones used for the fundamental solutions and/or due to the fact that the elastic limit has been exceeded.
3. Convert $\dot{\boldsymbol{\sigma}}_0$ to body force and traction increments $\dot{\mathbf{b}}_0$, $\dot{\mathbf{t}}_0$.
4. Compute new right hand side by evaluating the arising volume and surface integrals.
5. Solve for the new right hand side and compute a new increment of stress $\dot{\boldsymbol{\sigma}}$ inside the inclusion.
6. Repeat 2. to 5. until $\dot{\boldsymbol{\sigma}}_0$ is sufficiently small.

8.1 Elastic inclusions

Elastic inclusions can be modeled with the multi-region method (see for example [2]) and this involves an additional discretization and increases the number of unknowns. Here we include their treatment in the iterative process required for plasticity.

To compute the initial stress increment for the case where the inclusions have elastic properties which are different to the ones used for the fundamental solutions we use the relation between increments of stress $\dot{\boldsymbol{\sigma}}$ and strain $\dot{\boldsymbol{\epsilon}}$ in Voigt notation

$$\dot{\boldsymbol{\sigma}} = \mathbf{C}\,\dot{\boldsymbol{\epsilon}} \tag{11}$$
$$\dot{\boldsymbol{\epsilon}} = \mathbf{C}^{-1}\,\dot{\boldsymbol{\sigma}} \tag{12}$$

where $\mathbf{C}$ is the constitutive matrix for the domain used for the computation of the fundamental solutions. The difference in stress between the inclusion and the domain and therefore the initial stress increment can be computed by

$$\dot{\boldsymbol{\sigma}}_0 = (\mathbf{C}_i - \mathbf{C})\,\dot{\boldsymbol{\epsilon}} \tag{13}$$

where $\mathbf{C}_i$ is the constitutive matrix for the inclusion.

8.2 Inelastic behavior

If the inclusion experiences inelastic behavior then additional initial stresses are generated. Here we use the concept of visco-plasticity, but it is obvious that the method presented here can also be applied to elasto-plasticity. In visco-plasticity we specify a visco-plastic strain rate

$$\frac{\partial \boldsymbol{\epsilon}^{vp}}{\partial t} = \frac{1}{\eta}\Phi(F)\frac{\partial Q}{\partial \boldsymbol{\sigma}} \tag{14}$$

where η is a viscosity parameter, F is the yield function, Q the plastic potential [12]. It holds that

$$\Phi(F) = 0 \quad for \quad F < 0 \tag{15}$$
$$\Phi(F) = F \quad for \quad F > 0. \tag{16}$$

The visco-plastic strain increment during a time increment Δt can be computed by an explicit scheme

$$\dot{\boldsymbol{\epsilon}}^{vp} = \frac{\partial \boldsymbol{\epsilon}^{vp}}{\partial t}\Delta t. \tag{17}$$

The time step Δt can not be chosen freely and if chosen too large, oscillatory behavior will occur in the solution. Suitable time step values can be found in [13]. The initial stress increment is given by

$$\dot{\boldsymbol{\sigma}}_0 = \mathbf{C}\,\dot{\boldsymbol{\epsilon}}^{vp}. \tag{18}$$

8.3 Previous work

The first BEM formulation for inelastic problems has been proposed in [14]. The method has been improved substantially in [15], [16] and [17]. The latter approach proposed an initial strain formulation in which the consistent tangential operator [12] is used to obtain convergence of quadratic order for the iterative solution procedure.

The common approach for the evaluation of the necessary domain integration is to use cells which are identical to isoparametric finite elements. The cell based method for solving inelastic problems with the BEM is explained in detail in [18] and [2]. To overcome the need for a volume discretization, approaches such as the dual reciprocity BEM [19] or the use of radial basis functions [20] have been proposed. A comparison of these methods to the cell based approach found in [21] recommends the latter for accuracy and robustness. Moreover, radial basis

functions are not suitable for the analysis of infinite domains. The possibility to automatically generate cells in the inelastic region has been explored in [22]. In [23] and [24] the cell method is extended to cover the simulation of elastic inclusions with different material properties and in [25] applied to a fast BEM formulation.

All these approaches require the generation of a mesh of cells, which adds an additional effort. Therefore the main innovation presented here is that the concept of cells is abandoned and replaced by a geometry definition of the inclusion as will be explained. It is expected that this approach will not only make the simulation of these simulations more user friendly but we expect also an increase in accuracy of the results, because the approximation of initial stress inside cells with basis functions is avoided. An important aspect regarding accuracy is that most published cell based methods a continuous variation of initial stresses is assumed, regardless of the fact that the elasto-plastic boundary may cut through a cell and that in this case the variation of the initial stress is discontinuous.

In the following it is first outlined how the geometry of inclusions is defined using NURBS curves and how the arising volume and surface integrals are numerically evaluated.

9 Geometry definition for inclusions

The first task is the description of the geometry of the subdomain V_0. For this we propose to use a mapping method introduced recently for trimmed surfaces in [9] and [3]. This means that the subdomain is defined by two NURBS curves and a linear interpolation between them.

We establish a local coordinate system $\mathbf{s} = (s,t)^\intercal = [0,1]^2$ as shown in Figure 10 and perform all computations such as integration and differentiation in this system and then map it to the global x,y-system. Note that there is a one to one mapping between the coordinate $s = [0,1]$ and coordinate u of the red and green NURBS curve in Figure 10 . The global coordinates of a point $\mathbf{x}$ with the local coordinates $\mathbf{s}$ are given by

$$\mathbf{x}(s,t) = (1-t)\,\mathbf{x}^I(s) + t\,\mathbf{x}^{II}(s) \tag{19}$$

where

$$\mathbf{x}^I(s) = \sum_{k=1}^{K^I} R_k^I(s)\,\mathbf{x}_k^I \quad \text{and} \quad \mathbf{x}^{II}(s) = \sum_{k=1}^{K^{II}} R_k^{II}(s)\,\mathbf{x}_k^{II}. \tag{20}$$

The superscript I relates to the bottom (red) curve and II to the top (green) curve and $\mathbf{x}_k^I$, $\mathbf{x}_k^{II}$ are control point coordinates. K^I and K^{II} are the number of control points, $R_k^I(s)$ and $R_k^{II}(s)$ are NURBS basis functions. The derivatives are given by

$$\begin{aligned} \frac{\partial \mathbf{x}(s,t)}{\partial s} &= (1-t)\,\frac{\partial \mathbf{x}^I(s)}{\partial s} + t\,\frac{\partial \mathbf{x}^{II}(s)}{\partial s} \\ \frac{\partial \mathbf{x}(s,t)}{\partial t} &= -\mathbf{x}^I(s) + \mathbf{x}^{II}(s) \end{aligned} \tag{21}$$

where

$$\begin{aligned} \frac{\partial \mathbf{x}^I(s)}{\partial s} &= \sum_{k=1}^{K^I} \frac{\partial R_k^I(s)}{\partial s}\,\mathbf{x}_k^I \\ \frac{\partial \mathbf{x}^{II}(s)}{\partial s} &= \sum_{k=1}^{K^{II}} \frac{\partial R_k^{II}(s)}{\partial s}\,\mathbf{x}_k^{II}. \end{aligned} \tag{22}$$

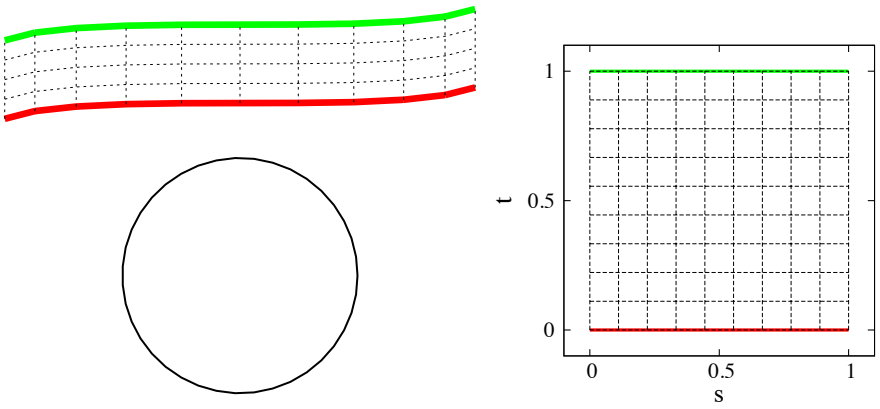


Figure 10: A circular excavation with an inclusion above it: Definition of the geometry of the inclusion with 2 NURBS curves in left global and right local coordinate space

The Jacobian matrix of this mapping is

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \\ \\ \frac{\partial x}{\partial t} & \frac{\partial y}{\partial t} \end{pmatrix} \tag{23}$$

and the Jacobian is $J(\mathbf{s}) = |\mathbf{J}|$.

Remark: The geometry definition just outlined is suitable for geological inclusions of tabular shape, as are defined in the following examples. However, the theory presented is not restricted to this geometry definition and any definition that involves a mapping to a local coordinate system $\mathbf{s}$ can be used.

10 Computation of $\{\mathbf{F}\}_0$

Here we discuss the computation of the right hand side during iteration. This involves the evaluation of the integrals in Equation (2) over S_0 and V_0 of the inclusion using numerical integration (Gauss quadrature). First the initial stresses have to be converted to tractions acting on boundary S_0 and body forces acting in the domain V_0. The traction increments are given by:

$$\dot{\mathbf{t}}_0 = \begin{pmatrix} \dot{\sigma}_{0x} & \dot{\tau}_{0xy} \\ \dot{\tau}_{0xy} & \dot{\sigma}_{0y} \end{pmatrix} \mathbf{n} \tag{24}$$

where $\mathbf{n}$ is the unit outward normal vector to the surface S_0. The body force increment $\dot{\mathbf{b}}_0$ can be computed by

$$\dot{\mathbf{b}}_0 = -\begin{pmatrix} \frac{\partial \dot{\sigma}_{0x}}{\partial x} + \frac{\partial \dot{\tau}_{0xy}}{\partial y} \\ \\ \frac{\partial \dot{\tau}_{0xy}}{\partial x} + \frac{\partial \dot{\sigma}_{0x}}{\partial y} \end{pmatrix} \tag{25}$$

It is convenient to compute the derivatives with respect to local coordinates $\mathbf{s}$ first and then transform them to global coordinates. For example the global derivatives of σ_x in terms of local derivatives are given by the transformation

$$\boldsymbol{\sigma}_{x,\mathbf{x}} = \mathbf{J}^{-1}\,\boldsymbol{\sigma}_{x,\mathbf{s}} \tag{26}$$

where

$$\boldsymbol{\sigma}_{x,\mathbf{x}} = \begin{pmatrix} \frac{\partial \sigma_x}{\partial x} \\ \frac{\partial \sigma_x}{\partial y} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\sigma}_{x,\mathbf{s}} = \begin{pmatrix} \frac{\partial \sigma_x}{\partial s} \\ \frac{\partial \sigma_x}{\partial t} \end{pmatrix} \tag{27}$$

and $\mathbf{J}$ is the Jacobian matrix (23).

The derivatives are numerically computed using finite differences. We define a regular grid of points inside the inclusion in the local coordinate system $\mathbf{s}$. For grid points that have other points left and right (or top and bottom) of them we use a central finite difference, whereas for points that only have one point on a side we use forward or backward finite differences.

Remark: In the following examples a simple linear interpolation between grid points and a linear extrapolation from grid points to the boundary S_0. Obviously more sophisticated schemes may be applied. However care has to be taken that the variation of the initial stress may be discontinuous. The simple scheme applied here leads to an increase in the accuracy for determination of the derivatives and for the evaluation of the associated integrals as the number of grid points is increased. The convergence of the solution as a function of the number of internal points is investigated in [26].

The numerical evaluation of $\{\mathbf{F}\}_0^{S_0}$ and $\{\mathbf{F}\}_0^{V_0}$ is explained in detail in [26].

11 Computation of results inside the inclusion

As mentioned above, the solution algorithm requires the evaluation of strains and stresses at internal points. For the initial solution (without body forces) the displacements at a point $\mathbf{y}_i$ inside the inclusion V_0 can be computed by

$$\mathbf{u}(\mathbf{y}_i) = \int_S \mathbf{U}(\mathbf{y}_i, \mathbf{x})\, \mathbf{t}(\mathbf{x})\, \mathrm{d}S - \int_S \mathbf{T}(\mathbf{y}_i, \mathbf{x})\, \mathbf{u}(\mathbf{x})\, \mathrm{d}S. \tag{28}$$

The strain at a point $\mathbf{y}_i$ inside the inclusion is

$$\boldsymbol{\epsilon}(\mathbf{y}_i) = \int_S \mathbf{S}(\mathbf{y}_i, \mathbf{x})\, \mathbf{t}(\mathbf{x})\, \mathrm{d}S - \int_S \mathbf{R}(\mathbf{y}_i, \mathbf{x})\, \mathbf{u}(\mathbf{x})\, \mathrm{d}S \tag{29}$$

where $\mathbf{S}$ and $\mathbf{R}$ are derived fundamental solutions [2]. The stresses can be computed using (11).

For the subsequent solution (including body forces) the displacements at a point $\mathbf{y}_i$ inside the inclusion is

$$\begin{aligned} \mathbf{u}(\mathbf{y}_i) = &\int_S \mathbf{U}(\mathbf{y}_i, \mathbf{x})\, \mathbf{t}(\mathbf{x})\, \mathrm{d}S - \int_S \mathbf{T}(\mathbf{y}_i, \mathbf{x})\, \mathbf{u}(\mathbf{x})\, \mathrm{d}S \\ &+ \int_{S_0} \mathbf{U}(\mathbf{y}_i, \bar{\mathbf{x}})\, \dot{\mathbf{t}}_0(\bar{\mathbf{x}})\, \mathrm{d}S_0 + \int_{V_0} \mathbf{U}(\mathbf{y}_i, \bar{\mathbf{x}})\, \dot{\mathbf{b}}_0(\bar{\mathbf{x}})\, \mathrm{d}V_0 \end{aligned} \tag{30}$$

The strain can be computed by

$$\begin{aligned} \boldsymbol{\epsilon}(\mathbf{y}_i) = &\int_S \mathbf{S}(\mathbf{y}_i, \mathbf{x})\, \mathbf{t}(\mathbf{x})\, \mathrm{d}S - \int_S \mathbf{R}(\mathbf{y}_i, \mathbf{x})\, \mathbf{u}(\mathbf{x})\, \mathrm{d}V \\ &+ \int_{S_0} \mathbf{S}(\mathbf{y}_i, \bar{\mathbf{x}})\, \dot{\mathbf{t}}_0(\bar{\mathbf{x}})\, \mathrm{d}S_0 + \int_{V_0} \mathbf{S}(\mathbf{y}_i, \bar{\mathbf{x}})\, \dot{\mathbf{b}}_0(\bar{\mathbf{x}})\, \mathrm{d}V_0 \end{aligned} \tag{31}$$

Again, for the evaluation of the integrals Gauss integration is used and is described in detail in [26].

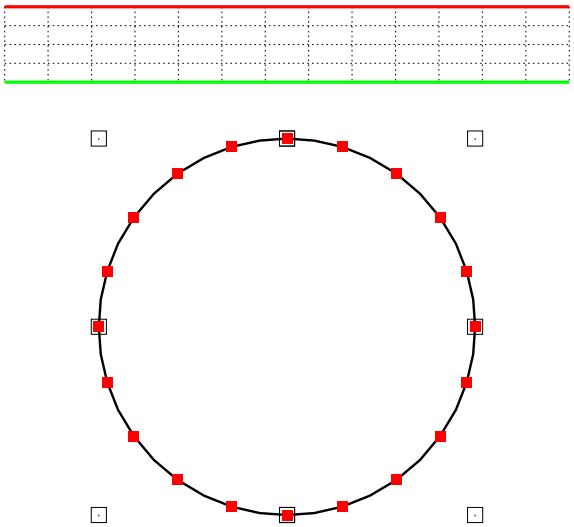


Figure 11: Test example 1: Definition of the geometry of the excavation and the inclusion with 2 NURBS curves. Control points for the excavation are shown as hollow squares, red squares indicate the collocation points used for the analysis.

12 Test examples

The theory is tested on two simple examples here and comparison is made with a coupled BEM/FEM analysis.

12.1 Example 1: Elastic inclusion above circular excavation

Here we test the capability of the method to simulate elastic inclusions. The example is that of a circular excavation in an infinite domain with the following properties:

E=1; Poissons ratio=0, virgin stress $\sigma_{x0} = 0$, $\sigma_{y0} = -1$.

Above the circular excavation there is an elastic inclusion with the elastic modulus of half the one assumed for the domain. Figure 11 shows the geometry definition of the excavation and the inclusion. The solution was achieved by order elevating the basis functions used for the description of the geometry three times. Figure 12 shows that the simulation converges after 8 iterations and that the end result compares well with a coupled Boundary/Finite element analysis using the software BEFE. The results of the simulation are shown in Figure 13.

12.2 Example 2: Elasto-plastic inclusion above circular excavation

This is the same example as the previous one except that the inclusion has the same elastic properties as the domain but has elasto-plastic properties. In this case a laminate plasticity model was applied with a horizontal orientation of the lamina and a Mohr-Coulomb yield condition. The properties assumed were:

Angle of friction $\phi : 10^\circ$, c: 0, non-associative flow law with $\psi : 0$.

The convergence of the maximum displacement after about 10 iterations is shown in Figure 14 and compared with a coupled analysis using BEFE.

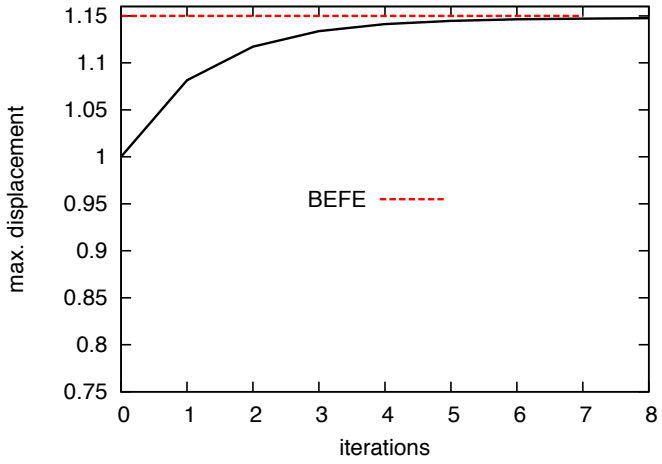


Figure 12: Convergence of maximum displacement as a function of the number of iterations

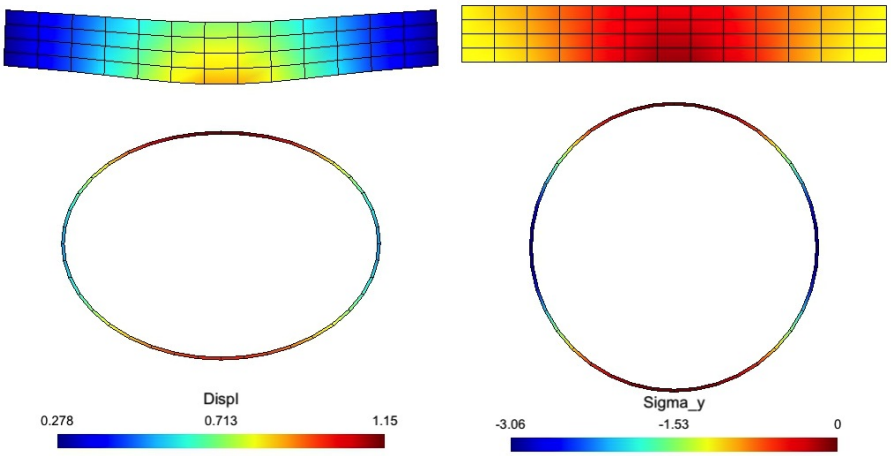


Figure 13: Results of the simulation. Left: Displaced shape; Right: Vertical stress

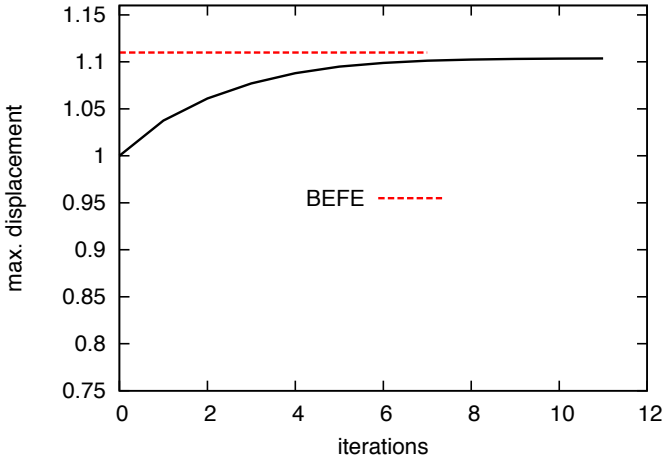


Figure 14: Example 2: Convergence of maximum displacement as a function of the number of iterations

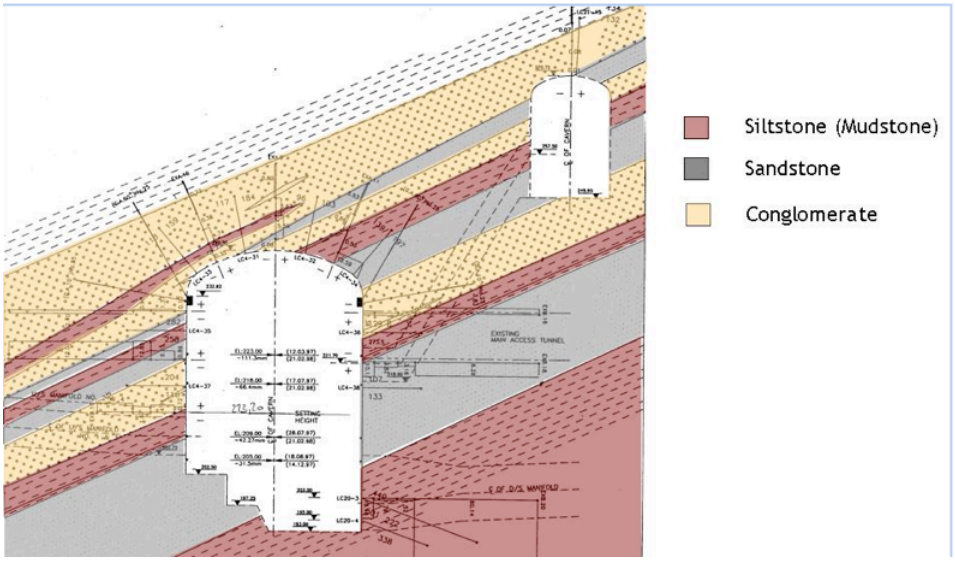


Figure 15: Practical example: Geometry of cavern showing the geology.

Rock mass	
Young's modulus	E=10000 MPa
Poisson's ratio	ν=0.20
Inclusion	
Young's modulus	E_i=6000 MPa
Poisson's ratio	ν_i=0.25
Mohr-Coulomb yield condition	
Angle of friction	ϕ=30°
Cohesion	c=0.73 MPa
Virgin stress field	σ_{xv}=−4 MPa
	σ_{yv}=−8 MPa

Table 1: Practical example: Material parameters and stress field

13 Practical example

The practical example is one that has been solved with a coupled BEM/FEM method and reported in [2]. It relates to the plane strain analysis of an underground power station cavern. Figure 15 shows a sketch of the final excavation stage together with the geology which basically consists of mudstone, sandstone and conglomerate. The mudstone is the weakest material and has the most profound effect on the ground behaviour. Therefore these layers have been modelled in the simulation reported in [2].

The material parameters assumed in the simulation are summarised in Table 1. For revisiting the coupled BEM/FEM analysis with the novel simulation method, using BEM only we consider an excavation stage to level 232 m and a limited extent of the inclusions. It should be noted here that no truncation of the mesh is necessary, as the solution obeys the radiation condition. The limited extent of the inclusions simply means that the effect of the mudstone layers diminishes at a distance from the excavation surface.

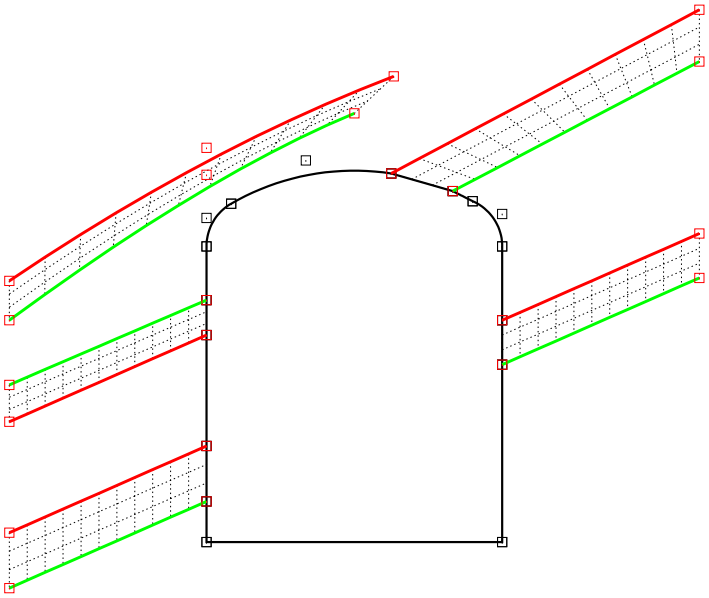


Figure 16: Practical example: The geometry of the cavern is described by 14 NURBS patches. The associated control points are shown as hollow black squares. Five inclusions are described by bounding curves with the associated control points shown as red hollow squares

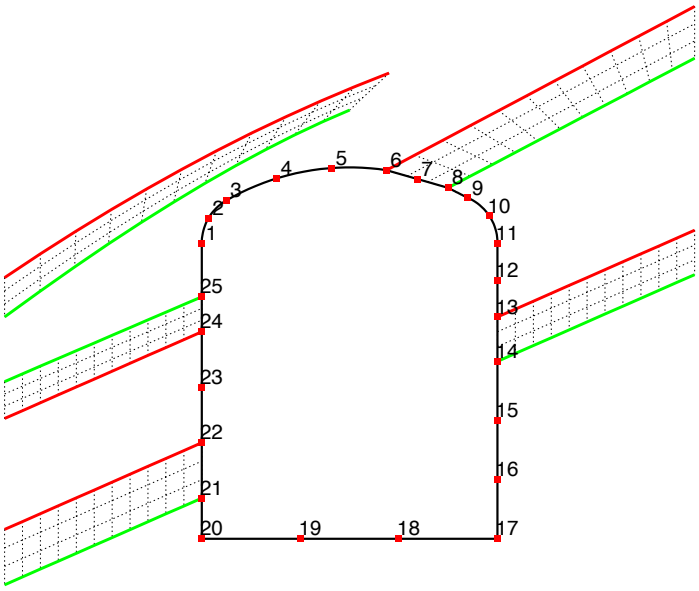


Figure 17: Practical example: Figure showing the location of collocation points

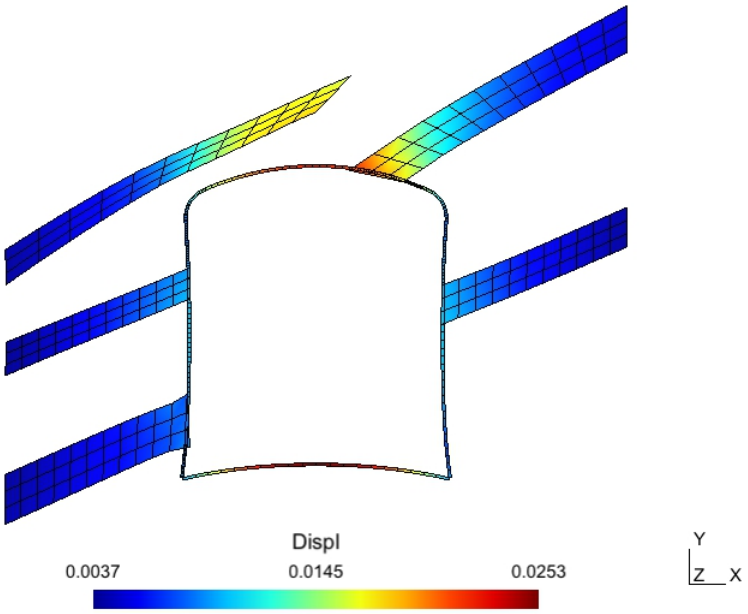


Figure 18: Practical example: Displaced shape

Figure 16 shows the geometry of the cavern with the discretization into 14 NURBS patches. For the analysis some basis functions for describing the geometry were order elevated, resulting the the collocation points shown in Figure 17. The resulting displacements are shown in Figure 18.

14 Summary and Conclusions

The aim of the paper was to show recent advances of the Boundary Element Method with applications in geomechanics. It has been demonstrated that the use of NURBS basis functions results in a significantly more efficient simulation with fewer parameters required for accurately describing the geometry and fewer degrees of freedom. It was also shown that NURBS technology can be used for the definition of geological inclusions with few parameters, resulting in more efficient simulation of these problems. The claims have been verified by test examples and a practical application. An extension of the theory to 3-D problems is underway.

REFERENCES

[1] T. Hughes, J. Cottrell, Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement, Computer Methods in Applied Mechanics and Engineering 194 (39–41) (2005) 4135–4195.
URL http://www.sciencedirect.com/science/article/pii/S0045782504005171

[2] G. Beer, I. Smith, C. Duenser, The Boundary Element Method with Programming, Springer-Verlag, Wien, 2008.

[3] G. Beer, Advanced numerical simulation methods - From CAD Data directly to simulation results, CRC Press/Balkema, 2015.

[4] R. Simpson, S. Bordas, J. Trevelyan, T. Rabczuk, A two-dimensional isogeometric boundary element method for elastostatic analysis, Computer Methods in Applied Mechanics and Engineering 209–212 (0) (2012) 87–100.
URL http://www.sciencedirect.com/science/article/pii/S0045782511002635

[5] M. Scott, R. Simpson, J. Evans, S. Lipton, S. Bordas, T. Hughes, T. Sederberg, Isogeometric boundary element analysis using unstructured T-splines, Computer Methods in Applied Mechanics and Engineering 254 (0) (2013) 197 – 221. doi:10.1016/j.cma.2012.11.001.
URL http://www.sciencedirect.com/science/article/pii/S0045782512003386

[6] B. Marussig, G. Beer, C. Duenser, Isogeometric boundary element method for the simulation in tunneling, Applied Mechanics and Materials 553 (2014) 495–500.

[7] G. Beer, B. Marussig, J. Zechner, C. Duenser, T.-P. Fries, Boundary Element Analysis with trimmed NURBS and a generalized IGA approach, in: E. Oñate, J. Oliver, A. Huerta (Eds.), 11th World Congress on Computational Mechanics (WCCM XI), 2014.

[8] B. Marussig, J. Zechner, G. Beer, T.-P. Fries, Fast isogeometric boundary element method based on independent field approximation, Computer Methods in Applied Mechanics and Engineering 284 (0) (2015) 458 – 488, isogeometric Analysis Special Issue.

[9] G. Beer, B.Marussig, J. Zechner, A simple approach to the numerical simulation with trimmed CAD surfaces, Computer Methods in Applied Mechanics and Engineering 285 (2015) 776–790.

[10] G. Beer, Mapped infinite patches for the NURBS based boundary element analysis in geomechanics, Computers and Geotechnics 66 (2015) 66–74.

[11] G. Beer, S. Bordas (Eds.), Isogeometric methods for numerical simulation, CISM lecture notes, Springer, 2014.

[12] J. Simo, T. Hughes, Computational Inelasticity, Springer, 1998.

[13] I. Cormeau, Numerical stability in quasi-static elasto-viscoplasticity, International Journal for Numerical Methods in Engineering 9 (1).

[14] J. Swedlow, T. Cruse, Formulation of boundary integral equations for three-dimensional elasto-plastic flow, International Journal of Solids and Structures 7 (12) (1971) 1673 – 1683. doi:10.1016/0020-7683(71)90006-0.
URL http://www.sciencedirect.com/science/article/pii/0020768371900060

[15] J. C. F. Telles, C. A. Brebbia, On the application of the boundary element method to plasticity, Applied Mathematical Modelling 3 (1979) 466–470.

[16] J. C. F. Telles, J. A. M. Carrer, Implicit procedures for the solution of elastoplastic problems by the boundary element method, Mathematical and Computer Modelling 15 (1991) 303–311.

[17] M. Bonnet, S. Mukherjee, Implicit bem formulations for usual and sensitivity problems in elasto-plasticity using consistent tangent operator concept, International Journal of Solids and Structures 33 (30) (1996) 4461–4480.

[18] X.-W. Gao, T. G. Davies, Boundary Element Programming in Mechanics, Cambridge University Press, Cambridge, 2002.

[19] D. P. Henry, P. K. Banerjee, A new bem formulation for two- and three-dimensional elastoplasticity using particular integrals, International Journal for Numerical Methods in Engineering 26 (9) (1988) 2079–2096. doi:10.1002/nme.1620260912.
URL http://dx.doi.org/10.1002/nme.1620260912

[20] X.-W. Gao, A boundary element method without internal cells for two-dimensional and three-dimensional elastoplastic problems, Journal of Applied Mechanics 69 (2) (2002) 154–160. doi:10.1115/1.1433478.
URL http://link.aip.org/link/?AMJ/69/154/1

[21] M. Ingber, A. Mammoli, M. Brown, A comparison of domain integral evaluation techniques for boundary element methods, International Journal for Numerical Methods in Engineering 52 (2001) 417–432.

[22] T. Ribeiro, G. Beer, C. Duenser, Efficient elastoplastic analysis with the boundary element method, Computational Mechanics 41 (2008) 715–732.
URL http://dx.doi.org/10.1007/s00466-007-0227-1

[23] K. Riederer, C. Duenser, G. Beer, Simulation of linear inclusions with the BEM, Engineering Analysis with Boundary Elements 33 (7) (2009) 959 – 965. doi:10.1016/j.enganabound.2009.01.003.
URL http://www.sciencedirect.com/science/article/pii/S0955799709000125

[24] K. Riederer, Modelling of Ground Support in Tunnelling using the BEM, Ph.D. thesis, Graz University of Technology (2010).

[25] J. Zechner, G. Beer, A fast elasto-plastic formulation with hierarchical matrices and the boundary element method, Computational Mechanics 51 (4) (2013) 443–453.

[26] G. Beer, B. Marussig, J. Zechner, C. Duenser, T. P. Fries, Isogeometric boundary element analysis with elasto-plastic inclusions. part 1: plane problems, Computer Methods in Applied Mechanics and Engineering.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ENERGETIC BEM FOR THE NUMERICAL SOLUTION OF DAMPED WAVE PROPAGATION EXTERIOR PROBLEMS

A. Aimi[1], **M. Diligenti**[1] **and C. Guardasoni**[1]

[1]Department of Mathematics and Computer Science, University of Parma, Italy
e-mail: {alessandra.aimi, mauro.diligenti, chiara.guardasoni}@unipr.it

Keywords: damped wave propagation, space-time boundary integral equation, energetic boundary element method

Abstract. *The analysis of damping phenomena that occur in many physics and engineering problems, such as fluid dynamics, kinetic theory and semiconductors, is of particular interest. For this kind of problems, one needs accurate and stable approximate solutions even on large time intervals. These latter can be obtained reformulating time-dependent problems modeled by partial differential equations (PDEs) of hyperbolic type in terms of boundary integral equations (BIEs) solved via boundary element methods (BEMs).*
In this context, starting from a recently developed energetic weak formulation of the space-time BIE modeling, in particular, classical wave propagation exterior problems [2, 3], we consider here an extension for the damped wave equation in 2D space dimension, based on successful simulations for the 1D case [6, 7].
In fact, the related energetic BEM reveals a robust time stability property, which is crucial in guaranteeing accurate numerical solutions on large time intervals. Several benchmarks will be presented and discussed.

1 INTRODUCTION

A variety of engineering and physical applications, such as the propagation or the scattering of acoustic or electromagnetic waves, leads to the problem of solving linear hyperbolic partial differential equations in two or three dimensional space. These problems are normally considered in an unbounded homogeneous domain and a method to tackle them is to reformulate the partial differential equation as an integral equation on the, usually bounded, boundary of the domain (BIE) which can then be solved using the boundary element method (BEM) ([11], [19]). The BIE-BEM combination is especially useful for solving problems of practical importance with irregular geometries. In some applications, the physically relevant data are given not by the solution in the interior of the domain but rather by the boundary values of the solution or its derivatives. These data can be obtained directly from the solution of boundary integral equations, whereas boundary values obtained from finite element method (FEM) solutions are in general not very accurate. Sometimes, however, a coupling of FEM and BEM proves to be useful. Also the analysis of damping phenomena that occur, for example, in fluid dynamics and in kinetic theory is of particular interest [8, 16, 23]. The use of advanced numerical techniques to solve PDEs, such as the finite element method (FEM) and the finite difference (FD) methods, for instance in structural mechanics and for fluid flow calculations, is now well established.
On the other side, both frequency-domain and time-domain boundary element method (BEM) can be used for hyperbolic boundary value problems [9, 10].
The most frequently adopted discretization scheme is the collocation technique [12, 15] with direct step-by-step evaluation of the time convolution, even if this approach has never reached the same level of maturity as in the frequency domain: the high computing costs of the time convolution and the limited robustness associated to instability phenomena have hindered considerably its diffusion. This is partly because accuracy and stability of the solution are affected by the time step size. Properties of the numerical solution by this method have been already studied in [14]. The former issue has been addressed in several ways. The time convolution can be evaluated applying the convolution quadrature method which has been developed initially in [20, 21] and has since then been successfully applied to many applications (see e.g.[22]). This method has the fundamental property of not using explicitly the expression of the kernel which is instead replaced by its Laplace transform.
On the contrary, the latter problem has been recently tackled in [1, 5, 4] where an energetic direct space-time Galerkin BEM for the discretization of retarded potential BIEs related to 1D, 2D and 3D wave propagation problems has been put forward and compared with the weak formulation due to Bamberger and Ha-Duong [9, 10, 17, 18]. The proposed technique is based on a natural energy identity satisfied by the solution of the corresponding differential problem, which leads to a space-time weak formulation of the BIEs with precise stability properties. Consequently, the integral problem can be discretized by unconditionally stable schemes via the so-called energetic BEM.
In this context, starting from the application of energetic BEM to classical wave propagation exterior problems [2, 3], we consider here an extension for the damped wave equation in 2D space dimension, based on successful simulations for the 1D case [6, 7]. Several benchmarks will be presented and discussed.

2 MODEL PROBLEM and its BIE ENERGETIC WEAK FORMULATION

We will analyze the following initial-boundary value problem for damped wave equation exterior to an open arc $\Gamma \subset \mathbf{R}^2$

$$\Delta u(x,t) - \frac{1}{c^2}\ddot{u}(x,t) - \frac{2D}{c^2}\dot{u}(x,t) - \frac{P}{c^2}u(x,t) = 0 \qquad x \in \mathrm{R}^2 \setminus \Gamma,\ t \in [0,T] \tag{1}$$

$$u(x,0) = 0 \qquad x \in \Gamma \tag{2}$$

$$\dot{u}(x,0) = 0 \qquad x \in \Gamma \tag{3}$$

$$u(x,t) = \bar{u}(x,t) \qquad x \in \Gamma,\ t \in (0,T) \tag{4}$$

where overhead dots indicate derivatives with respect to time, c is the propagation velocity of a perturbation in the domain, $\bar{u}(x,t)$ is the Dirichlet boundary datum; the viscous damping term is characterized by the coefficient D, while the material damping is modeled by the coefficient P.

Since the goal of this paper is to approximate u using a BEM technique, we have to obtain a boundary integral reformulation of the problem (1)-(4) over Γ.

For this purpose, using classical arguments [13], let us consider the single layer space-time integral representation formula of $u(x,t)$:

$$u(x,t) = \int_\Gamma \int_0^t G(r, t-\tau)\phi(\xi,\tau)d\tau d\gamma_\xi =: \mathcal{V}\phi(x,t), \qquad x \in \mathrm{R}^2 \setminus \Gamma,\ t \in [0,T] \tag{5}$$

where $r = \|x-\xi\|_2$, $\phi(x,t) = \left[\frac{\partial u}{\partial \mathbf{n}_x}(x,t)\right]_\Gamma$ is the jump of $u(x,t)$ along Γ and $G(x,t)$ is the forward fundamental solution of the two-dimensional damped wave operator, that is

$$G(r,t) = \begin{cases} \frac{c}{2\pi}\, e^{-Dt}\, \frac{\cos\left(\frac{\sqrt{P-D^2}}{c}\sqrt{c^2t^2-r^2}\right)}{\sqrt{c^2t^2-r^2}} H[ct-r]\,, & P \geq D^2 \\ \\ \frac{c}{2\pi}\, e^{-Dt}\, \frac{\cosh\left(\frac{\sqrt{D^2-P}}{c}\sqrt{c^2t^2-r^2}\right)}{\sqrt{c^2t^2-r^2}} H[ct-r]\,, & P < D^2 \end{cases} \tag{6}$$

with $H[\cdot]$ the Heaviside distribution.

Now, with a limiting process for x tending to Γ and using the Dirichlet datum (4), one obtains the following BIE

$$\mathcal{V}\phi(x,t) = \overline{u}(x,t)\,, \quad x \in \Gamma,\ t \in [0,T] \tag{7}$$

whose energetic weak formulation reads (see [2]):

find $\phi \in H^0([0,T], H^{-1/2}(\Gamma))$ *such that*

$$\langle(\dot{\mathcal{V}}\phi), \psi\rangle_{\Gamma\times L^2([0,T])} = \langle\dot{\bar{u}}, \psi\rangle_{\Gamma\times L^2([0,T])}\,, \tag{8}$$

where ψ *is a suitable test function belonging to the same functional space of* ϕ.

Note that, performing a time integration by parts in the sense of distributions, the equation (8) can be equivalently rewritten as

$$\langle(\mathcal{V}\phi), \dot{\psi}\rangle_{\Gamma\times L^2([0,T])} = \langle\overline{u}, \dot{\psi}\rangle_{\Gamma\times L^2([0,T])}\,. \tag{9}$$

3 SPACE-TIME GALERKIN APPROXIMATION

For time discretization we consider a uniform decomposition of the time interval $[0,T]$ with time step $\Delta t = T/N_{\Delta t}$, $N_{\Delta t} \in \mathrm{N}^+$, generated by the $N_{\Delta t}+1$ time instants: $t_k = k\Delta t,\ k = 0,\ldots,N_{\Delta t}$ and we choose piecewise constant shape functions for the time approximation of ϕ.

In particular, for $k = 0, \dots, N_{\Delta t} - 1$, time shape functions for the approximation of ϕ will be defined as

$$\overline{\varphi}_k(t) = H[t - t_k] - H[t - t_{k+1}]\,. \tag{10}$$

For the space discretization we consider a polygonal approximation of boundary Γ denoted by $\widehat{\Gamma}$ and constituted by $N_{\Delta x}$ straight elements $e_i, i = 1, \dots, N_{\Delta x}$, with $length(e_i) \leq \Delta x$, $e_i \cap e_j = \emptyset$ if $i \neq j$ and such that $\cup_{i=1}^{N_{\Delta x}} \bar{e}_i = \widehat{\Gamma}$.
The functional background compels one to choose spatial shape functions belonging to $L^2(\hat{\Gamma})$ for the approximation of ϕ. Hence, having defined $\mathcal{P}_{d_i}$ the space of algebraic polynomials of degree d_i on the element e_i of $\widehat{\Gamma}$, we consider the space of piecewise polynomial functions

$$X_{\Delta x}^{-1} := \{\widetilde{\varphi}(x) \in L^2(\widehat{\Gamma}) \,:\, \widetilde{\varphi}_{\,|e_i} \in \mathcal{P}_{d_i}\,,\, \forall\, e_i \subset \widehat{\Gamma}\,\}\,.$$

Hence, denoting with M, the number of unknowns on $\widehat{\Gamma}$, and having introduced in $X_{\Delta x}^{-1}$ the standard piecewise polynomial boundary element basis functions $\varphi_j(x)$, $j = 1, \cdots, M$, the approximate solutions of the problem at hand will be expressed as

$$\widehat{\phi}(x,t) := \sum_{k=0}^{N_{\Delta t}-1} \sum_{i=1}^{M} \alpha_i^k\, \widetilde{\varphi}_i(x)\, \overline{\varphi}_k(t)\,. \tag{11}$$

Energetic Galerkin BEM is obtained inserting the introduced discretization into the weak problem (9) and its algebraic reformulation consists in a linear system: $A\Phi = b$ of order $(M \cdot N_{\Delta t})$, whose matrix elements are linear combination of integrals of the form

$$\int_{\widehat{\Gamma}} \widetilde{\varphi}_j(x) \int_0^T \dot{\overline{\varphi}}_h(t) \int_{\widehat{\Gamma}} \widetilde{\varphi}_i(\xi) \int_0^t G(r, t-\tau)\overline{\varphi}_k(\tau) d\tau d\gamma_\xi dt d\gamma_x\,. \tag{12}$$

Specifying the choice made for time basis function, from (12) one obtains a combination of integrals of the form

$$\int_{\widehat{\Gamma}} \widetilde{\varphi}_j(x) \int_{\widehat{\Gamma}} \widetilde{\varphi}_i(\xi) \int_0^{t_h} G(r, t_h - \tau) H[\tau - t_k] d\tau d\gamma_\xi d\gamma_x\,. \tag{13}$$

The analysis of kernel singularities has been performed for the case of the wave equation without damping, i.e. for $P = D = 0$, in [2]; in particular the presence of the Heaviside distribution $H[ct - r]$ and of the square root $\sqrt{c^2t^2 - r^2}$ can cause a lot of numerical troubles that in [2] have been solved by suitable splitting of the outer integral over $\widehat{\Gamma}$ and using quadrature schemes which regularize integrand functions with mild singularities for the second nested integral in space variable. Since we expect a similar behavior for the damped kernel, having set

$$G_0(r,t) = \frac{c}{2\pi} \frac{1}{\sqrt{c^2t^2 - r^2}} H[c\,t - r] \tag{14}$$

the fundamental solution related to the classical wave operator, we consider the expansion of $G(r,t)$ with respect to damping parameters, centered in $P = D = 0$, and we rewrite (13) as

$$\begin{aligned}
&\int_{\widehat{\Gamma}} \widetilde{\varphi}_j(x) \int_{\widehat{\Gamma}} \widetilde{\varphi}_i(\xi) \int_0^{t_h} [G(r, t_h - \tau) - G_0(r, t_h - \tau)] H[\tau - t_k] d\tau d\gamma_\xi d\gamma_x \\
&+ \int_{\widehat{\Gamma}} \widetilde{\varphi}_j(x) \int_{\widehat{\Gamma}} \widetilde{\varphi}_i(\xi) \int_0^{t_h} G_0(r, t_h - \tau) H[\tau - t_k] d\tau d\gamma_\xi d\gamma_x\,,
\end{aligned} \tag{15}$$

in such a way that the problematic issues described above are confined in the second term, whose time integral can be evaluated analytically. The time integral of the first term is instead performed numerically. At last, the outer integrals over $\widehat{\Gamma}$ are numerically treated by suitable quadrature schemes as in [2].

4 Numerical Results

In the following, we will present some numerical results obtained for 2D exterior problems starting from the proposed energetic weak formulation. We consider the problem (1)-(4) with $\Gamma = \{(x,0)\,|\,x \in [-1,1]\}$ and the Dirichlet boundary datum

$$g(x,t) = -H[t-kx]f(t-kx)x, \quad \text{where} \quad f(z) = \begin{cases} \sin^2\left(\dfrac{\omega z}{2}\right), & \text{if} \quad 0 \le z \le \dfrac{\pi}{\omega} \\ 1\,, & \text{if} \quad z \ge \dfrac{\pi}{\omega} \end{cases}$$

with $\omega = 8\pi$, $k = \cos\theta$ and $\theta = \pi/2$ as suggested in [2]. Further the velocity is fixed as $c = 1$. For the energetic BEM, constant shape and test functions in space and time are here always adopted.
At first, we chose a uniform decomposition of Γ in 80 straight elements ($\Delta x = 0.025$) and we set $\Delta t = 0.025$. We show in Figure 1, the time history of the solutions $\phi(x,t)$ on the straight element e_{20} and on the time interval $[0,4]$, for $P = 0$ varying $D = 0, 1, 10, 20$ on the left and for $D = 0$ varying $P = 0, 10, 20$ on the right. Note the effects of increasing viscous and material damping which substantially change the aspect of the solution related to the classical wave equation, visible in the graphs for trivial parameters.

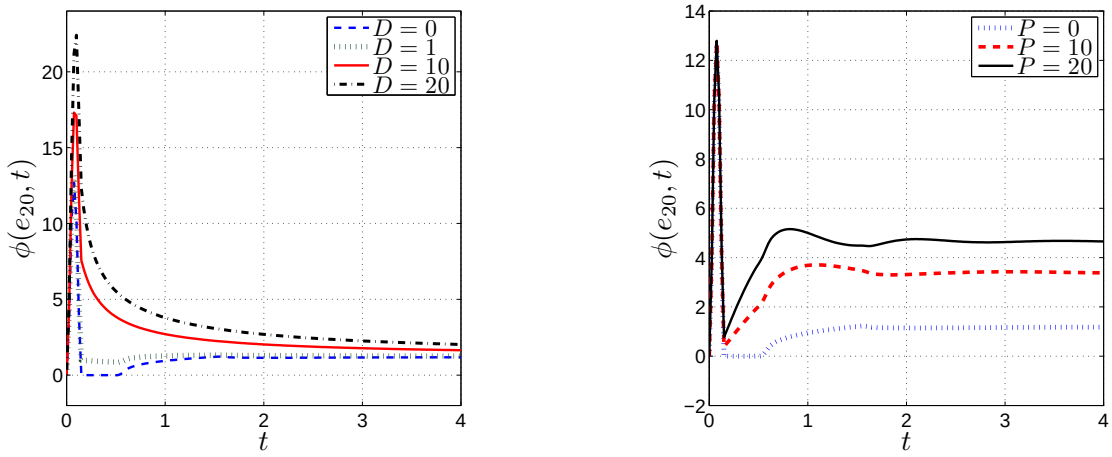


Figure 1: Time history of $\phi(x,t)$ on straight element e_{20}.

Then, we chose a uniform decomposition of Γ in 40 subintervals ($\Delta x = 0.05$) and enlarge the observation time interval, fixing $T = 10$ and $\Delta t = 0.1$. Since in this benchmark the Dirichlet datum becomes independent of time, for $P = 0$ we expect that the BIE transient solution $\phi(x,t)$ on Γ tends to the stationary one $\phi_\infty(x)$ (Fig. 2, left), i.e. the solution of the BIE related to the following Dirichlet problem for the Laplace equation:

$$\begin{aligned} &\Delta u_\infty(x) = 0 \qquad x \in \mathrm{R}^2 \setminus \Gamma \\ &u_\infty(x) = -x \qquad x \in \Gamma \\ &u_\infty(x) = O(1) \quad \|x\| \to \infty \end{aligned}$$

Looking at the graphs of the time history of $\|\phi(\cdot,t) - \phi_\infty(\cdot)\|_{L^1(\Gamma)}$, shown in Figure 2 (right),

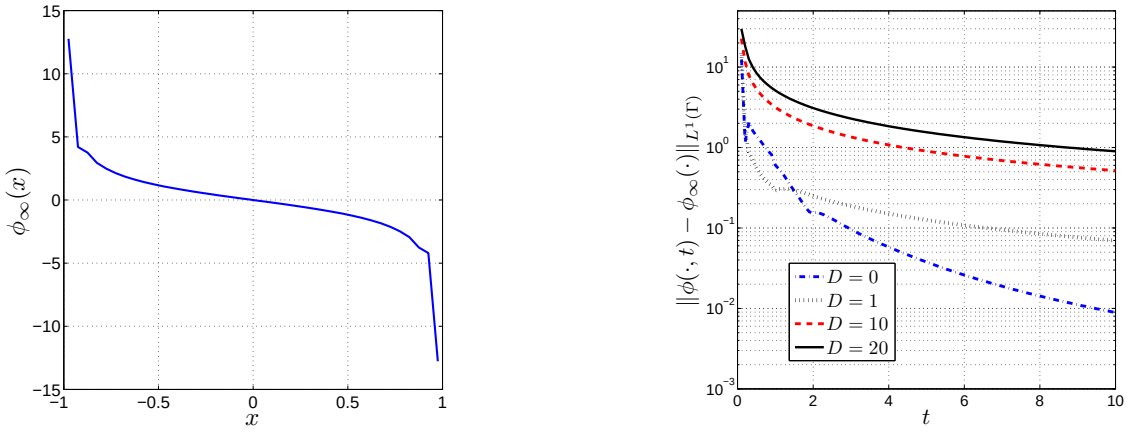


Figure 2: Stationary solution $\phi_\infty(x)$ on Γ (left); convergence of damped BIE solution towards the solution of Laplace problem, when $P = 0$ (right)

we observe the expected convergence, that becomes slower for increasing values of parameter D.
Analogously, for $D = 0$ and $P \neq 0$ we expect that the transient solution $\phi(x,t)$ on Γ tends to the stationary one $\overline{\phi}_\infty(x)$ (Fig. 3), i.e. the solutions of the BIE related to the following Dirichlet problem for the Helmholtz equation:

$$\begin{aligned} &\Delta v_\infty(x) + k^2 v_\infty(x) = 0 \quad && x \in \mathrm{R}^2 \setminus \Gamma \\ &v_\infty(x) = -x && x \in \Gamma \\ &v_\infty(x) = O(\|x\|^{-1}) && \|x\| \to \infty \end{aligned}$$

with wave number $k = \sqrt{-P}/c^2$.

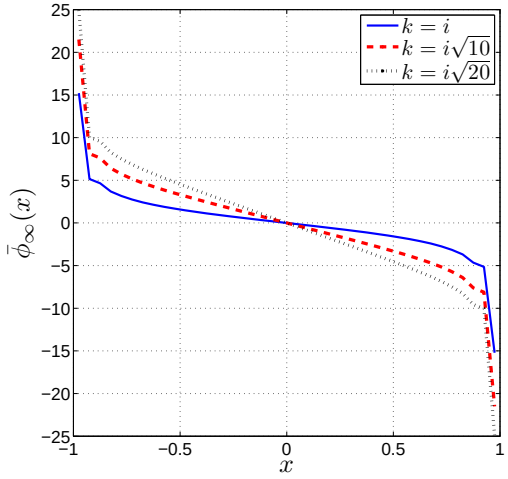


Figure 3: Stationary solutions $\overline{\phi}_\infty(x)$ on Γ.

Looking at the graphs of the time history of $\|\phi(\cdot,t) - \bar{\phi}_\infty(\cdot)\|_{L^1(\Gamma)}$, in Figure 4, we observe

the expected convergence, that becomes more oscillating for increasing values of parameter P.

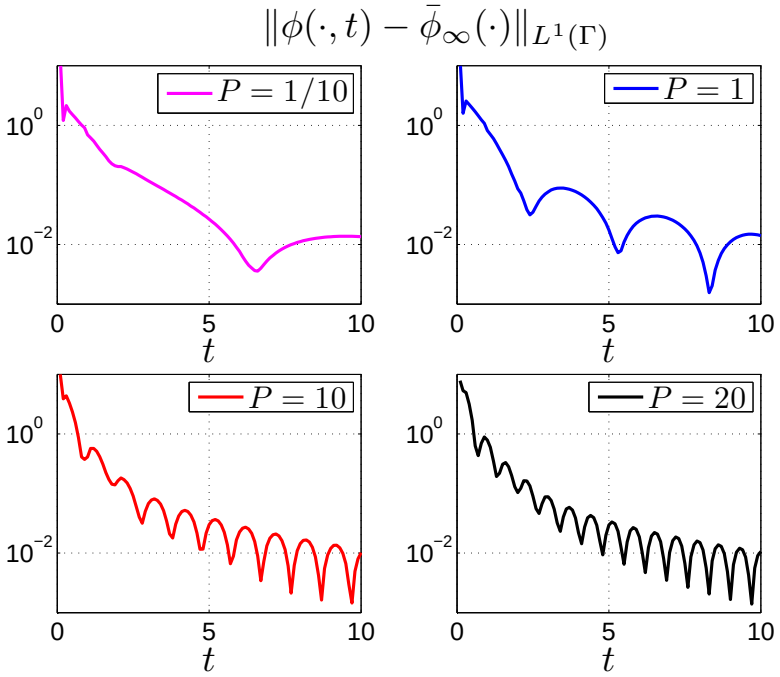


Figure 4: Convergence of damped BIE solution towards the solution of Helmholtz problem, when $D=0$ and $P\neq 0$.

REFERENCES

[1] A. Aimi, M. Diligenti, A new space-time energetic formulation for wave propagation analysis in layered media by BEMs, *Int. J. Numer. Meth. Engng.* **75**, 1102–1132, 2008.

[2] A. Aimi, M. Diligenti, C. Guardasoni, I. Mazzieri, S. Panizzi, An energy approach to space-time Galerkin BEM for wave propagation problems, *Int. J. Numer. Meth. Engng*, **80**, 1196–1240, 2009.

[3] A. Aimi, M. Diligenti, S. Panizzi, Energetic Galerkin BEM for wave propagation Neumann exterior problems, *CMES* **58**, 185-219, 2010.

[4] A. Aimi, M. Diligenti, A. Frangi, C. Guardasoni, A stable 3D energetic Galerkin BEM approach for wave propagation interior problems, *Engineering Analysis with Boundary Elements*, **36** 4, 1756–1765, 2012.

[5] A. Aimi, S. Gazzola, C. Guardasoni, Energetic boundary element method analysis of wave propagation in 2D multilayered media, *Mathematical Methods in the Applied Sciences*, **35** 110, 1140–1160, 2012.

[6] A. Aimi and S. Panizzi, BEM-FEM Coupling for the 1D Klein-Gordon Equation, *Numer. Methods Partial Differential Eq.*, **30**, 2042–2082, (2014).

[7] A. Aimi, M. Diligenti and C. Guardasoni, Energetic BEM-FEM coupling for the numerical solution of the damped wave equation, *International Journal for Numerical Methods in Engineering*, submitted, (2015).

[8] H. Antes, G. Beer, W. Moser, Soil-structure interaction and wave propagation problems in 2D by a Duhamel integral based approach and the convolution quadrature method, *Comput. Mech*, **36** 6, 431–443, 2005.

[9] A. Bamberger, T. Ha Duong, Formulation variationelle espace-temps pour le calcul par potential retardé de la diffraction d'une onde acoustique (I), *Mathematical Methods in the Applied Sciences*, **8**, 405–435, 1986.

[10] A. Bamberger, T. Ha Duong, Formulation variationelle pour le calcul de la diffraction d'une onde acoustique par une surface rigide, *Mathematical Methods in the Applied Sciences*, **8**, 598–608, 1986.

[11] P.K. Banerjee, P.K. Butterfield, Boundary Element Methods in Engineering, McGraw-Hill U.K. Ltd. 1981.

[12] B. Birgisson, E. Siebrits and A.P. Peirce, Elastodynamic direct boundary element methods with enhanced numerical stability properties, *Int. J. Numer. Meth. Engrg.*, **46**, 871–888, 1999.

[13] M. Costabel, *Time-dependent problems with the boundary integral equation method*, Encyclopedia of Computational Mechanics, John Wiley and Sons, E. Stein et al. eds., 1–28, 2004.

[14] D.A.S. Curran, M. Cross, B.A. Lewis, Solution of Parabolic Differential Equations by the Boundary Element Method Using Discretization in Time, *Applied Mathematical Modelling*, **4**, 398–400, 1980.

[15] P.J. Davies and D.B. Duncan, Stability and Convergence of Collocation Schemes for Retarded Potential Integral Equations, *SIAM J. Numer. Anal.*, **42**, 1167–1188, 2004.

[16] A. Frangi, G. Novati, On the numerical stability of time-domain elastodynamic analyses by BEM, *Computer Methods in Applied Mechanics and Engineering*, **173**, 403–417, 1999.

[17] T. Ha Duong, On retarded potential boundary integral equations and their discretization. In: P. Davis, D. Duncan, P. Martin, B. Rynnm editors. Topics in Computational wave propagation. Direct and inverse problems. Springer-Verlag, 301–336, 2003.

[18] T. Ha Duong, B. Ludwig and I. Terrasse, A Galerkin BEM for transient acoustic scattering by an absorbing obstacle, *Int. J. Numer. Methods Engnrg.*, **57** (13), 1845–1882, 2003.

[19] F. Hartmann, *Introduction to Boundary Element Theory Method in Engineering*, McGraw-Hill, U.K. Ltd. 1981.

[20] C. Lubich, Convolution quadrature and discretezed operatorial calculus. I and II. *Numer. Math.* **,** 129–145, 413–425, 1988.

[21] C. Lubich, Convolution quadrature revised, *BIT*, **44**(3), 503–514, 2004.

[22] M. Schanz and H. Antes, Application of Operational Quadrature Methods in Time Domain Boundary Element Methods, Mechanica, **32**, (1997), 179–186.

[23] M. Schanz, Application of 3D Boundary Element Formulation to Wave Propagation in Poroelastic Solids, *Eng. Anal. Boundary Elem.*, **25**, 363–376, 2001.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MODELLING THE PROCESS OF SEQUENTIAL EXCAVATION WITH THE BOUNDARY ELEMENT METHOD

C. Duenser[1] and G. Beer[1,2]

[1]Institute for Structural Analysis, Graz University of Technology
Lessingstrasse 25, 8010 Graz, Austria
e-mail: duenser@tugraz.at

[2] Centre for Geotechnical and Materials Modeling University of Newcastle
Callaghan, NSW 2308, Australia
e-mail: beer@tugraz.at

Keywords: Sequential Excavation, BEM, BETI, DDM.

Abstract. *The Boundary Element Method (BEM) is ideally suited for the simulation of underground constructions like tunnels or caverns. Such structures are modelled with the BEM inside an infinite or semi-infinite domain. As the radiation condition is fulfilled by the BEM no truncation of the domain is necessary. Only the surface of the structure (e.g. tunnel) has to be discretised by boundary elements (BE). An accurate simulation of the tunnelling process has to consider the sequential excavation where parts of the rock mass are excavated at different time and location. This special constructional condition has a direct influence onto the simulation model. In this work different methods are presented which consider the sequential excavation.*
The first method is the discretisation of the problem by multiple BE regions (MRBEM). Each region, which will be excavated during the excavation process, is discretised by a separate finite BE region. These regions are embedded inside an infinite region which represent the infinite extend of the domain. Thus, a system of BE regions arise which have to be coupled at their common interfaces. Two coupling strategies, the Boundary Element Tearing and Interconnecting Method (BETI) and the method of Interface Coupling (IC) will be presented to solve the sequential tunnel excavation.
The second method uses only a single BE region (SRBEM) for every step of excavation. For each load step the geometry/mesh has to be updated. Thus, the mesh of the previous load step will be extended by the surface of the new excavation volume of the current load step. Beside the geometry update an essential part of this method is an accurate evaluation of the excavation loading. The excavation loads for each excavation step are tractions applied at the part of the boundary surface just generated by the geometry update. These tractions depend on all previous load steps and will be evaluated by a calculation of internal results in the interior of the single region. The internal results can be either stresses or displacements.
In this work the modelling strategies of the MRBEM and SRBEM approach will be presented. On a realistic tunnel example the accuracy of the results for the mentioned methods will be shown as well as the performance of the calculations.

1 INTRODUCTION

Among other characteristics the construction of tunnels within the New Austrian Tunnelling Method is characterised by the sequential excavation of the tunnel. A typical sequential excavation is shown in the Fig. 1 where top heading and bench excavation is performed. Thus the cross section of the tunnel is divided into two parts, top heading and bench. Within the sequential excavation those volume parts are excavated at different time and location. As shown in the figure, in the longitudinal direction of the tunnel the top heading excavation is more advanced as the bench excavation at the same time.

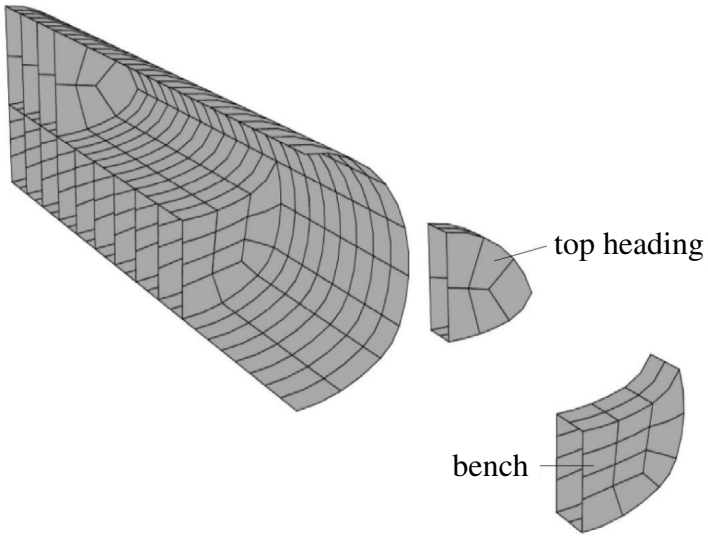


Figure 1: Sequential tunnel excavation

The modelling of the tunnel excavation is done either in an infinite or semi-infinite domain, depending on the depth of the tunnel from the ground surface. Using the Boundary Element Method (BEM) for the simulation of such problems the radiation condition is fulfilled implicitly within the solution, thus no truncation of the mesh and special boundary conditions are necessary. Using the BEM only the surface of the problem has to be discretised. Thus the mesh generation is drastically reduced for such simulation problems. This will be shown by a 3D example in section 4.

2 BOUNDARY INTEGRAL EQUATION

The basis of the current work is the displacement boundary integral equation (see [1]) which is:

$$\mathbf{C}(\mathbf{y})\mathbf{u}(\mathbf{y}) + \cancel{\int_{\Gamma}} \mathbf{T}(\mathbf{y},\mathbf{x})\mathbf{u}(\mathbf{x})\,\mathrm{d}\Gamma = \int_{\Gamma} \mathbf{U}(\mathbf{y},\mathbf{x})\mathbf{t}(\mathbf{x})\,\mathrm{d}\Gamma \tag{1}$$

where $\mathbf{U}(\mathbf{y},\mathbf{x})$ and $\mathbf{T}(\mathbf{y},\mathbf{x})$ are the fundamental solutions for displacements and tractions and $\mathbf{u}(\mathbf{x})$ and $\mathbf{t}(\mathbf{x})$ are the boundary displacements and tractions, respectively. $\mathbf{C}(\mathbf{y})$ is an integral free term depending on the geometrical conditions at the boundary point $\mathbf{y}$. The boundary integral equation (1) is valid for a single region whose boundary is discretised with boundary elements. Due to the discretisation the boundary Γ is divided into elements E and nodes N.

Thus, the integral equation (1) is transformed to:

$$\mathbf{C}\mathbf{u}_i + \sum_{e=1}^{E}\sum_{n=1}^{N} \mathbf{\Delta T}^e_{ni}\mathbf{u}^e_n = \sum_{e=1}^{E}\sum_{n=1}^{N} \mathbf{\Delta U}^e_{ni}\mathbf{t}^e_n \tag{2}$$

$\mathbf{\Delta T}^e_{ni}$ and $\mathbf{\Delta U}^e_{ni}$ are integrated kernel matrices with respect to the collocation node i and element n. Eq. 2 is evaluated for all collocation points i and the coefficient matrices $\mathbf{\Delta T}^e_{ni}$ and $\mathbf{\Delta U}^e_{ni}$ are assembled into matrices $[\Delta T]$ and $[\Delta U]$, whereas the following equation arise:

$$[\Delta T]\{u\} = [\Delta U]\{t\} \tag{3}$$

Eq. 3 allows the solution of a single boundary element region. At the nodes of the boundary either displacements or tractions are known. The values for the unknown boundary conditions (BC's) are solved by rearranging Eq. 3. The corresponding columns of matrices $[\Delta T]$ and $[\Delta U]$ related to the unknown BC' are shifted to the left side and the columns of the matrices $[\Delta T]$ and $[\Delta U]$ related to the known BC's are multiplied with the values of the known BC's and form the right hand side vector $\{f\}$ of the following equation:

$$[A]\{x\} = \{f\} \tag{4}$$

In case of a mixed boundary value problem the content of the solution vector $\{x\}$ are either displacements or tractions, matrix $[A]$ is filled up either with columns of matrix $[\Delta T]$ or $[\Delta U]$.

3 MODELLING THE SEQUENTIAL EXCAVATION PROBLEM

In the particular case of a sequential excavation rock volumes are excavated gradually. Thus the simulation is performed step wise. In each step in general one or more rock volumes may be excavated. The simulation model has to meet these requirements to evaluate the deformation and stress state [2] of the tunnel progress. With the BEM this process can be simulated by applying a Multiple Region BEM (MRBEM) or a Single Region BEM (SRBEM). Using a MRBEM approach each region removed during the excavation process is discretised by a BEM region. The regions are coupled at their common interfaces and form the MREBM system. Using the SRBEM only a single region is necessary to solve the tunnelling problem for a certain step of excavation.

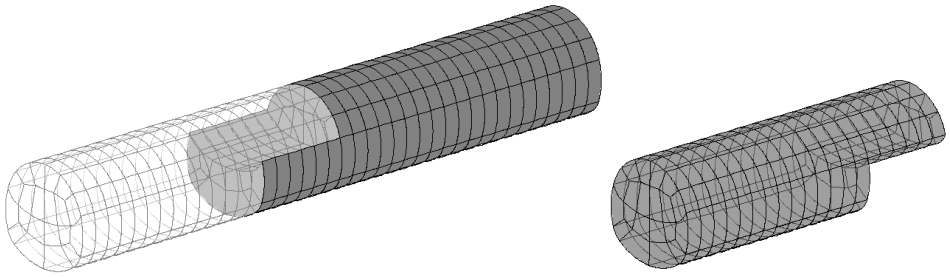


Figure 2: MRBEM model (left) and SRBEM model (right)

In Fig. 2 the meshes are shown which are necessary to solve the tunnelling problem for a certain excavation state either with the MRBEM (left mesh) or with the SRBEM method (right mesh) [3]. The mesh of the MRBEM model consists of several finite regions which are

embedded in an infinite region. The finite regions are deactivated from the calculation model during the simulation process according to a predetermined sequence.

With the SRBEM model a single BEM region is discretised which represents the actual excavation state of the tunnel, which is the surface of the volume already excavated. The mesh size at the beginning of the analysis is very small and is increasing together with the progress of the tunnel excavation.

3.1 Multiple region BEM

Currently two coupling methods are used to simulate the sequential excavation.These two methods are explained next.

3.1.1 Interface coupling method (IC)

For the interface coupling method [1] stiffness matrices are evaluated which are related only to the interface of a region.

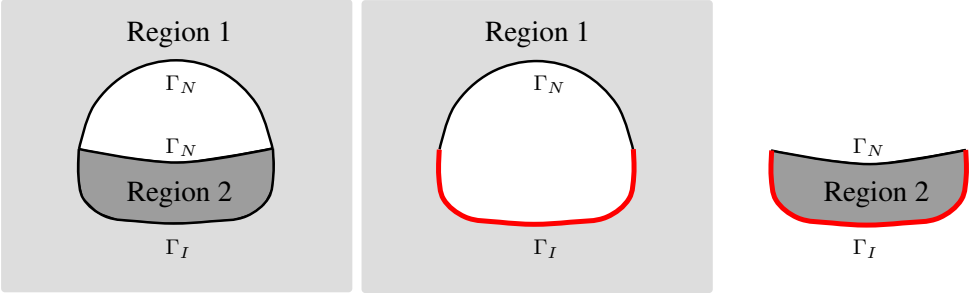


Figure 3: Interface coupling (IC)

Starting with the assembled equation system (3), which is

$$[\Delta T]\{u\} = [\Delta U]\{t\} \tag{5}$$

and assuming that no Dirichlet boundary conditions are present (which in tunnelling is usually the case) the boundary can be divided into a Neumann and Interface part. This is shown in Fig. 3 where 2 regions are coupled at the common interface Γ_I. Due to the different boundary condition types Eq. 5 can be separated to:

$$\begin{bmatrix}[U_I]\ [U_N]\end{bmatrix} \begin{Bmatrix}\{t_I\}\\ \{t_N\}\end{Bmatrix} = \begin{bmatrix}[T_I]\ [T_N]\end{bmatrix} \begin{Bmatrix}\{u_I\}\\ \{u_N\}\end{Bmatrix} \tag{6}$$

The traction t_N are the known boundary condition values and will be shifted to the right side together with the kernel matrix $[U_N]$. u_N is the unknown vector of displacements of the Neumann boundary and this will be shifted to the left together with the kernel matrix $[T_N]$.

$$\begin{bmatrix}[U_I]\ -[T_N]\end{bmatrix} \begin{Bmatrix}\{t_I\}\\ \{u_N\}\end{Bmatrix} = \begin{bmatrix}[T_I]\ -[U_N]\end{bmatrix} \begin{Bmatrix}\{u_I\}\\ \{t_N\}\end{Bmatrix} \tag{7}$$

The matrix on the left side of Eq. 7 is renamed by $[A]$ and on the right hand side the vector of interface displacements $\{u_I\}$ are separated from the traction vector $\{t_N\}$ belonging to the

Neumann boundary:

$$\begin{bmatrix} \quad A \quad \end{bmatrix} \begin{Bmatrix} \{t_I\} \\ \{u_N\} \end{Bmatrix} = [T_I]\,\{u_I\} + \begin{bmatrix} -[U_N] \end{bmatrix} \{\ \{t_N\}\ \} \tag{8}$$

Solving this equation for $\{t_I\}$ and $\{u_N\}$ will result to the following:

$$\begin{Bmatrix} \{t_I\} \\ \{u_N\} \end{Bmatrix} = [A]^{-1}[T_I]\,\{u_I\} + [A]^{-1}\big[-[U_N]\big] \{\ \{t_N\}\ \} \tag{9}$$

The matrix product $[A]^{-1}[T_I]$ will result to the stiffness matrix $[K]^*$ and a flexibility matrix $[D]$, and the second part of the right hand side of Eq. 9, as vector $\{t_N\}$ is known, will give a known vector shown next:

$$\begin{Bmatrix} \{t_I\} \\ \{u_N\} \end{Bmatrix} = \begin{bmatrix} [K]^* \\ [D] \end{bmatrix} \{u_I\} + \begin{Bmatrix} \{t_{IN}\} \\ \{u_{NN}\} \end{Bmatrix} \tag{10}$$

The first equation of 10 represents the interface tractions $\{t_I\}$. Multiplying it with the mass matrix $[M]_r$ of region r, the tractions are replaced by equivalent nodal point forces. A coupling to a Finite Element region would be possible know. This equation is shown next:

$$\{f_I\}_r = [K]_r\{u_I\}_r + \{f_{IN}\}_r \tag{11}$$

The final forces $\{f_I\}_r$ at the interface of region r are the forces due to the interface displacements $\{u_I\}_r$ plus the forces at the interface due to the loading (given tractions $\{t_N\}_r$. The interface forces $\{f_I\}_r$ as well as the interface displacements $\{u_I\}_r$ are unknown at the present state. These unknowns are evaluated together with all other regions. Thus, $[K]_r$ and $\{f_{IN}\}_r$ are evaluated for every region r of the multiple region system and the final system of equation can be assembled under the following conditions:

– Equilibrium of forces at the interface:

$$\{f_I\}_1 + \{f_I\}_2 + ... + \{f_I\}_r = 0 \tag{12}$$

Eq.12 states that the forces at the interface of all adjacent regions are in equilibrium.

– Compatibility of displacements at the interface:

$$\{u_I\}_m = \{u_I\}_n \tag{13}$$

Equation (13) states that the displacements at the interface of adjoining regions n and m are equal.

Considering these conditions, Eq.11 of every region is assembled into a global equation system which is shown as following:

$$\{f_I\}_r = [K]^{sys}\{u_I\} + \{f_N\} = 0 \tag{14}$$

where $[K]^{sys}$ is the assembled stiffness matrix related to all coupling interfaces of the system. $\{f_{IN}\}$ is the right hand side vector related to the loading of the system and $\{u_I\}$ is the vector of interface displacements. From this equation the interface displacements are solved. Once $\{u_I\}$ is known all remaining unknowns (tractions at the interface $\{t_I\}$ and displacements at the Neumann boundary $\{u_N\}$) can be evaluated using Eq.10.

3.1.2 Boundary element tearing and interconnecting method (BETI)

The Boundary Element Tearing and Interconnecting Method (BETI) is a domain decomposition method based on the Symmetric Galerkin BEM [4] similar to the Finite Element Tearing and Interconnecting Method (FETI) for the FEM introduced by [5]. In this work the collocation BEM is used to set up the equation system similar to the original BETI method. Applying this method for each region a stiffness matrix based on the DOFs of the whole region surface is calculated which is in contrast to the method IC applied in section 3.1.1, where the stiffness matrix is based on the coupled DOFs only. From Eq.3 the boundary tractions of a region are calculated as following:

$$[\Delta U]^{-1}[\Delta T]\{u\} = \{t\} \tag{15}$$

As in the IC method tractions are converted to work equivalent nodal point forces by multiplying Eq.15 with the mass matrix $[M]$:

$$[M][\Delta U]^{-1}[\Delta T]\{u\} = [M]\{t\} = \{f\} \tag{16}$$

where the stiffness matrix $[K]$ is:

$$[K] = [M][\Delta U]^{-1}[\Delta T] \tag{17}$$

Inserting this into Equation (16) will result in the well known relation between displacements and forces:

$$[K]\{u\} = \{f\} \tag{18}$$

As in the IC method two conditions have to be satisfied to formulate a coupled system consisting off r boundary element regions:

- Equilibrium
- Compatibility

The equilibrium state of a region can be described by using Equation (18):

$$[K]\{u\} = \{f_N\} + [B]^T\{\lambda\} \tag{19}$$

whereas the force vector on the right hand side of Equation (18) is split into:

$$\{f\} = \{f_N\} + [B]^T\{\lambda\} \tag{20}$$

and inserted into Equation (19). $[K]\{u\}$ are the forces at the boundary of the region due to deformation, $\{f_N\}$ is the force vector of the given loading (Neumann boundary conditions) and $[B]^T\{\lambda\}$ are the coupling forces (Lagrange multipliers) to the neighbouring regions.

The compatibility of a system of r regions can be written in following form:

$$[B]_1\{u\}_1 + [B]_2\{u\}_2 + \cdots + [B]_r\{u\}_r = \{b\} \tag{21}$$

Equation (21) either guaranties that the displacements at the interface of adjacent regions are equal or that the displacements at the Dirichlet boundary are equal to the applied Dirichlet boundary conditions which are entries of vector $\{b\}$.

The final system of equation of a coupled system of r boundary element regions is shown as following:

$$\begin{bmatrix} [K]_1 & & 0 & & -[B]_1^T \\ & [K]_2 & & & -[B]_2^T \\ 0 & & \ddots & & \vdots \\ & & & [K]_r & -[B]_r^T \\ [B]_1 & [B]_2 & \cdots & [B]_r & 0 \end{bmatrix} \cdot \begin{Bmatrix} \{u\}_1 \\ \{u\}_2 \\ \vdots \\ \{u\}_r \\ \{\lambda\} \end{Bmatrix} = \begin{Bmatrix} \{f_N\}_1 \\ \{f_N\}_2 \\ \vdots \\ \{f_N\}_r \\ \{b\} \end{Bmatrix} \tag{22}$$

Equations 1 to r of Eq.22 are representing the equilibrium of each region and the last equation of Eq.22 guaranties compatibility of displacements at every node at the interface of adjacent regions and at the nodes corresponding to the Dirichlet boundary.

In the implementation of the BETI method Eq.22 is not assembled to an equation system. The equation system Eq.22 is condensed to the solution of the coupling forces λ (Lagrange multipliers). This is done by inserting equations 1 to r into the last equation of Eq.22. From this equation λ is solved either directly or iteratively with a BiCGSTAB iterative solver. As the stiffness matrix $[K]_r$ of a finite region (floating region) is singular [6] special treatment of rigid body motions have to be considered. The whole solution formulation is shown in detail by [4].

3.1.3 Comparison of coupling methods

The main advantage of the BETI method is that the stiffness matrix of each region has to be calculated only once and in the case of a sequential tunnel excavation these matrices can be used for each load step of excavation. Using the BETI method for this application type the stiffness matrices are independent on the changing boundary conditions. Changing boundary conditions due to sequential excavation are considered by the coupling matrix $[B]_r$ of Equation (22). The coupling matrices have to be computed repeatedly for each calculation step. As those matrices are sparsely populated they are implemented as sparse matrices. The effort to set up those matrices is small and it is insignificant compared to the overall computing time. The way how the equation system is formulated makes the treatment of operations related to a region independently from the other regions. Thus, the BETI method is ideally suited for parallelisation techniques.

The advantage of the IC method is that the size of the stiffness matrix is related to the number of DOFs at the interface of the coupled system of regions. In the case of a sequential excavation the coupling surfaces are reduced from one excavation step to the other. In each load step one or more regions are deactivated from the simulation model. Due to the deactivation the boundary condition of surfaces adjacent to the deactivated regions change from Interface condition to Neumann condition. Thus, the size of the assembled system stiffness matrix reduces from one load step to the other and the solution of the equation system gets faster. For regions for which a change of boundary conditions happens the stiffness matrix has to be calculated newly. Compared to the BETI method stiffness matrices do not remain constant throughout the entire analysis of such an excavation simulation.

3.2 Single region BEM

In Fig.4 the excavation using single regions [7] are shown on a simple example in 2D for two excavation steps. The simulation considers a top heading excavation in the first excavation

step and a bench excavation in the second step by using a single BEM region for each step. The region for the first step is the boundary of the top heading and the region for the second step is the boundary of the final tunnel surface (the outer boundary of top heading and bench together).

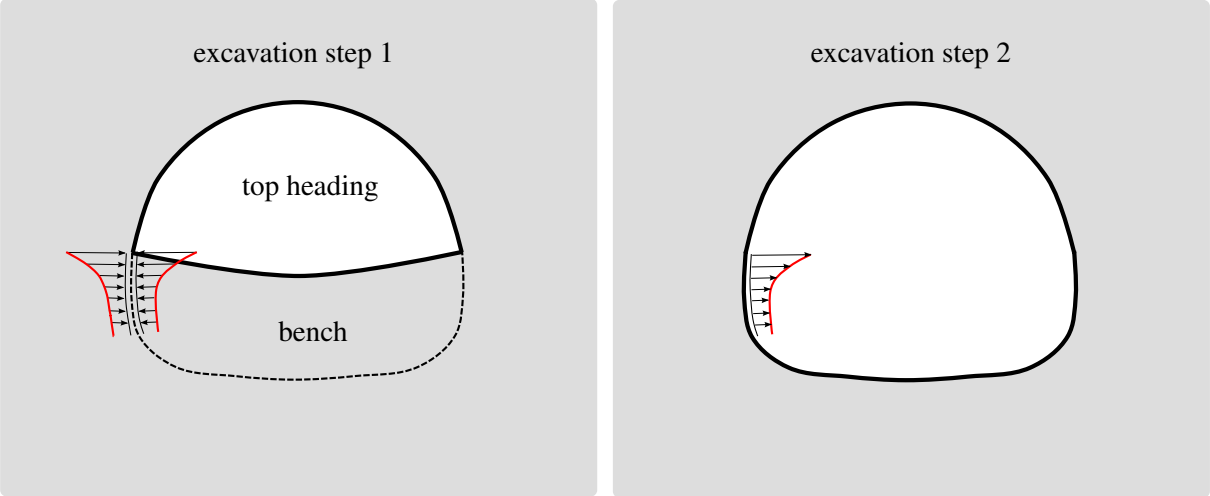


Figure 4: Single region BEM (SRBEM) excavation

The crucial task is the evaluation of the excavation loading (boundary tractions) for the subsequent excavation step. For the first step the excavation loading (boundary traction) is given from the known primary stress field at the whole surface of the top heading region. Applying Eq.3 with known tractions the displacements are calculated for the top heading boundary. In a post-processing algorithm the tractions along the boundary of the bench (dashed line in Fig.4) inside of region 1 (top heading region) are calculated. These tractions are applied at the region of the subsequent load step (shown in Fig.4 on the right) and again Eq.3 is used to solve this load step.
For the evaluation of the tractions inside a region two methods have been investigated. The simpler method is the direct evaluation of the stress at points inside the region using the stress boundary integral equation. As the stress at the corners of the bench (shown in Fig.4) are singular, an evaluation of the stress is difficult as well as the distribution of the stress towards that corner point. The results due to this method are not satisfactory, thus a second method was investigated where displacements are calculated at the inner points. With this method an intermediate calculation for the region which is excavated (only the bench region!) is necessary. The mentioned displacements will be the known Dirichlet boundary conditions and the tractions will be calculated by this intermediate single region calculation. In excavation step 2 this traction will be applied as loading. A detailed explanation about the key issue of evaluating excavation loads can be found in [8].

In comparison to the the coupling methods with this method the mesh of the model has to be adapted at each step of excavation due to the change of the tunnel surface. The simulation is starting with a very small mesh and in every subsequent excavation step the existing mesh has to be extended by the surface of the removed volume of excavation. At the end of the excavation simulation (all regions are excavated) the mesh size is the same as for the coupling methods. This will be shown in the 3D example shown next.

4 EXAMPLE - CROSSING PASSAGE - NEW YORK UNDERGROUND

In the following an example of a sequential tunnel excavation is shown. More precisely it is the crossing passage of two tunnels of the New York Underground. In the planning phase 3D-CAD planning documents were available. In order to create the BEM model these data was used to mesh the geometry. The initial CAD model of the construction design was too detailed in order to use it for the numerical discretisation and therefore the CAD geometry description had to be simplified. After some adaptation and simplification of the CAD model it was imported to the pre-processor CUBIT 2014 [9] and the mesh was automatically generated. The final mesh is shown in Fig.5 and was modelled with linear (4 nodes) quadrilateral boundary elements (BE). At all four ends of the simulation model infinite BE's [10] are situated to consider correct boundary conditions in the longitudinal direction of the tunnels. The mesh consists of approximately 20 000 nodes and 7 800 elements.

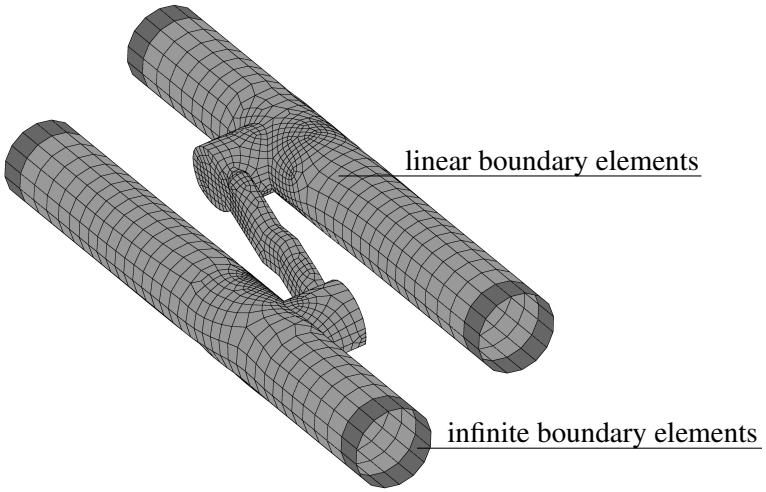


Figure 5: Mesh for the BEM analysis

In Fig.6 the sequence of the sequential excavation is shown. In sum 12 load cases (indicated by LC#) are considered. In the first 6 load cases the main tunnels are excavated, from LC7 to LC12 the crossing passage is excavated. The meshes for each LC shown in Fig.6 correspond to the single regions used in the SRBEM simulation method. The excavation model starts with a small mesh which is extended from LC to LC.

For both coupling methods (IC and BETI) the mesh of LC12 (shown in Fig.6) would be the infinite region, which is filled with finite regions, the region volumes which are excavated from LC to LC. Thus, the simulation for the coupling methods starts with a large model (with all coupled regions) which is reduced by finite regions during the excavation process.

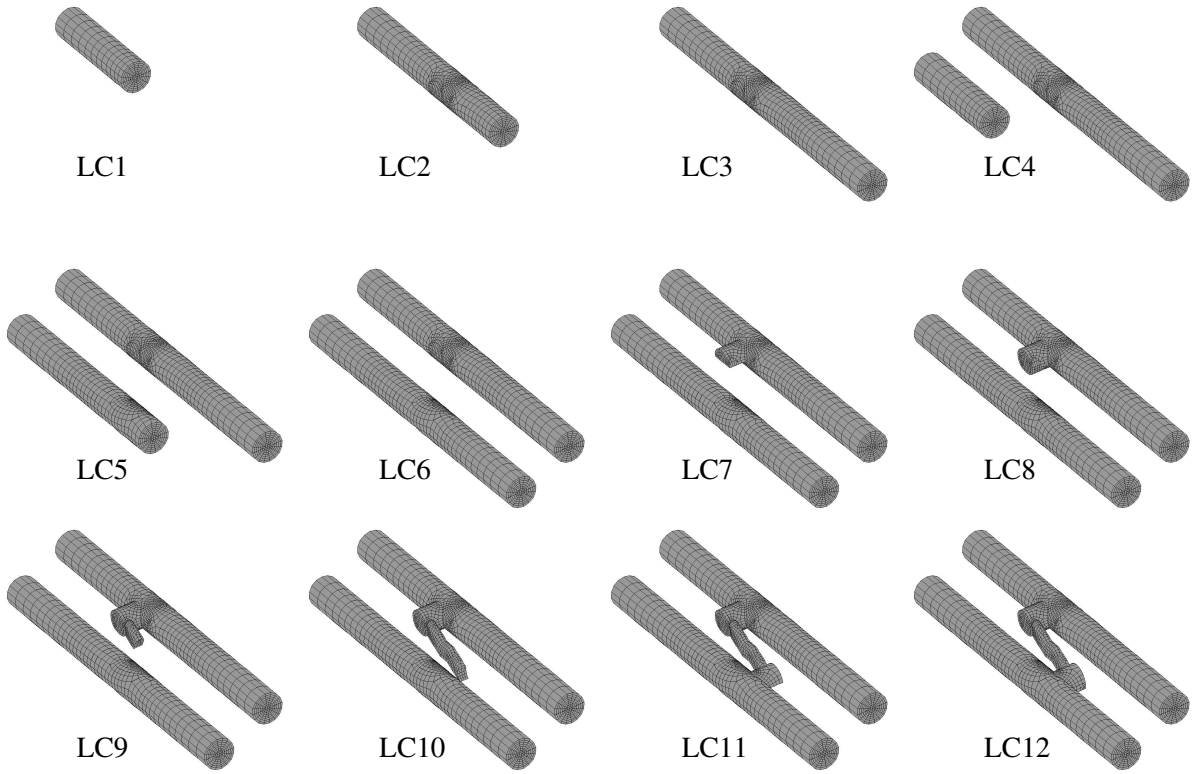


Figure 6: Excavation sequence LC1 to LC12

In Fig.7 contour lines of displacements in z-direction for the fully excavated tunnels and crossing passage are shown. The maximum settlement is 8.1 cm at the joining openings of the crossing passage and the upper main tunnel. The maximum heave is about 8.3 cm at the joining tunnels of the crossing passage and the bottom main tunnel.

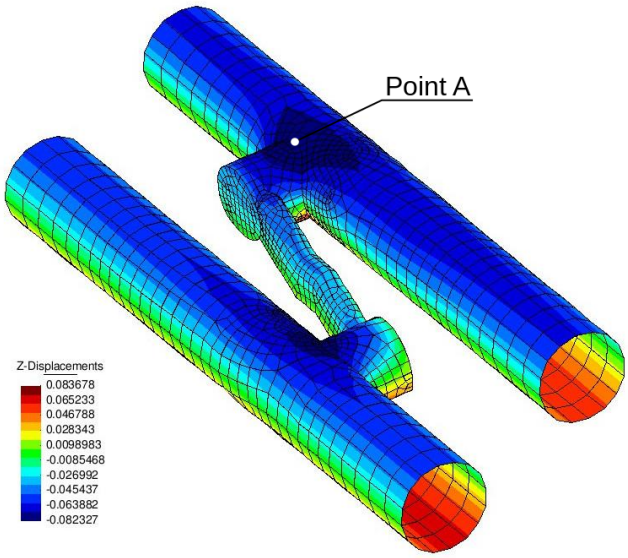


Figure 7: Contour lines of z-displacements for LC12

In the following diagram of Fig.8 vertical displacements for node A (location shown in Fig.7) of all load steps are displayed. The calculations were done for the three methods of excavation explained in section 3:

- Single region BEM method (SRBEM)
- Multiple region BEM method / interface coupling (MRBEM IC)
- Multiple region BEM method / BETI (MRBEM BETI)

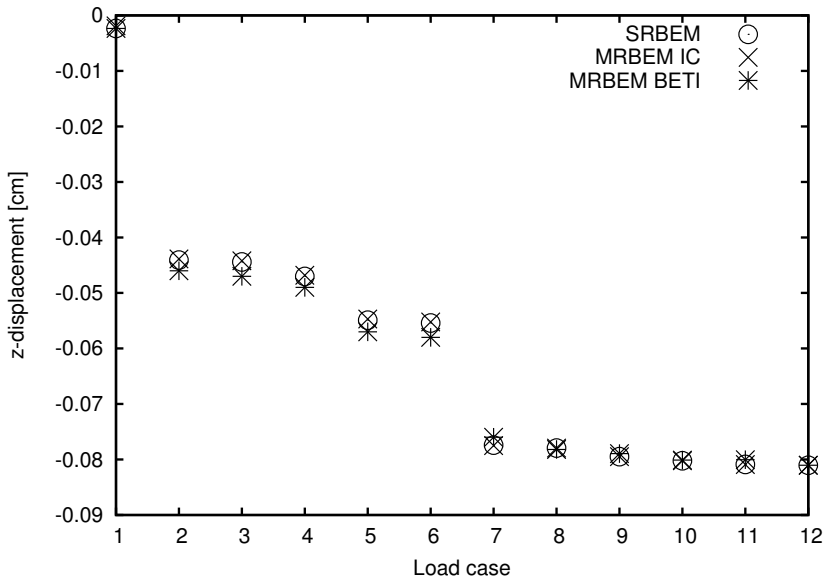


Figure 8: z-displacements of point A for LC1 to LC12

The displacements are displayed for each load case and each simulation method. It is shown that for the SRBEM and for the MRBEM IC very similar results are obtained. The results for the MRBEM BETI are slightly different from the two other methods.

Modelling method		calculation time [h]		
	1 CPU	8 CPU's	16 CPU's	32 CPU's
SRBEM	3,51	1,06	0,76	0,66
MRBEM – IC	12,39	2,50	1,67	1,40
MRBEM – BETI	5,41	0,87	0,55	0,58

Table 1: Computation times for different number of CPU's

The calculation times for the three simulation methods are shown in Table1 depending on the number of CPU's used. Without any parallelisation (calculation with 1 CPU) the calculation time for the MRBEM IC is the largest of the three methods. For this method the stiffness matrix of BEM regions with changing boundary conditions has to be calculated again from one load step to the other. This is the main reason for the lack of efficiency of this method (IC). For the present example the calculation time for the SRBEM is similar to the one of the MRBEM

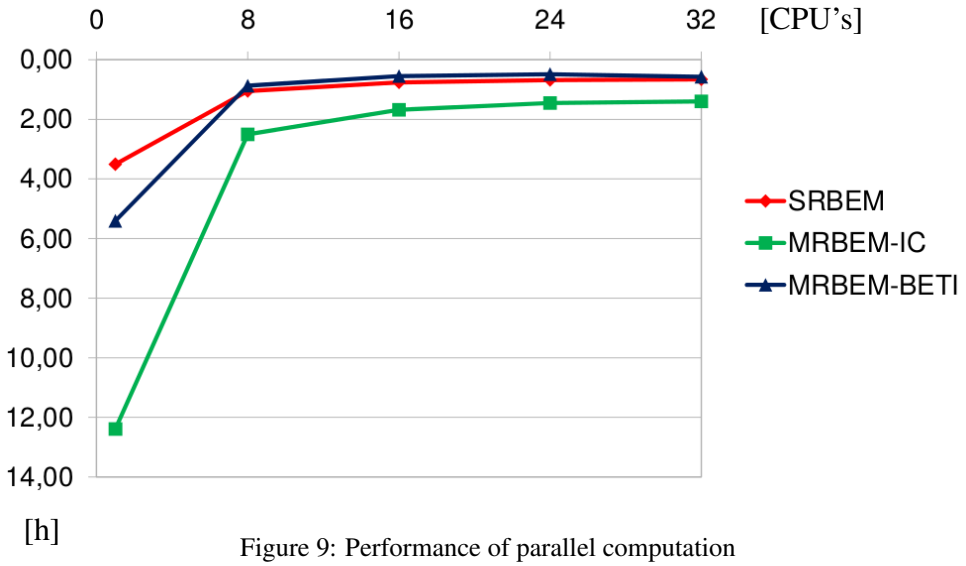


Figure 9: Performance of parallel computation

BETI. At the first excavation step the SRBEM method always starts with a very small mesh (small equations system), the size of the equation system increases from step to step as the surface of the excavation volume grows. For the MRBEM BETI method the stiffness matrices of all regions are calculated only once at the beginning of the analysis. They dont change any more during the whole analysis. This is the main advantage of the MRBEM BETI against the MRBEM IC method.

In Fig.9 the performance of the three simulation models is shown taking advantage of parallelisation. As shown the time difference for the overall computation of this excavation problem is decreasing as the number of CPU's increase.

At the current state of experience an objective statement cant be made whether SRBEM or MRBEM BETI is favourable in relation to computing performance. Maybe for an example of an increased number of load steps it can be expected that the MRBEM BETI method has an advantage over the SRBEM method.

5 CONCLUSIONS

In this work three simulation methods for the calculation of the sequential tunnel excavation using the BEM have been presented. Two coupling methods, the classical multiple region boundary element method with interface coupling (MRBEM IC) and the BETI coupling method (MRBEM BETI) are explained in brevity as well as the single region method (SRBEM).

On a practical example in 3D the results for the three methods are verified and it has been shown that the accuracy of the solutions of all methods is excellent and corresponds well to each other. The performance of parallelisation is demonstrated, whereas the methods SRBEM and MRBEM-BETI have distinct advantages in computing efficiency over the classical interface coupling. Causes and consequences of the three models for the simulation of the sequential tunnel excavation related to their implementation and the final performance have been worked out and are demonstrated by a realistic example in 3D.

REFERENCES

[1] G. Beer, I. Smith and C. Duenser, *The Boundary Element Method with Programming*. SpringerWienNewYork, 2008.

[2] C. Duenser, K. Thoeni, K. Riederer, B. Lindner, G. Beer, New Developments of the Boundary Element Method for Underground Constructions, *International Journal of Geomechanics* **12** (6) 665–675, 2012.

[3] C. Duenser, G. Beer, New Algorithms for the Simulation of the Sequential Tunnel Excavation with the Boundary Element Method, *Computational Methods in Tunnelling - EURO:TUN 2007*, 1–11, 2007.

[4] U. Langer and O. Steinbach, Boundary Element Tearing and Interconnecting Methods. *Computing*, **71**, 205–228, 2003.

[5] C. Farhat, F.-X. Roux, A method of finite element tearing and interconnecting and its parallel solution algorithm, *International Journal for Numerical Methods in Engineering*, **32** (6), 1205–1227, 1991.

[6] G. Of, O. Steinbach, The all-floating boundary element tearing and interconnecting method, *Journal of Numerical Mathematics* **17**, 277–298, 2009.

[7] C. Duenser, B. Lindner, G. Beer, Simulation of sequential Tunnel Excavation/Construction with the Boundary Element Method, in: *Computational Methods in Tunneling - EURO:TUN 2009, Aedificatio Publishers*, 147–154, 2009.

[8] C. Duenser and G. Beer, Simulation of sequential excavation with the Boundary Element Method. *Computers and Geotechnics*, **44**, 157–166, 2012.

[9] CUBIT, Preprocessor - Geometry and Mesh Generation Toolkit, Scandia National Laboratories, https://cubit.scandia.gov/

[10] W. Moser, C. Duenser, G. Beer, Mapped infinite elements for three-dimensional multi-region boundary element analysis, *International Journal for Numerical Methods in Engineering* **61** (3), 317–328, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

GREEN'S FUNCTIONS FOR THE EVALUATION OF ANCHOR LOSSES IN MEMS

A. Frangi[1] **and M. Cremonesi**[1]

[1]Department of Civil and Environmental Engineering
Politecnico di Milano
Piazza Leonardo da Vinci, 32, 20133, Milano, Italy
e-mail: {attilio.frangi,massimiliaano.cremonesi}@polimi.it

Keywords: MEMS, anchor losses, Green's functions.

Abstract. *The issue of dissipation has a peculiar importance in micro-electro-mechanical-structures (MEMS). Among the sources of damping that affect their performance, the most relevant are [1]: thermoelastic coupling, air damping, intrinsic material losses, electrical loading due to electrode routing, anchor losses. Moreover, recent experimental results indicate the presence of additional temperature dependent dissipation mechanisms which are not yet fully understood (see e.g. [2, 12]). In a resonating structure the quality factor Q is defined as:*

$$Q = 2\pi W/\Delta W \tag{1}$$

where ΔW and W are the energy lost per cycle and the maximum value of energy stored in the resonator, respectively. According to eq.(1), the magnitude of Q ultimately depends on the level of energy loss (or damping) in a resonator. The focus of the present contribution is set on anchor losses and the impact they have in the presence of axial loads. Anchor losses are due to the scattering of elastic waves from the resonator into the substrate. Since the latter is typically much larger than the resonator itself, it is assumed that all the elastic energy entering the substrate through the anchors is eventually dissipated. The semi-analytical evaluation of anchor losses has been addressed in several papers with different levels of accuracy [3, 6]. These contributions consider a resonator resting on elastic half-spaces and assume a weak coupling, in the sense that the mechanical mode, as well as the mechanical actions transmitted to the substrate, are those of a rigidly clamped resonator. The displacements and rotations induced in the half-space are provided by suitable Green's functions. Photiadis, Judge et al. [7] studied analytically the case of a 3D cantilever beam attached either to a semi-infinite space or to a semi-infinite plate of finite thickness. Their results are based on the semi-exact Green's functions established in [4]. More recently Wilson-Rae et al. [9, 10] generalized all these approaches using the involved framework of radiation tunnelling in photonics. Unfortunately, these contributions provide estimates of quality factors that differ quantitatively. In this paper we revisit the procedure of [7], which rests on simple mechanical principles, but starting from the exact Green's functions for the half space studied by Pak [14]. Through a careful analysis utilizing the theory of residues and inspired by the work of Achenbach [15], we show that the results obtained coincide exactly with those of [9], but for the case of torsion.

1 INTRODUCTION: ANALYTICAL ESTIMATE OF DISSIPATION

Following a rather standard procedure [6, 7, 8], in this Section we describe the simplest possible analytical (or semi-analytical) approach based on a decoupling assumption. Let us consider a structure, like the beam of Figure 1, attached to semi-infinite elastic spaces and vibrating in one of its fundamental modes with angular frequency ω. The number of anchor points is irrelevant and the procedure must be identically repeated for all of them.

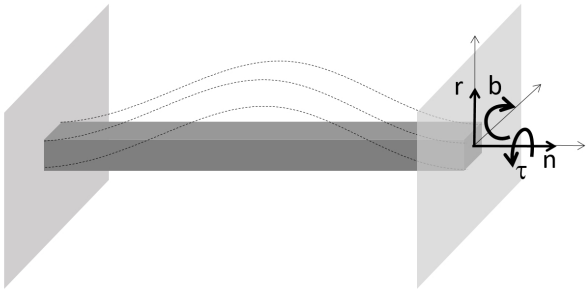


Figure 1: Sketch of a bistable doubly clamped beam

As a consequence we focus on a specific attachment point and start considering it as perfectly rigid. Standard theories of structural mechanics permit to express concentrated forces and couples exerted by the structure on the support. These generally include a constant component (due for instance to pre-stresses or initial deformation) and a sinusoidal varying contribution (see Figure 1 for the notation):

$$\text{axial force}: \quad n(t) = n_0 + Ne^{i\omega t} \tag{2}$$

$$\text{shear force}: \quad r(t) = r_0 + Re^{i\omega t} \tag{3}$$

$$\text{bending couple}: \quad b(t) = b_0 + Be^{i\omega t} \tag{4}$$

$$\text{torque}: \quad \tau(t) = \tau_0 + Te^{i\omega t} \tag{5}$$

The shear force and bending couple have in general two components which are treated in the same manner.

We now introduce the decoupling assumption, according to which frequencies, forces and couples are not significantly altered if the rigid support is replaced with a deformable half space. These concentrated actions induce displacements and rotations:

$$n(t) \quad \rightarrow \quad d(t) = d_0 + De^{i\omega t} \tag{6}$$

$$r(t) \quad \rightarrow \quad v(t) = v_0 + Ve^{i\omega t} \tag{7}$$

$$b(t) \quad \rightarrow \quad \phi(t) = \phi_0 + \Phi e^{i\omega t} \tag{8}$$

$$\tau(t) \quad \rightarrow \quad \psi(t) = \psi_0 + \Psi e^{i\omega t} \tag{9}$$

where D, V, Φ, Ψ are in general complex variables and denote the amplitude of the time dependent part of, respectively, axial and tangential displacements, bending and torsional rotations. These are known as Green's functions for the half elastic space and it is worth stressing that the real part of these kinematic quantities is in general unbounded.

However, the dissipation over one cycle due to the scatter of elastic waves in the infinite half-space is:

$$\Delta W = -\pi \left(\text{Im}[D]N + \text{Im}[V]R + \text{Im}[\Phi]B + \text{Im}[\Psi]T \right)$$

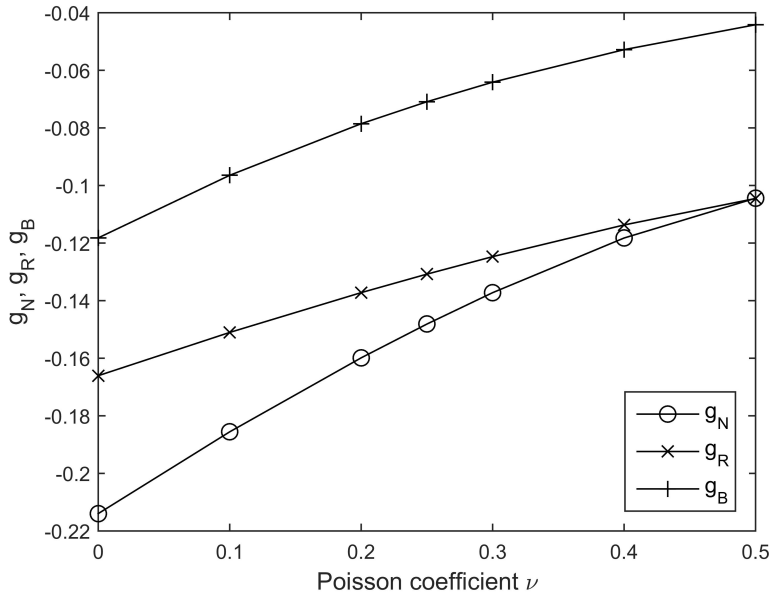


Figure 2: Functions g_N (eq.21), g_R (eq.28) and g_B (eq.24)

and only depends of the bounded imaginary part of the Green's functions.

In particular, setting $k_\mathrm{T} = \omega/c_\mathrm{T}$, $c_\mathrm{T} = \sqrt{\mu/\rho}$, in Section 2 we show that

$$\mathrm{Im}[D] = N\frac{k_\mathrm{T}}{\mu}g_N(\nu) \qquad \mathrm{Im}[V] = R\frac{k_\mathrm{T}}{\mu}g_R(\nu) \tag{10}$$

$$\mathrm{Im}[\Phi] = B\frac{k_\mathrm{T}^3}{\mu}g_B(\nu) \qquad \mathrm{Im}[\Psi] = -T\frac{k_\mathrm{T}^3}{\mu}\frac{1}{12\pi} \tag{11}$$

where g_N, g_R and g_B are plotted versus the Poisson coefficient ν in Figure 2. Most of these functions, rigorously established starting from the work by Pak [14], are similar to analogous results published in [7], but differ quantitatively. On the contrary, the numerical values of the g functions coincide, but for $\mathrm{Im}[\Psi]$, with the expressions given in [9] starting from a totally different perspective.

Finally, summing over all the anchor points:

$$\Delta W = -\pi \sum_i \left(N_i^2 \frac{k_\mathrm{T}}{\mu} g_N(\nu) + R_i^2 \frac{k_\mathrm{T}}{\mu} g_R(\nu) + B_i^2 \frac{k_\mathrm{T}^3}{\mu} g_B(\nu) - T_i^2 \frac{k_\mathrm{T}^3}{\mu} \frac{1}{12\pi} \right) \tag{12}$$

In order to apply eq.(12) only the expressions of R, N, B, T are required. In many cases fully analytical estimates are available, like for a cantilever beam in axial or bending vibrations. More in general, a numerical tool is required, as for the case of the buckled beam of interest in this paper, discussed in the following section.

2 DISPLACEMENTS AND ROTATIONS DUE TO POINT LOADS ON AN ELASTIC HALF-SPACE

The general procedure proposed by Pak [14] is here employed to derive the surface displacements and rotations induced by harmonic point forces and couples exerted on the surface of an elastic half-space of outward normal $\mathbf{e}_z$. In [14] forces, represented as stress discontinuities,

are distributed on a circle of radius a. By suitable specifying their form and taking the limit $a \to 0$, all the concentrated loads of interest herein can be recovered. Starting from formulas (18) and (25) provided by [14], the Green's functions for bending and torque couples are also obtained. Next, focusing on the imaginary part of interest for the dissipation, free terms are evaluated using the theory of residua and results are provided in the form of at most weakly singular integrals, in general functions of the Poisson coefficient. All the results are presented graphically in Figure 2.

Here we set $k_{\mathrm{T}} = \omega/c_{\mathrm{T}}, k_{\mathrm{L}} = \omega/c_{\mathrm{L}}, \eta = k/k_{\mathrm{T}}$ (c_T and c_L are shear and longitudinal wave velocities) and

$$\alpha^2 = \frac{c_T^2}{c_L^2} = \frac{1-2\nu}{2(1-\nu)} \tag{13}$$

$$q^2 = k^2 - \omega^2/c_{\mathrm{T}}^2 = k^2 - k_{\mathrm{T}}^2 = k_{\mathrm{T}}^2(\eta^2 - 1) \tag{14}$$

$$p^2 = k^2 - \omega^2/c_{\mathrm{L}}^2 = k^2 - k_{\mathrm{L}}^2 = k_{\mathrm{T}}^2(\eta^2 - \alpha^2) \tag{15}$$

$$G = (k^2+q^2)^2 + 4k^2pq = k_{\mathrm{T}}^4\left((2\eta^2-1)^2 + 4\eta^2\sqrt{\eta^2-\alpha^2}\,\sqrt{\eta^2-1}\right) \tag{16}$$

$$F = (k^2+q^2)^2 - 4k^2pq = k_{\mathrm{T}}^4\left((2\eta^2-1)^2 - 4\eta^2\sqrt{\eta^2-\alpha^2}\,\sqrt{\eta^2-1}\right) = k_{\mathrm{T}}^4 f(\eta) \tag{17}$$

Vertical displacement due to a vertical force. The distribution of surface stresses:

$$t_z = \frac{1}{\pi a^2}$$

over the circle of radius a centered at the origin induces a unit vertical force. From [14], the resulting displacement w along the z axis is:

$$w(r) = \frac{1}{\pi\mu a}\int_0^\infty \Omega(k) J_1(ak) J_0(kr)\mathrm{d}k \tag{18}$$

where J_m is the m-th order Bessel function, $r = |\boldsymbol{y} - \boldsymbol{x}|$ and

$$\Omega = \frac{1}{2k_T^2}\left(-p + \frac{k^2}{q} - \frac{G}{F}\left(p + \frac{k^2}{q}\right) + \frac{8k^2p(k^2+q^2)}{F}\right) = -\frac{1}{k_T}\frac{\sqrt{\eta^2-\alpha^2}}{f(\eta)} \tag{19}$$

where $f(\eta)$ has been defined in eq.(17). Equation (18) can be rewritten:

$$w(r) = -\frac{1}{\pi\mu a}\int_0^\infty \frac{\sqrt{\eta^2-\alpha^2}}{f(\eta)} J_0(k_T\eta r) J_1(k_T\eta a)\mathrm{d}\eta \tag{20}$$

It is worth stressing that the real part has a potential singular behavior at the origin, as expected. However we are interested only in the imaginary part of $D = \lim_{r\to 0} w(r)$ which is smooth, since the integrand in eq.(18) is real for $\eta > 1$. Hence the limits $r \to 0$, $a \to 0$ can be safely taken.

The integral has a pole in $\eta = \eta_R$ such that $f(\eta_R) = 0$, which corresponds to Rayleigh waves, with $\eta_R > 1$. Following [15] this gives the free term:

$$D^R(r) = -\frac{k_T}{2\pi\mu}\left(-\mathrm{i}\pi\frac{\eta_R\sqrt{\eta_R^2-\alpha^2}}{f'(\eta_R)}\right) = \mathrm{i}\frac{k_T}{\mu}\left(\frac{\eta_R\sqrt{\eta_R^2-\alpha^2}}{2f'(\eta_R)}\right)$$

The integral in eq.(20) has to interpreted in the Cauchy principal value sense. Globally:

$$\mathrm{Im}[D] = N\frac{k_\mathrm{T}}{\mu}g_N(\nu) \tag{21}$$

with:

$$g_N(\nu) = -\frac{1}{2\pi}\mathrm{Im}\left[\int_0^1 \frac{\eta\sqrt{\eta^2-\alpha^2}}{f(\eta)}\mathrm{d}\eta\right] + \frac{\eta_R\sqrt{\eta_R^2-\alpha^2}}{2f'(\eta_R)}$$

It is worth stressing that the final results coincides with that provided by Achenbach ([15], eq.77) using a different procedure.

Rotation due to bending moment. The distribution of surface stresses

$$t_z = \frac{4}{\pi a^4} r\cos\theta = \frac{2}{\pi a^4} re^{i\theta} + \frac{2}{\pi a^4} re^{-i\theta}$$

over the circle of radius a is equivalent to a unit bending couple around the $\theta = \pi/2$ axis. Indeed, in order to apply the procedure of [14], t_z needs to be expressed in polar coordinates over the circle as:

$$t_z = \sum_m t_{z,m}(r)e^{im\theta}$$

The induced radial displacement on the surface is:

$$w(r) = \frac{4\cos\theta}{\pi\mu a^2}\int_0^\infty \Omega(k)J_2(ak)J_1(kr)\mathrm{d}k \tag{22}$$

which, by differentiation, yields the rotation ϕ around the $\theta = \pi/2$ axis:

$$\phi = \left.\frac{\partial w(r)}{\partial r}\right|_{\theta=0} = \frac{2}{\pi\mu a^2}\int_0^\infty \Omega J_2(ak)k(J_0(kr) - J_2(kr))\mathrm{d}k \tag{23}$$

Since we are only interested in the imaginary part of Φ and the integrand is real for $\eta > 1$, the rotation for $r \to 0$, $a \to 0$, is bounded and:

$$\mathrm{Im}[\Phi] = B\frac{k_\mathrm{T}^3}{\mu}g_B(\nu) \tag{24}$$

with:

$$g_B(\nu) = -\frac{1}{4\pi}\mathrm{Im}\left[\int_0^1 \frac{\eta^3\sqrt{\eta^2-\alpha^2}}{f(\eta)}\mathrm{d}\eta\right] + \frac{\eta_R^3\sqrt{\eta_R^2-\alpha^2}}{4f'(\eta_R)}$$

where the second term represents the contribution of Rayleigh waves at $f(\eta_R) = 0$.

Tangential displacement due to tangential force. Like the vertical force also this case is already treated in [14]. The surface stress distribution

$$t_\theta = -\frac{1}{\pi a^2}\sin\theta \qquad t_r = \frac{1}{\pi a^2}\cos\theta$$

over the circle of radius a, which is equivalent to a unit horizontal force, generates the radial displacement u_r:

$$u_r = \frac{\cos\theta}{2\pi\mu a}\left[\int_0^\infty (\gamma_1+\gamma_2)J_0(kr)J_1(ak)\mathrm{d}k + \int_0^\infty (\gamma_2-\gamma_1)J_2(kr)J_1(ak)\mathrm{d}k\right] \tag{25}$$

with

$$\gamma_1 = \frac{1}{2k_T^2}\left(\frac{k^2}{p} - q - \frac{G}{F}\left(\frac{k^2}{p}+q\right) + \frac{8k^2q(k^2+p^2)}{F}\right) = -\frac{1}{k_T}\frac{\sqrt{\eta^2-1}}{f(\eta)} \tag{26}$$

$$\gamma_2 = \frac{1}{q} = \frac{1}{k_T}\frac{1}{\sqrt{\eta^2-1}} \tag{27}$$

The displacement V at the point of application of the load and $\theta = 0$ is the limit of u_r for $r \to 0$, $a \to 0$, and

$$\mathrm{Im}[V] = R\frac{k_\mathrm{T}}{\mu}g_R(\nu) \tag{28}$$

having set:

$$g_R(\nu) = -\frac{1}{4\pi}\int_0^1 \frac{\eta h_1(\eta)}{f(\eta)\sqrt{\eta^2-1}}\mathrm{d}\eta + \frac{\eta_R h_1(\eta_R)}{4f'(\eta_R)\sqrt{\eta_R^2-1}} \tag{29}$$

with:

$$h_1(\eta) = -2 - 4\eta^4 + \eta^2\left(5 + 4\sqrt{\eta^2-1}\sqrt{\eta^2-\alpha^2}\right) \tag{30}$$

Rotation due to torque. Finally, the distribution of surface stresses

$$t_\theta = \frac{2}{\pi a^4}r$$

over the circle of radius a is equivalent to a unit torque and induces the circumferential displacement

$$u_\theta = \frac{2}{\pi\mu a^2}\int_0^\infty \gamma_2(k)J_2(ak)J_1(kr)\mathrm{d}k \tag{31}$$

and the torque angle

$$\psi = \frac{\partial u_\theta}{\partial r} = \frac{2}{\pi\mu a^2}\int_0^\infty \gamma_2(k)J_2(ak)(J_0(kr)-J_2(kr))k\mathrm{d}k \tag{32}$$

with γ_2 defined in eq.(27). In this case the free term due to Rayleigh waves is absent and, all the limits taken, the imaginary part of the rotation Ψ at the origin is

$$\mathrm{Im}[\Psi] = T\frac{k_\mathrm{T}^3}{\mu}g_T \tag{33}$$

with:

$$g_T = \frac{1}{8\pi}\mathrm{Im}\left[\int_0^1 \frac{\eta^3}{\sqrt{\eta^2-1}}\mathrm{d}\eta\right] = -\frac{1}{12\pi}$$

It is worth stressing that g_T is independent of the Poisson coefficient, like in [7], but differs from the results of [9].

3 APPLICATIONS: AXIAL AND BENDING MODES

Simple applications of these formulas give estimates of the quality factors of cantilever beams of length L and cross section area A, resting on an elastic half space. For simplicity, the half-space is assumed to be made of the same isotropic material as the beam. In the case of axial vibrations for a cantilever on a rigid support the axial displacement reads:

$$u = U \sin(k_{\mathrm{B}} x) e^{i\omega t}$$

with:

$$k_{\mathrm{B}} = (1+2m)\frac{\pi}{2L}, \quad \omega = \sqrt{\frac{E}{\rho}}\, k_{\mathrm{B}}$$

The maximum value of the stored elastic energy is:

$$W = \frac{1}{2} U^2 E A k_{\mathrm{B}}^2 \frac{L}{2}$$

and the force exerted on the support $N = EAk_{\mathrm{B}}U$. Assuming that ω, N, W are not significantly altered if the rigid support is replaced with a deformable half space, N induces a displacement D of the half space given by eq.(21) and the dissipation is:

$$\Delta W = -\pi N \,\mathrm{Im}[D] = \pi E^2 A^2 k_{\mathrm{B}}^2 U^2 \frac{1}{\mu} k_{\mathrm{T}}\, g_N(\nu)$$

leading to:

$$Q = \frac{L}{EA} \frac{\mu c_{\mathrm{T}}}{g_N(\nu)} \frac{1}{\omega} \tag{34}$$

Similarly, for a bending mode $\psi(x)U$ characterized by a given wave number k_B (e.g. $k_B L = 1.875$ in the first mode) and normalised such that

$$\int_0^L \psi^2 \mathrm{d}s = L,$$

the maximum stored energy is $W = (1/2) EIL k_B^4 U^2$, while the bending couple and shear force read

$$B = 2EIk_B^2 U, \qquad R = 2EIk_B^3 \beta U, \quad \text{with} \quad \beta = \frac{\sin k_B L - \sinh k_B L}{\cos k_B L + \cosh k_B L}$$

If only the shear force is considered (bending dissipation is usually negligible):

$$Q = \frac{L}{4EJk_B^2\beta} \frac{\mu c_{\mathrm{T}}}{g_R(\nu)} \frac{1}{\omega} \tag{35}$$

where $\omega = \sqrt{EI/(\rho A)} k_B^2$. Formulas (34) and (35) coincide with the ones given by [9], Table 1.

4 APPLICATION: ANALYSIS OF A BISTABLE BEAM

The motion of a beam of length L subjected to a compressive force P and undergoing large displacements is governed, as a first approximation, by the non-linear equation:

$$\rho A \frac{\partial^2 w}{\partial t^2} + EI \frac{\partial^4 w}{\partial x^4} + \left[P - \frac{EA}{2L} \int_0^L \left(\frac{\partial w}{\partial x} \right)^2 \mathrm{d}x \right] \frac{\partial^2 w}{\partial x^2} = 0 \tag{36}$$

where w is the beam deflection, A is the cross-section area, I is the inertia modulus and E is the Young modulus. In eq.(36) the classical equation of slender beams has been corrected for the presence of the uniform axial force

$$n(t) = -P + \frac{EA}{L} \int_0^L \frac{1}{2} \left(\frac{\partial w}{\partial x} \right)^2 \mathrm{d}x$$

In particular we start considering the quasi-static post-buckling response of a clamped-clamped beam, i.e. the evolution beyond the critical load $P_c = 4\pi^2 EI/(L^2)$ as P is slowly increased.

Let ϕ denote the buckling mode:

$$\phi = \frac{1}{2} \left(1 - \cos \frac{2\pi x}{L} \right)$$

Assuming that the beam deflection has the expression $w(x) = b\phi(x)$, one obtains a non-linear relationship between P and b:

$$P = P_c + b^2 \frac{EA}{2L} \int_0^L \frac{\mathrm{d}^2 \phi}{\mathrm{d}x^2} \mathrm{d}x$$

We now study the small vibrations around a postbuckling state, characterized by a given b:

$$w(x,t) = b\phi(x) + u(x,t), \qquad |u(x,t| << b|\phi(x)| \tag{37}$$

A linearization of eq.(36) yields:

$$\rho A \frac{\partial^2 u}{\partial t^2} + EI \frac{\partial^4 u}{\partial x^4} + P_c \frac{\partial^2 u}{\partial x^2} - \frac{EAb^2}{L} \frac{\mathrm{d}^2 \phi}{\mathrm{d}x^2} \int_0^L \frac{\mathrm{d}\phi}{\mathrm{d}x} \frac{\partial u}{\partial x} \mathrm{d}x = 0 \tag{38}$$

An explicit and simple expression for the quality factor can be obtained with very good approximation for small values of b/V. In this case it is reasonable to assume that $\Phi(x)$ is the buckling mode. Assuming that the greatest contribution to dissipation is due to axial loads:

$$N = \alpha EA \frac{b}{L} \frac{\pi^2}{2} \quad \rightarrow \quad \Delta W = N^2 \frac{b}{\mu L^2} \left(4\pi^3 \sqrt{\frac{1+\nu}{3}} g_N(\nu) \right)$$

and the quality factor is:

$$Q \frac{HV^2}{L^3} = \frac{1}{\tilde{b}} \left(\frac{1}{8\pi^3} \frac{1}{g_N(\nu)} \frac{1}{1+\nu} \sqrt{\frac{3}{1+\nu}} \right) \tag{39}$$

REFERENCES

[1] Ardito R., Comi C., Corigliano A., Frangi A., Solid damping in micro electro mechanical systems, *Meccanica*, 43, 419, 428

[2] Segovia-Fernandez J., Cremonesi M., Cassella C., Frangi A., Piazza G. Anchor Losses in AlN Contour Mode Resonators, *JMEMS*, 24, 265–275, 2015

[3] Cross M.C., Lifshitz R., Elastic wave transmission at an abrupt junction in a thin plate with application to heat transport and vibrations in mesoscopic systems, *Physical Review B*, 64, 085324, 1-22

[4] Bycroft G.N., Forced Vibrations of a Rigid Circular Plate on a Semi-Infinite Elastic Space and on an Elastic Stratum, *Phil. Transactions of the Royal Society of London. Series A*, 248, 327–368, 1956

[5] Miller G.F., Pursey H. The Field and Radiation Impedance of Mechanical Radiators on the Free Surface of a Semi-Infinite Isotropic Solid, *Proc. of the Royal Society of London. Series A*, 223, 521–541, 1954

[6] Jimbo Y., Itao K., Energy loss of a cantilever vibrator, *J. Horological Inst. Jpn.*, 47, 1–15, 1968

[7] Judge J.A., Photiadis D.M., Vignola J.F., Houston B.H., Jarzynski J., Attachement losses of micromechanical and nanomechanical resonators in the limits of thick and thin support structures, *J. Applied Physics*, 101, 013521, 2007

[8] Photiadis D.M., Judge J.A., Attachement losses of high Q oscillators, *Applied Physics Letters*, 85, 482–484, 2004

[9] Wilson-Rae I., Intrinsic dissipation in nanomechanical resonators due to phonon tunnelling, *Physics review B*, 77, 245418, 2008

[10] Cole G.D., Wilson-Rae I., Werbach K., Vanner M.R., Aspelmeyer M., Phonon-tunnelling dissipation in mechanical resonators, *Nature Communications*, 2, Article number: 231, 2011

[11] Frangi A., Bugada A., Martello M. and Savadkoohi P.T. Validation of PML-based models for the evaluation of anchor dissipation in MEMS resonators, *Eur. J. Mech.*, 37, 256–265, 2013

[12] Frangi A., Cremonesi M., Jaakkola A. and Pensala T. Analysis of anchor and interface losses in piezoelectric MEMS resonators, *Sens. Actuators A, Phys.*, 190, 127–135, 2013

[13] Frangi A., De Masi B., Confalonieri F. Zerbini S., Threshold shock sensor based on a bistable mechanism: design, modelling and measurements, *JMEMS*, 24, 2019–2026, 2015

[14] Pak R.Y.S., Asymmetric wave propagation in an elastic half-space by a method of potentials, *Journal of Applied Mechanics*, 54, 121–126, 1987

[15] Achenbach J.D., Calculation of wave fields using elastodynamic reciprocity, *Int. J. Solids Structures*, 37, 7043–7053, 2000

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EFFICIENT BOUNDARY ELEMENT FORMULATION OF THERMOELASTICITY

R. Rott[1], **M. Schanz**[1]

[1]Institute of Applied Mechanics, Graz University of Technology
Technikerstraße 4, 8010 Graz, Austria
e-mail: {r.rott, m.schanz}@tugraz.at

Keywords: Uncoupled Quasistatic Thermoelasticity, Boundary Element Method, Convolution Quadrature Method

Abstract. *The problem of thermoelasticity is present in many different areas of solid mechanics. It describes the effects of thermal as well as mechanical loads on an elastic structure. We use the uncoupled quasistatic formulation of thermoelasticity (UQT), in a linear model and apply the Boundary Element Method. The UQT formulation is applicable in most cases, where the mechanical load is constant or slowly varying in time. Here, the influence of the elastic deformations on the heat distribution is neglected. This leaves us with a decoupled system of differential equations, consisting of the heat equation and an elastic equation, which accounts for thermal and mechanical loads. In the elastic equation the thermal field variables are introduced via convolutions. We apply three different methods for the calculation of these convolutions to find the elastic field variables and compare their computation times.*

1 INTRODUCTION

For a large class of problems in solid mechanics not only mechanical loads but also thermal influences have to be considered to describe the behavior of a structure correctly. For an elastic deformation under the influence of heat, the theory of thermoelasticiy is well known for over 5 decades (see for example [1–3]). Today we can apply the theory in many different fields, for example in structural engineering or in the description of a hot forming tool. This tool is used in the production of vehicle parts. The process of hot forming makes use of the thermomechanical properties of the material. In the process, the tool is directly heated and cooled down thereafter. For this industrial application the thermal problem has been simulated and a fast method was developed in a previous work by Messner and Schanz [4]. It is work in progress to extend this thermal problem to the elastic deformations, caused by the temperature change.

Therefore, in our studies we aim to develop a fast numerical method to calculate the effect of such thermomechanical loads and to simulate the elastic deformations. The model we use is linear in its geometry, as well as in its elastic and thermal behavior and thus applicable only to small deformations. For this purpose the Boundary Element Method (BEM) is a good framework for our numerical model. In the BEM the discretization of the body is reduced to a discretization of the body's surface. Results at an arbitrary point of the volume can be calculated in a subsequent step. In particular, we use the theory of Uncoupled Quasistatic Thermoelasticity (UQT), which is a special case of thermoelasticity that meets the requirements of our applications. The UQT has been studied [3, 5, 6] and applied by different authors (see [7, 8]).

In the following, we will describe the theoretical approach we use and the different methods to solve the equations. We show numerical results and compare them to a known analytic solution. Additionally, we present a comparison of computation times for these results.

2 BASIC EQUATIONS

The general case of coupled thermoelasticity can be described by the following set of coupled differential equations [6]

$$\kappa\,\theta_{,jj} - \dot{\theta} - \kappa\,\alpha\,\dot{u}_{j,j} + \psi = 0 \tag{1}$$

$$\mu\,u_{i,jj} + (\lambda+\mu)\,u_{j,ij} - (3\lambda+2\mu)\,\alpha\,\theta_{,i} + f_i = \rho\,\ddot{u}_i \quad , \tag{2}$$

with

u_i	displacement	λ, μ	Lamé isothermal elastic constants
θ	temperature	α	thermal expansion coefficient
f_i	body forces	κ	thermal diffusivity
ψ	body heat sources	ρ	mass density.

For this set of coupled differential equations, a simplification can be justified, which applies to many technical applications. In a process of slowly varying mechanical load, the heat production or consumption due to mechanical stress can be neglected. For such a slow or quasistatic process the elastic inertia terms can be neglected as well. With these assumptions we end up with the set of uncoupled quasistatic differential equations of thermoelasticity (in the following denoted by UQT) [5], where the term describing the influence from mechanical stress on the heat evolution vanishes. Additionally, we do not consider body forces or inner heat sources in

this model

$$\kappa\, \theta_{,jj} - \dot{\theta} = 0 \tag{3}$$

$$\mu\, u_{i,jj} + (\lambda + \mu)\, u_{j,ij} - (3\lambda + 2\mu)\, \alpha\, \theta_{,i} = 0 \quad . \tag{4}$$

With this set of equations, the thermal evolution in equation (3) is completely independent, or uncoupled, from the mechanical variables. This provides us with the possibility to solve this differential equation independently. As a result, the thermal quantities are determined and are known in advance for the solution of the elastic differential equation (4), where they can be considered as a thermal load.

To apply the Boundary Element Method we need to formulate boundary integral equations for this set of differential equations, which are [7]

$$c_{\theta\theta}(\xi)\, \theta(\xi,t) = \int_\Gamma \{\, [g_{\theta\theta} * q](x,t) - \, [f_{\theta\theta} * \theta](x,t)\} \; d\Gamma \tag{5}$$

$$c_{ij}(\xi)\, u_j(\xi,t) = \int_\Gamma \{\, g_{ij} t_j(x,t) - f_{ij} u_j(x,t) + \, [g_{i\theta} * q](x,t) - \, [f_{i\theta} * \theta](x,t)\} \; d\Gamma \;, \tag{6}$$

where $i, j = 1, 2, 3$. The functions g and f are the fundamental solutions, which can be found in [1, 3, 6]. While the purely thermal equation (5) is scalar, the elastic equation (6) is vectorial. The jump terms $c_{\theta\theta}$ and c_{ij} follow from shifting the field point ξ from the domain Ω to the boundary Γ.

In our studies, we will focus on the numerical solution of the elastic equation (6), since a fast method for the thermal equation (5) already exists [4].

Spatial Discretization

For a numerical description of our problem we need discretizations of both, space and time. The geometry description is based on an element formulation, establishing the Boundary Element Method. This method uses the triangulation of the geometry, yielding disjoint, adjacent, finite elements τ_l on the boundary Γ

$$\Gamma \approx \Gamma_L = \sum_{l=0}^{L} \tau_l \quad . \tag{7}$$

In contrast to other element methods, for the BEM the discretization of the surface Γ is sufficient, due to the boundary integral equations used. As a consequence, the volume description only depends on the quality of the surface description. The field $y(x,t)$ to be investigated is formed by a linear combination of shape functions $\phi(x)$. Depending on the choice of the shape functions, the physical quantity is determined by a sum over element functions ϕ_n and nodal values y_n,

$$y(x,t) = \sum_{n=0}^{N} \phi_n(x)\; y_n(t) \quad , \tag{8}$$

where N is the number of shape functions. The field variable y represents all four fields of interest, the displacement u_i, the temperature θ, the tractions t_i and the heat flux q.

3 TREATMENT OF THERMAL QUANTITIES

In the elastic equation, the thermal variables are introduced via convolutions over time. The boundary integral equation (6) convolves known boundary values with the fundamental solutions of the coupling terms as a right hand side. The treatment of these convolutions can be of different kind.

3.1 Analytic integration in time

We introduce discrete time steps

$$t_m = m\Delta t \qquad m = 0, 1, ..M \tag{9}$$

and shape functions $\psi_m(t)$ in time. The field is consequently a linear combination of these time shape functions

$$y_n(t) = \sum_{m=0}^{M} \psi^m(t)\, y^m \quad , \tag{10}$$

where M denotes the number of contributing time shape functions. Inserting these shape functions in the first convolution integral in 6, yields

$$\int_0^t g_{i\theta}(x, t-\tau) q(x,\tau)\, d\tau = \sum_{m=1}^{M} \int_{(m-1)\Delta t}^{m\Delta t} g_{i\theta}(x, M\Delta t - \tau)\, q^m(x)\psi^m(t)\, d\tau \; . \tag{11}$$

Choosing constant discontinuous shape functions in time results in variables, which are only dependent on the present interval. As a consequence, it becomes possible to pull out the field quantity from the time integral [7]. An analytic integration of the fundamental solutions is favorable. It exists for the heat-coupling fundamental solution and was developed by Dargush [7]. With these assumptions the convolution involves only the multiplication of matrices and boundary value vectors

$$\begin{aligned} \int_0^t g_{i\theta}(x, t-\tau) q(x,\tau)\, d\tau &= \sum_{m=1}^{M} q^m(x) \int_{(m-1)\Delta t}^{m\Delta t} g_{i\theta}(x, M\Delta t - \tau)\, d\tau \\ &= \sum_{m=1}^{M} q^m(x)\, G_{i\theta}^{M+1-m}(x) \; , \end{aligned} \tag{12}$$

where $G_{i\theta}$ is the time integrated fundamental solution. A triangular block Toeplitz structure is formed. This is a convenient method involving nothing but variables in the time domain.

3.2 The Convolution Quadrature Method (CQM)

The Convolution Quadrature Method was developed by Lubich [9, 10], to numerically evaluate a convolution, making use of one of the functions in the Laplace domain. The basic idea of the CQM is to express the convolution as a quadrature formula. The weights are derived from one of the function's Laplace transform

$$\mathcal{L}\{g\}(s) = G(s) = \int_0^\infty g(t)\, \mathrm{e}^{-st}\, dt \quad . \tag{13}$$

The entire derivation of the CQM is shown in the Appendix. The strategy leads to the following formulation. First, we discretize in time

$$t_n = n\Delta t \quad , \tag{14}$$

with $n = 0, 1, ..., N$, for the discrete convolution of $N + 1$ time intervals of equal size. Hence, it is

$$\int_0^t g_{i\theta}(x, t-\tau) q(x, \tau)\, d\tau \approx \sum_{k=0}^{n} \omega_{n-k}(\Delta t) q(k\Delta t) \quad . \tag{15}$$

$\omega_n(\Delta t)$ depends on the Laplace transform of the function, $G_{i\theta}(\frac{\gamma(z)}{\Delta t})$, where $\gamma(z)$ is the characteristic function of the underlying multistep method.

3.3 Fast Fourier Transformation (FFT)

The Fourier transform is another method which shifts the problem to the frequency domain, where the convolution becomes simply a multiplication

$$\mathcal{F}\{g * q\}(\omega) = \mathcal{F}\{g\}(\omega)\, \mathcal{F}\{q\}(\omega) = Y(\omega) \quad . \tag{16}$$

Here $\mathcal{F}$ denotes the Fourier transformation of the form

$$\mathcal{F}\{g\}(\omega) = G(\omega) = \frac{1}{\sqrt{2\pi i}} \int_{-\infty}^{\infty} g(t)\, \mathrm{e}^{-i\omega t}\, dt \quad . \tag{17}$$

As the thermoelastic problem is causal, there is no solution at negative times, and the Fourier transformation becomes almost equivalent to the Laplace transform for setting the parameter $s = i\omega$.

Therefore, it is possible to make use of the fundamental solution, which exists for the Laplace domain in an analytic form, and use it for the transformation into Fourier space as well. Finally, an inverse Fourier transformation has to be applied to the product $Y(\omega)$, which yields the result $y(t)$ in time domain

$$\mathcal{F}^{-1}\{Y\}(\omega) = y(t) \quad . \tag{18}$$

For a numerical algorithm of the discrete Fourier transformation there exists the very efficient Fast Fourier Transform algorithm (FFT). It operates at an effort of order $O(NlogN)$.

Even though all methods are given for one of the convolution integrals exemplary here, both convolution integrals in the elastic equation (6) are treated in the same way.

4 RESULTS

4.1 Problem Validation

To validate the results of the uncoupled quasistatic thermoelastic problem, we want to consider a test problem for which an analytic solution exists. In literature there exist several simple and well known problems for the solution of the heat equation alone but not so many with a simple thermoelastic solution as well. There is one popular test case which goes back to Timoschenko and Goodier [11] and was also used by Dargush [7] and Chatterjee [8] for validation.

The problem setting is the following. A cubic body of homogeneous, isotropic material properties is fixed by roller bearings on 5 of its faces. At these faces the heat flux is zero, thus the system is thermally insulated there. The remaining 6th face of the cube is mechanically unconstrained (figure 1). Initially, the cube is in an equilibrium state at zero degrees. At time

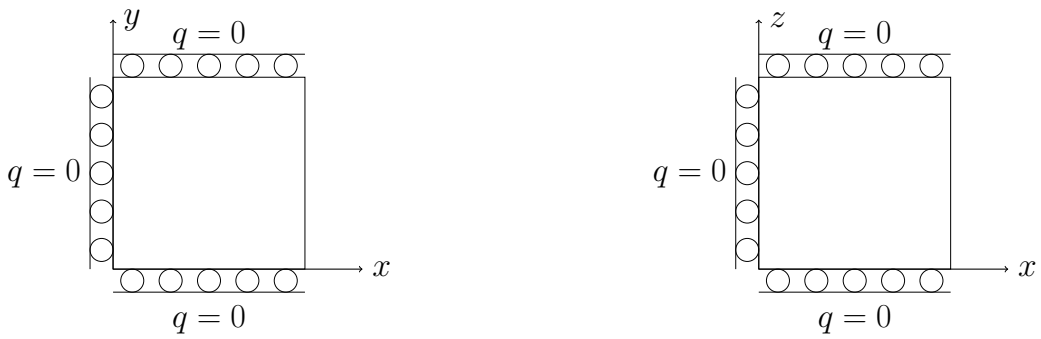


Figure 1: Mechanical and thermal constraints of the unit test cube, in top view and front view.

zero the temperature at the mechanically unconstrained face 6 is raised by 1 degree. As a result, the body expands elastically, as the temperature increases. After a certain time a stationary state is reached. Due to the mechanical boundary conditions the elastic expansion of the cube is only possible in x-direction. The three-dimensional model shows, therefore, a one dimensional behavior, for which solutions are known. The behavior is reduced to a one dimensional problem of the coordinate x with the following analytic solutions for the temperature θ and the displacement u

$$\theta(x,t) = 1 - \frac{4}{\pi}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)} \quad \exp\Big(-\frac{(2n+1)^2\pi^2\kappa\, t}{4L^2}\Big) \quad \cos\Big(\frac{(2n+1)\pi\, x}{2L}\Big)\,, \tag{19}$$

$$\begin{aligned} u(x,t) &= \frac{(3\lambda+2\mu)\alpha}{\lambda+2\mu}\int_0^x \theta(x,t)\,\mathrm{d}y \\ &= \frac{(1+\nu)\alpha}{(1-\nu)}\left(x - \frac{8L}{\pi^2}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)}\,\exp\Big(-\frac{(2n+1)^2\pi^2\kappa\, t}{4L^2}\Big)\,\sin\Big(\frac{(2n+1)\pi\, x}{2L}\Big)\right), \end{aligned} \tag{20}$$

where L is the length of the cube in x-direction. With the relations of the Lamé coefficients of an isotropic material, the prefactor depends only on Poisson's ratio ν and the thermal expansion coefficient α.

For the following calculations the heat equation was solved in a first step, using the CQM. Since a fast method for the thermal problem exists, we want to focus on the deformations only. The elastic part was solved in a second step using the three different methods presented before for the convolution terms. Each method is going to be compared to the analytic solution of our test problem. As our test body, we choose the unit cube (figure 1), with its mechanically unconstrained face at $x = 1$. We show the elastic displacement over time for a point on this surface and denote the surface by $x = 1$.

Discretization The Boundary element mesh chosen for our surface approximation consists of triangular elements. The basic mesh has 24 elements, 4 triangles on each cube face. From this basic mesh 2 refinement levels are created, each splitting a previous element into 4 smaller triangles. In this manner, we receive refinement 1 with 96 boundary elements and refinement 2 of 384 elements. The meshes are shown in figure 2.

Time stepping For choosing an appropriate time step it is necessary to resolve the time dependence sufficiently. Therefore, the time step size highly depends on the material parameters,

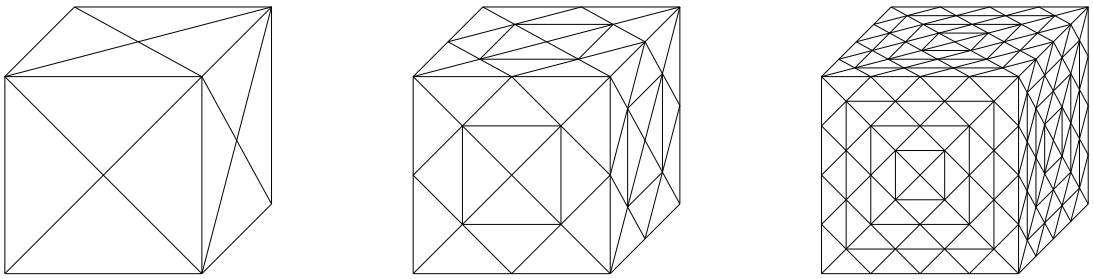


Figure 2: Basic mesh, refinement 1 and refinement 2.

especially on the thermal diffusivity κ and the heat expansion coefficient α. For the test case we choose the material parameters α and κ to be 1 and the Poisson's ratio ν to be zero. With these parameters a time step size of 0.01s is adequate.

Analytic integration in time (AIt)

As described in section 3.1, the fundamental solutions were analytically integrated for the elastic convolutions, which couple the heat quantities to the elastic deformations. The displacement u_x is shown for the basic mesh (24 elements), refinement 1 (96 elements) and refinement 2 (384 elements) and compared to the analytic solution, at a point on the free surface $x = 1$. In figure 3 a good agreement with the analytic solution is shown and in figure 4 convergence is observable.

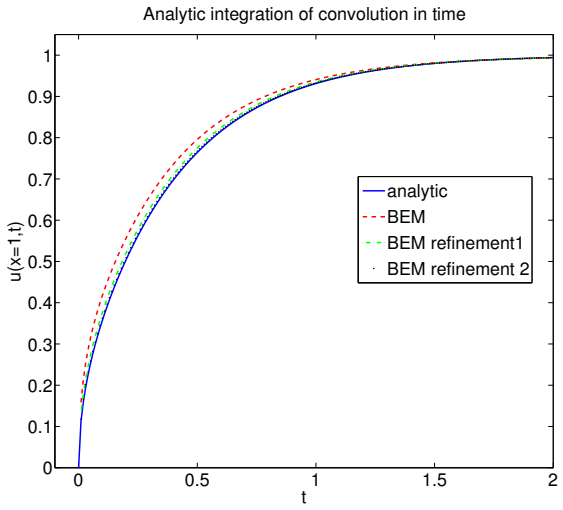


Figure 3: Convolution calculated by analytic integration with a time step size of Δt=0.01 seconds.

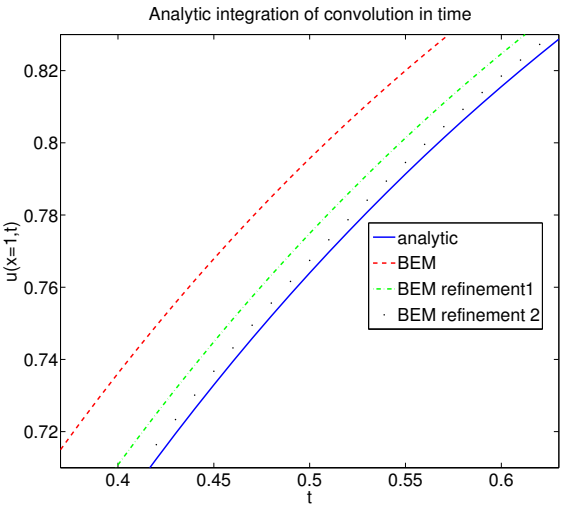


Figure 4: The zoom into the curve of figure 3 shows the convergence with the refinement.

Convolution Quadrature Method

The CQM yields high precision and convergence to the analytic solution as well. The quality is of the same order as for the analytic integration, for each mesh refinement, as shown in figures 5 and 6.

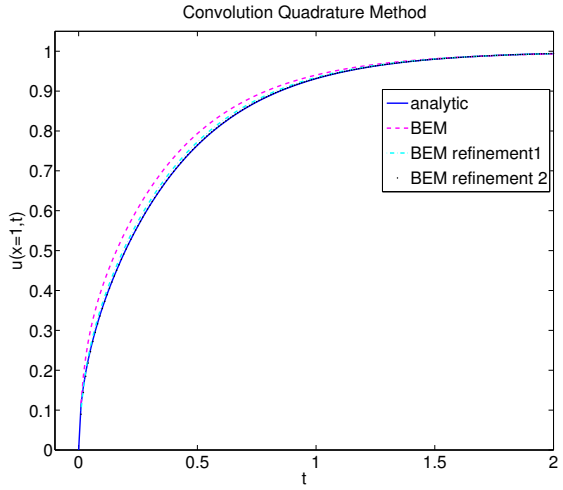


Figure 5: Displacement $u_x(x = 1)$, calculated using the CQM with Δt =0.01s.

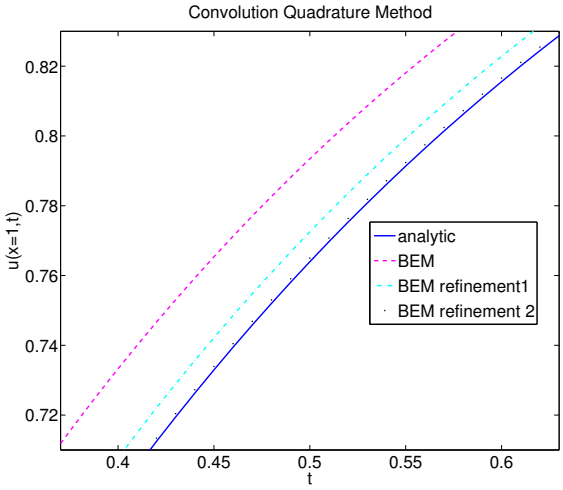


Figure 6: Zoom into figure 5 at around 0.5 seconds.

Fast Fourier Transformation

Calculating the elastic convolutions with the FFT, we once more receive results of comparable quality. The results for the test case are shown in figures 7 and 8.

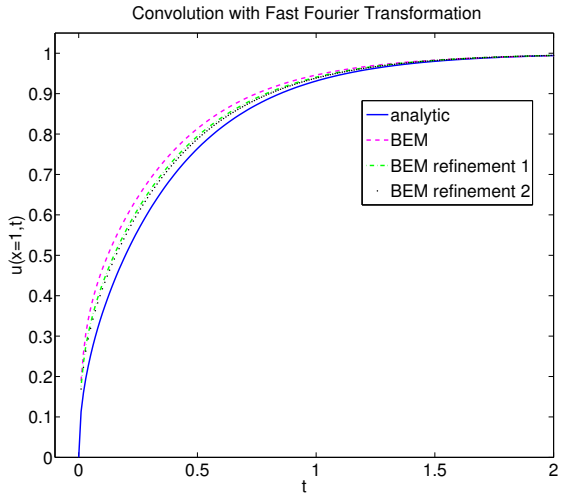


Figure 7: Displacement $u_x(x = 1)$, for the FFT convolution Δt =0.01s.

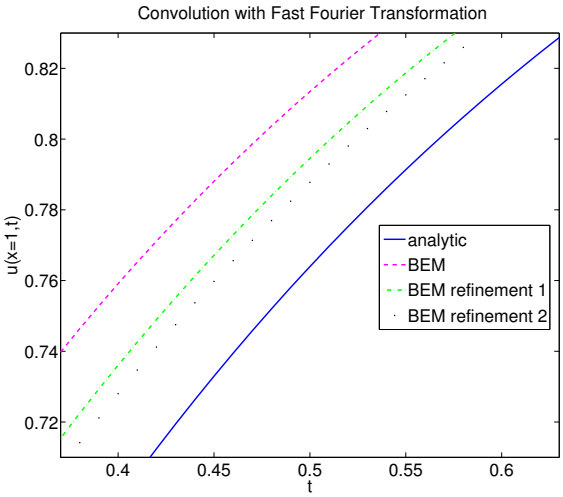


Figure 8: Zoom into figure 7 shows the convergence with the mesh refinement.

The result of refinement 2 is not as good as with CQM or the analytic integration method. This can be improved by choosing the parameters for the FFT differently.

Timings

Finally, we compare the computation time of the elastic equation for the three different methods of the convolution calculation. For each mesh refinement level the timings are shown, where AIt is our first method, which integrates the fundamental solutions analytically in time. The timings were taken on a standard PC using 6 cores. We use the parameters $\Delta t = 0.01$ seconds for the time step, for a total duration of 5 seconds.

	AIt	CQM	FFT
basic mesh	38	35	35
refinement 1	495	457	457
refinement 2	5921	5450	5250

Table 1: Comparison of computation times in seconds for the three convolution methods.

All three methods have similar computation times. The FFT is the fastest, followed by CQM and the analytic integration in time AIt, which is still fast.

5 CONCLUSIONS AND OUTLOOK

We showed the convergence of the displacement results to an analytic solution for the elastic equation of UQT, using three different methods for the convolution, an analytic integration, the CQM and the FFT. The results show a good agreement with the analytic solution, especially for high spatial discretization but already for quite coarse meshes as well. The time steps need to be chosen accordingly, depending on the thermal diffusivity and the heat expansion coefficient. Setting the material parameters to 1, we get a good resolution with Δt =0.01s. The computation time for the elastic equation is of the same order for all three methods. Still an even more efficient method is desirable, which accounts for the decrease of the rate of change of the field variables over time.

For time steps of variable sizes there exists the Generalized Convolution Quadrature Method (GCQM) introduced by Fernandez and Sauter [12]. We plan to apply this method to our problem. For the matrix assembling there are various fast methods available in the literature. It is work in progress to implement further methods for efficiency improvements.

A Appendix: CQM

The Convolution Quadrature Method was introduced by Lubich [9], in 1988 and has since been adapted by other authors such as Banjai and Sauter [13]. The following derivation of the CQM is taken from [14], where further details can be found.

We want to determine the convolution of two functions f and g

$$[f * g](t) = \int_0^t f(t-\tau)\, g(\tau)\, d\tau \quad . \tag{21}$$

The first step is to replace the function f by its inverse Laplace transformation $\hat{f}(s)$ like

$$\int_0^t f(t-\tau)\, g(\tau)\, d\tau = \frac{1}{2\pi i} \lim_{R\to\infty} \int_{c-iR}^{c+iR} \hat{f}(s) \underbrace{\int_0^t \mathrm{e}^s (t-\tau) g(\tau) d\tau}_{x(s,t)}\, ds \tag{22}$$

The term $x(s,t)$ can be interpreted as the solution of a differential equation of first order with zero initial condition

$$\frac{dx(s,t)}{dt} = sx(s,t) + g(t) \quad . \tag{23}$$

This differential equation can be solved by a linear multistep method in a discrete formulation. The introduction of $N+1$ discrete time steps of length Δt leads to a discrete formulation of

our convolution

$$y(t_n) = \frac{1}{2\pi i} \lim_{R\to\infty} \int_{c-iR}^{c+iR} \hat{f}(s) x(s, t_n)\, ds \quad . \tag{24}$$

The solution of $x(s, t_n)$ by a linear multistep method, such as the Backward Differential Formula of order 1 or 2 (BDF1, BDF2), together with some reformulations, which can be found in [14], leads us to the following representation of $x(s, t_n) = x_n$ as a power series

$$\sum_{n=0}^{\infty} x_n z^n = \frac{1}{\frac{\gamma(z)}{\Delta t} - s} \sum_{n=0}^{\infty} g(t_n) z^n \quad , \tag{25}$$

where $\gamma(z)$ is the characteristic function of the chosen multistep method. Plugging x_n into the discrete convolution in equation (24)

$$\sum_{n=0}^{\infty} y_n z^n = \frac{1}{2\pi i} \lim_{R\to\infty} \int_{c-iR}^{c+iR} \hat{f}(s) \frac{1}{\frac{\gamma(z)}{\Delta t} - s} ds \sum_{n=0}^{\infty} g(t_n) z^n \,, \tag{26}$$

with $y(t_n) = y_n$, and applying the residue theorem gives

$$\sum_{n=0}^{\infty} y_n z^n = \hat{f}\left(\frac{\gamma(z)}{\Delta t}\right) \sum_{n=0}^{\infty} g(t_n) z^n \quad . \tag{27}$$

Since we are interested in the solution y_n of the convolution for every time step, we use Cauchy's integral formula for the coefficients of the power series

$$\omega_n(\Delta t) = \frac{1}{2\pi i} \int_{|z|=\rho} \hat{f}\left(\frac{\gamma(z)}{\Delta t}\right) z^{-n-1} dz \quad , \tag{28}$$

where ρ is the radius of analyticity of $\hat{f}\left(\frac{\gamma(z)}{\Delta t}\right)$. Each of the discrete solutions y_n can then be determined by a comparison of coefficients

$$\sum_{n=0}^{\infty} y_n z^n = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \omega_{n-k}(\Delta t) g(k\Delta t) z^n \quad , \tag{29}$$

resulting in

$$y_n = \sum_{k=0}^{n} \omega_{n-k}(\Delta t) g(k\Delta t) \tag{30}$$

for all discrete times $n = 0, 1, ..., N$. With this reformulation we gained a simple multiplication of the function g in time domain by weights, depending on the Laplace transform of the function f and the chosen linear multi-step method for the term $x(s, t_n)$.

A further reformulation has been introduced by Banjai and Sauter [13], which leads to a system of decoupled linear equations in the Laplace domain. By inserting the weights ω_n into equation (30) a representation as an inverse transformation of the CQM can be found. An engineering application of this can be found in [15].

REFERENCES

[1] V. Kupradze, *Three-dimensional problems of elasticity and thermoelasticity*. Elsevier, 2012.

[2] V. Kupradze, T. Burchuladze, Dynamical problems of the theory of elasticity and thermoelasticity. *Itogi Nauki i Tekhniki. Seriya" Sovremennye Problemy Matematiki. Noveishie Dostizheniya"*, **7**, 163–294, 1975.

[3] W. Nowacki, Green's functions for a thermoelastic medium (quasi-static problem). *IASI, Buletinului Institutului Politehnic*, **12**, 83–92, 1966.

[4] M. Messner, M. Schanz, J. Tausch, A fast Galerkin method for parabolic space–time boundary integral equations. *Journal of Computational Physics*, **258**, 15–30, 2014.

[5] V. Sladek, J. Sladek, Boundary integral equation method in thermoelasticity Part III: Uncoupled thermoelasticity. *Applied mathematical modelling*, **8**, 413–418, 1984.

[6] W. Nowacki, *Dynamic problems of thermoelasticity*. Springer, 1975.

[7] P. Dargush, G.F. Banerjee, Boundary element methods in three-dimensional thermoelasticity. *International journal of solids and structures*, **26**, 199–216, 1990.

[8] J. Chatterjee, F. Ma, D. Henry, P. Banerjee, Two-and three-dimensional transient heat conduction and thermoelastic analyses by BEM via efficient time convolution. *Computer methods in applied mechanics and engineering*, **196**, 2828–2838, 2007.

[9] C. Lubich, Convolution quadrature and discretized operational calculus I. *Numerische Mathematik*, **52**, 129–145, 1988.

[10] C. Lubich, Convolution quadrature and discretized operational calculus II. *Numerische Mathematik*, **52**, 413–425, 1988.

[11] J. Timoshenko, S. Goodier, *Theory of elasticity*. McGraw-Hill, 1951.

[12] M. Lopez-Fernandez, S. Sauter, Generalized convolution quadrature with variable time stepping. *IMA Journal of Numerical Analysis*, **33**, 1156–1175, 2013.

[13] L. Banjai, S. Sauter, Rapid solution of the wave equation in unbounded domains. *SIAM Journal on Numerical Analysis*, **47**, 227–249, 2008.

[14] M. Schanz, *Wave propagation in viscoelastic and poroelastic continua: a boundary element approach*, Vol. 2. Springer, 2012.

[15] M. Schanz, W. Ye, J. Xiao, Comparison of the convolution quadrature method and enhanced inverse FFT with application in elastodynamic boundary element method. *Computational Mechanics*, 1–14, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A CHEBYSHEV INTERPOALTION-BASED FAST MULTIPOLE METHOD FOR POROELASTICITY

Barbara Knöbl[1], **Thomas Traub**[1], **and Martin Schanz**[1]

[1] Institute of Applied Mechanics, Graz University of Technology
Technikerstraße 4/II, 8010 Graz, Austria
e-mail: {barbara.knoebl, thomas.traub, m.schanz}@tugraz.at

Keywords: Boundary Element Method, Fast Multipole Method, Poroelasticity, Chebyshev polynomials.

Abstract. *The treatment of porous media with the BEM exhibits some problems, which are common also for the Finite Element Method, but also some BEM specific ones. The three main problems are: First, the multiphase system requires four degrees of freedom per node, leading to large matrices even for small problems. Since the BEM matrices are densely populated, this makes the method prohibitive for large problem sizes. Second, due to the different physical nature of the degrees of freedom the matrix entries vary over several orders of magnitude. Third, the fundamental solution of poroelasticity is computationally expensive.*

We present a FM-BEM that circumvents those points: The Chebyshev interpolation-based FMM significantly reduces the memory consumption of the system matrix and thus allows for larger problem sizes to be treated. As well, it requires fewer evaluations of the fundamental solution. To employ an iterative solver, the use of a transformation of the material data is mandatory.

1 INTRODUCTION

Porous media occur frequently in nature as well as construction materials; probably the best-known example for a natural porous medium is soil. The understanding of wave propagation in such media is of distinct interest for oil and gas explorisation but also for earthquake analysis. In these particular fields the treatment of unbounded domains, such as a halfspace, is required. The Boundary Element Method (BEM) is advantageous over the Finite Element Method (FEM) for the numerical treatment of such geometries, since it requires only the discretisation of the boundary and inherently fulfills the radiation condition [1].

In the most straightforward description the saturated porous media problem shows six unknowns, i.e. the solid's and fluid's displacements $\mathbf{u}$ and $\mathbf{u}^f$, respectively. The description can be reduced to four, using the fluid's pressure p instead of its displacements $\mathbf{u}^f$, which is only possible in the Laplace domain. The BEM requires the knowledge of the fundametal solution of the underlying partial differential equation, which is accordingly also solely available in the Laplace domain. To derive the temporal behavior the Convolution Quadrature Method [2, 3] can be employed.

A review of poroelastic models and their numerical treatment is given in [4]. Early BEM formulations were given by Manolis and Beskos, who presented an $(\mathbf{u}, \mathbf{u}^f)$-description in Laplace domain, and by Cheng et al. [5] and Domínguez [6], publishing frequency domain $(\mathbf{u}, p)$-formulations. An extensive discussion of the collocation BEM for linear poroelasticity is given by Schanz [7]. A symmetric Galerkin BEM, which is advantageous for FEM-BEM-Coupling, was presented in [8].

The application of so-called fast methods is necessary for the efficient treatment of the porous media problem with BEM for the following two reasons: The complicated structure of the fundamental solution makes the evaluation of the matrix entries computationally expensive, and the multiphase system requires four degrees of freedom per node, leading to large matrices even for small problems. Furthermore, it is well known that the BEM matrices are densely populated. Hence, methods are needed that reduce memory and the number of function evaluations; so-called blackbox methods would be preferable. The most promising candidates are the Adaptive Cross Approximation (ACA) [9] and the Fast Multipole Method (FMM). The classical FMM [11] uses an expansion into spherical harmonics to construct a separable expansion of the kernel function. A review of the application of the FMM in BEM is given in [12]. Regarding vector-valued problems, elastodynamics has been successfully treated with the classical FMM, e.g. in [13, 14]. Among the FMM variants the Chebyshev interpolation FMM [10] seems to be favorable for poroelasticity, since the fundamental solution is only evaluated at a small set of spatial points, avoiding an analytical expansion.

In this paper we present a Chebyshev interpolation-based FM-BEM for poroelasticity. In the next section, we briefly summarize the BEM formulation for poroelasticity in Laplace domain. Then we describe the employed Chebyshev interpolation-based FMM where we focus on the differences to the standard FMM. Finally we present numerical experiments to validate the presented algorithms and to demonstrate their advantage over the standard BEM.

2 BOUNDARY ELEMENT FORMULATION IN LAPLACE DOMAIN

In this work we study the linearized behavior of a porous medium saturated with one fluid in the open domain $\Omega \subset \mathbb{R}^3$. In the following we give a brief summary of the problem at hand. A detailed dicussion of Biot's theory [15] and a BEM formulation thereof can be found in [7].

The governing equations of the poroelastic problem are stated for the solid's displacement $\mathbf{u}$

and the fluid's pressure p and read as

$$\mathcal{L}\mathbf{u}^g(\mathbf{x}) = 0 \tag{1}$$

where $\mathbf{u}^g = (\mathbf{u}, p)^\top$ and

$$\mathcal{L} = \begin{bmatrix} \mathcal{L}^S + s^2(\rho - \beta\rho^f)\mathbf{I} & (\alpha - \beta)\nabla \\ s(\alpha - \beta)\nabla^\top & \frac{\beta}{s\rho^f}\nabla^2 + \frac{s\phi^2}{R} \end{bmatrix}. \tag{2}$$

$\mathcal{L}^S$ is the differential operator of the linear elastodynamic problem, $s \in \mathbb{C}, \Re(s) > 0$ is the Laplace parameter, β is defined as

$$\beta = \frac{\kappa\rho^f\phi^2 s}{\phi^2 + s\kappa\left(0.66\,\phi\rho^f + \phi\rho^f\right)}. \tag{3}$$

The remaining symbols are material parameters listed in Table 1.

ρ	density	2458	$[\mathrm{kg/m^3}]$
ρ^f	density of the fluid	1000	$[\mathrm{kg/m^3}]$
α	Biot parameter	0.778	$[\mathrm{N/m^2}]$
R	Biot parameter	$4.885 \cdot 10^8$	$[\mathrm{N/m^2}]$
ϕ	porosity	0.19	$[-]$
λ	Lamé parameter	0	$[\mathrm{N/m^2}]$
μ	Lamé parameter	$7.2 \cdot 10^9$	$[\mathrm{N/m^2}]$
κ	permeability	$1.9 \cdot 10^{-10}$	$[\mathrm{N/m^2}]$

Table 1: Material parameters of Berea sandstone [16]

In this paper we focus on the examination of the different BEM operators which are needed to solve a mixed boundary value problem (BVP) through the so-called direct approach. For simplicity we restrict ourselves to the Dirichlet problem, i.e.

$$\mathbf{u}^g(\mathbf{x}) = \mathbf{g}^D \quad \forall \mathbf{x} \in \Gamma, \tag{4}$$

where $\Gamma = \partial\Omega$.

A solution $\mathbf{u}^g(\mathbf{x})$ for $\mathbf{x} \in \Omega$ can be constructed by the indirect Single-Layer-Potential (SLP) approach

$$\mathbf{u}^g(\mathbf{x}) = \int_\Gamma \left(\mathbf{U}^*(\mathbf{x}, \mathbf{y})\right)^\top \mathbf{w}(\mathbf{y}) ds_\mathbf{y} \quad \forall \mathbf{x} \in \Omega, \tag{5}$$

or by the indirect Double-Layer-Potential (DLP) approach

$$\mathbf{u}^g(\mathbf{x}) = \int_\Gamma \left(\mathcal{T}_y^* \mathbf{U}^*(\mathbf{x}, \mathbf{y})\right)^\top \mathbf{v}(\mathbf{y}) ds_\mathbf{y} \quad \forall \mathbf{x} \in \Omega, \tag{6}$$

where the traction operator is defined as

$$\mathcal{T}_y^* = \begin{bmatrix} \mathcal{T}_y^S & s\,\alpha\mathbf{n}_y \\ -\beta\mathbf{n}_y^\top & \frac{\beta}{s\rho^f}\frac{\partial}{\partial\mathbf{n}_y}. \end{bmatrix} \tag{7}$$

with the traction operator $\mathcal{T}_y^S$ of elasicity. The corresponding boundary integral equation (BIE) for the unknown densities $\mathbf{w}(\mathbf{x})$, $\mathbf{v}(\mathbf{x})$ reads as

$$\mathbf{u}^g(\mathbf{x}) = \int_\Gamma \left(\mathbf{U}^*(\mathbf{x},\mathbf{y})\right)^\top \mathbf{w}(\mathbf{y}) ds_\mathbf{y} \quad \forall \mathbf{x} \in \Gamma \tag{8}$$

and

$$\mathbf{u}^g(\mathbf{x}) = \int_\Gamma \left(\mathcal{T}_y^* \mathbf{U}^*(\mathbf{x},\mathbf{y})\right)^\top \mathbf{v}(\mathbf{y}) ds_\mathbf{y} + (\sigma - 1)\mathbf{v}(\mathbf{x}) \quad \forall \mathbf{x} \in \Gamma, \tag{9}$$

where $\sigma = 0.5$ almost everywhere. The function $\mathbf{U}^*(\mathbf{x},\mathbf{y})$ is the *fundamental solution* of the differential equation, i.e. the Green's function of the adjoint operator $\mathcal{L}^*$. It is a four-by-four-tensor-valued function, where

$$\mathbf{U}^*(\mathbf{x},\mathbf{y}) = \begin{bmatrix} U^S & U^f \\ P^S & P^f \end{bmatrix} \tag{10}$$

with U^S being a three-by-three tensor, U^f a three-by-one tensor, P^s a one-by-three tensor and P^f a scalar. The explicit expressions of the four subblocks can be found in [7].

To achieve a numerical solution of the unknown densities $\mathbf{w}(\mathbf{x})$, $\mathbf{v}(\mathbf{x})$, we approximate $\mathbf{w}(\mathbf{x})$ using constant discontinuous functions φ^0 on a partition of the boundary (triangulation)

$$\mathbf{w}(\mathbf{x}) \approx \sum_{j=1}^{N} \varphi_j^0(\mathbf{x})\, \mathbf{w}_j \tag{11}$$

and $\mathbf{v}(\mathbf{x})$ using linear continous functions φ^1

$$\mathbf{v}(\mathbf{x}) \approx \sum_{j=1}^{M} \varphi_j^1(\mathbf{x})\, \mathbf{v}_j, \tag{12}$$

respectively. Note that for the vector valued problem at hand, the coefficients $\mathbf{w}_j$ and $\mathbf{v}_j$ are accordingly vector valued. The supports of the functions are restricted to either one or some neighboring triangulation elements. Hence, the boundary integral equation can be reformulated to its discretized form

$$\mathbf{u}^g(\mathbf{x}) = \sum_{j=1}^{N} \int_{supp(\varphi_j^0)} \left(\mathbf{U}^*(\mathbf{x},\mathbf{y})\right)^\top \varphi_j^0(\mathbf{y}) ds_\mathbf{y}\, \mathbf{w}_j \quad \forall \mathbf{x} \in \Gamma \tag{13}$$

and

$$\mathbf{u}^g(\mathbf{x}) = \sum_{j=1}^{M} \int_{supp(\varphi_j^1)} \left(\mathcal{T}_y^* \mathbf{U}^*(\mathbf{x},\mathbf{y})\right)^\top \varphi_j^1(\mathbf{y}) ds_\mathbf{y}\, \mathbf{v}_j + (\sigma - 1)\varphi_j^1(\mathbf{x})\mathbf{v}_j \quad \forall \mathbf{x} \in \Gamma, \tag{14}$$

respectively. This yields $N(M)$ vector-valued unknowns $\mathbf{w}_j$ ($\mathbf{v}_j$). The collocation method is employed to solve the BIE for the unknown coefficients: The equation is evaluated for N (M) $\mathbf{x}_i \in \Gamma$, yielding a set of equations for $\mathbf{w}$

$$\mathbf{u}^g(\mathbf{x}_i) = \sum_{j=1}^{N} \int_{supp(\varphi_j^0)} \left(\mathbf{U}^*(\mathbf{x}_j,\mathbf{y})\right)^\top \varphi_j^0(\mathbf{y}) ds_\mathbf{y}\, \mathbf{w}_j =: \mathbf{V}_{ij}\, \mathbf{w}_j \tag{15}$$

and $\mathbf{v}$

$$\mathbf{u}^g(\mathbf{x}_i) = \sum_{j=1}^{M} \int_{supp(\varphi_j^1)} \left(\mathcal{T}_y^* \mathbf{U}^*(\mathbf{x}_i, \mathbf{y})\right)^\top \varphi_j^1(\mathbf{y}) ds_\mathbf{y}\, \mathbf{v}_j + (\sigma - 1)\varphi_j^1(\mathbf{x}_i)\mathbf{v}_j \tag{16}$$

$$=: \left(\mathbf{K}_{ij} + (\sigma - 1)\mathbf{I}\right)\, \mathbf{v}_j, \tag{17}$$

Note that the indices i, j refer to four-by-four blocks (*block entry*) of the fundamental solution and to a subvector $\mathbf{w}_i$ of lentgh four.

The system matrix $\mathbf{V}_{ij}$ ($\mathbf{K}_{ij}$) is densely populated and hence, the standard collocation BEM scales quadratically in the number of unknowns. It can be solved using a direct solver, which scales like N^3 (M^3), or by means of an iterative solver, which involves a matrix-vector-product (MVP) in each iteration, scaling quadratically, while the number of iterations may be kept approximately constant with respect to N (M) when using a preconditioner [1].

For the measured material data of Berea sandstone [16], c.f. Table 1, the entries of the system matrix $\mathbf{V}_{ij}$ ($\mathbf{K}_{ij}$) differ by magnitudes. Numerical tests show that the direct solver employed (Eigen's `PartialPivLU` [17]) indeed yields an accurate solution of the resulting ill-conditioned system. The use of iterative solvers, on the other hand, requires a better conditioning and is mandatory for the use of fast methods like the FMM.

Numerical tests show that a simple transformation of the material data as described in [18] is sufficient to achieve that, i.e.

$$\tilde{x}_i = \frac{x_i}{A} \quad \tilde{t} = \frac{t}{B} \quad \tilde{\lambda} = \frac{\lambda}{C} \quad \tilde{\mu} = \frac{\mu}{C} \quad \tilde{\rho} = \frac{A^2}{B^2\, C}\rho \quad \tilde{\kappa} = \frac{B\, C}{A^2}\kappa \tag{18}$$

with $A = 1$, $B = 1$, $C = \mu(3\lambda + 2\mu)/(\lambda + mu)$.

Similar to iterative solvers, compression methods like the truncated Singular Value Decomposition (SVD) are likewise sensitive to such differing matrix entries, whereas the use of the transformed material data allows for its immediate application. Hence, in the following, we use the transformed material data for all calculations.

3 INTERPOLATION-BASED FM-BEM FOR POROELASTICITY

Solving (15) for $\mathbf{w}_j$ using an iterative solver involves evaluating the matrix-vector-product many times. The basic idea of the FMM is to construct a fast summation scheme by the use of a separable expansion (often denoted "degenerate expansion") of the considered function.

The FMM used in this paper is a Chebyshev interpolation-based FMM based on the *black-box FMM* by Darve [10]. In the case of the BEM the considered function is the discretized boundary integral operator $\mathbf{V}_{ij}$, or $\mathbf{K}_{ij}$, respectively. Thus, it involves the integration over the function at hand. A separable expansion of the fundamental solution allows to shift its evaluation outside the integral. Such an expansion is only possible for distant points $\mathbf{x}, \mathbf{y}$, i.e. for off-diagonal blocks of the system matrix. Therefore the standard geometrical clustering scheme of the FMM [11] is employed. For the FM-BEM we cluster over the collocation points, which is straightforward, and over the shape functions. Because of the strictly geometrical clustering the supports of some shape functions are not entirely enclosed within the cluster's boundaries. Thus the cluster's bounding box is extended such that the supports are just enclosed within.

Before we describe the different FM-operators, we briefly summarize the Chebyshev interpolation of an arbitrary function $f : \mathbb{R} \to \mathbb{C}$,

$$f(x) \approx \sum_{m=1}^{p} S_p(\bar{x}_m, x)\, f(\bar{x}_m). \tag{19}$$

The interpolation function $S_p(\bar{x}_m, x)$ is defined on the interval $[-1, 1]$ and reads as

$$S_p(\bar{x}_m, x) = \frac{1}{p} + \frac{2}{p} \sum_{n=1}^{p-1} T_n(x)\, T_n(\bar{x}_m) \quad \forall x \in [-1, 1]. \tag{20}$$

It can easily be extended to arbitrary intervals using a linear transformation. $T_n(x)$ denote the Chebyshev polynomials of order p,

$$T_n(x) = \cos(n \arccos(x)) \quad \forall x \in [-1, 1], \tag{21}$$

and $\bar{x}_m$ the corresponding roots,

$$\bar{x}_m = \cos\left(\frac{(m - \frac{1}{2})\pi}{p}\right). \tag{22}$$

For the sake of readability we do not distinct between the standard interpolation function and its transformed variant and denote both $S_p(\bar{x}_m, x)$, since it is clear from the context which one is used. This interpolation scheme can easily be extended to functions on $\mathbb{R}^3$ using

$$S_p(\bar{\mathbf{x}}_\mathbf{m}, \mathbf{x}) = \prod_{i=1}^{3} S_p(\bar{x}_{m_i}, x_i), \tag{23}$$

where $\bar{\mathbf{x}}_m = (\bar{x}_{m_1}, \bar{x}_{m_2}, \bar{x}_{m_3})$ and $\mathbf{m}$ is a multi-index. The straightforward way to approximate non-scalar functions like the tensor-valued fundamental solution $\mathbf{U}^*(\mathbf{x}, \mathbf{y})$ is to interpolate all entries using the same intepolation scheme, which is equivalent to multiplying the whole block with the scalar interpolation function S_p.

We concentrate on the differences between the standard FMM and the FM-BEM and hence omit the description of the multilevel scheme which can be found in [11]. To illustrate the method we present the idea of the Chebyshev interpolation-based FMM for one block entry of the discretized SLP operator

$$\mathbf{V}_{ij} = \int_{supp(\varphi_j^0)} \left(\mathbf{U}^*(\mathbf{x}_i, \mathbf{y})\right)^\top \varphi_j^0(\mathbf{y}) ds_\mathbf{y}\,. \tag{24}$$

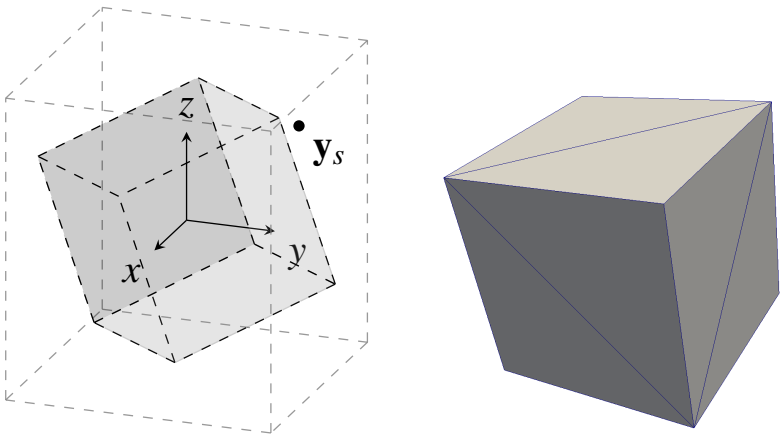
Interpolation with respect to $\mathbf{y}$ yields

$$\mathbf{V}_{ij} = \sum_\mathbf{m} \left(\mathbf{U}^*(\mathbf{x}_i, \bar{\mathbf{y}}_\mathbf{m})\right)^\top \underbrace{\int_{supp(\varphi_j^0)} S_p(\bar{\mathbf{y}}_\mathbf{m}, \mathbf{y}) \varphi_j^0(\mathbf{y}) ds_\mathbf{y}}_{\text{P2M-operator}}\,. \tag{25}$$

Further, interpolation with respect to $\mathbf{x}$ yields

$$\mathbf{V}_{ij} = \underbrace{\sum_\mathbf{n} S_p(\mathbf{x}_i, \bar{\mathbf{x}}_\mathbf{m})}_{\text{L2P-operator}} \underbrace{\sum_\mathbf{m} \left(\mathbf{U}^*(\bar{\mathbf{x}}_\mathbf{n}, \bar{\mathbf{y}}_\mathbf{m})\right)^\top}_{\text{M2L-operator}} \int_{supp(\varphi_j^0)} S_p(\bar{\mathbf{y}}_\mathbf{m}, \mathbf{y}) \varphi_j^0(\mathbf{y}) ds_\mathbf{y}\,. \tag{26}$$

The points $\bar{\mathbf{x}}_\mathbf{n}, \bar{\mathbf{y}}_\mathbf{m} \in \mathbb{R}^3$ are the three-dimensional Chebyshev nodes transformed to the bounding box. Obviously, using such an interpolation, the number of kernel evaluations can be reduced if the submatrix associated with the corresponding clusters of $\mathbf{x}_i$ and φ_j^0 is larger than



(a) domain and source point (b) coarsest mesh

Figure 1: Computation domain.

the associated M2L-operator of size $p^3 \times p^3$. The reduction is actually significantly larger, since for the numerical integration in the standard BEM (15) the kernel is evaluated for each quadrature point. Using the FMM this integration arises in the P2M-operator, which involves only the numerical integration of a polynomial. We emphasize that the matrix approximation is implemented in terms of MVPs, and that an aforementioned multi-level scheme is used.

The traction fundamental solution explicitly depends on the normal vector of an associated surface, i.e. on the normal vector of the triangulation element in the case of the discretized boundary integral operator. Hence a direct interpolation as presented for the SLP is not possible. Recalling (16) the DLP operator is defined as

$$\mathbf{K}_{ij} = \int_{supp(\varphi_j^1)} \left(\mathcal{T}_y^* \mathbf{U}^*(\mathbf{x}_i, \mathbf{y})\right)^\top \varphi_j^1(\mathbf{y}) ds_\mathbf{y} \tag{27}$$

and can thus be interpolated using the interpolation of $\mathbf{U}^*$

$$\mathbf{K}_{ij} = \sum_\mathbf{m} \left(\mathbf{U}^*(\mathbf{x}_i, \bar{\mathbf{y}}_\mathbf{m})\right)^\top \int_{supp(\varphi_j^1)} \left(\mathcal{T}_y^{*\top} S_p(\bar{\mathbf{y}}_\mathbf{m}, \mathbf{y})\right) \varphi_j^1(\mathbf{y}) ds_\mathbf{y}. \tag{28}$$

The traction operator $\mathcal{T}_y^*$ can be shifted onto the interpolation functions, which leads to a tensorial interpolation operator, whereas all other operations are identical to the SLP. Hence, the calculation of the far field of the DLP-operator only involves the evalution of the fundametal solution $\mathbf{U}^*$ itself.

4 NUMERICAL EXAMPLES

To study the FMM for the poroelastic problem we seek a numerical solution of a known solution to the homogenous differential equation. For a source point $\mathbf{y}_s$ outside the domain, the individual rows of the fundamental solution $\mathbf{U}^*(\mathbf{x}, \mathbf{y}_s)$ fulfill (1) and thus we use a linear combination of all rows as boundary condition. To study a general case, the calculation domain is a unit cube which is rotated along the axes of the coordination system, c.f. Figure 1.

We create a sequence of triangulations by global uniform refinement starting at the coarsest discretization using 12 elements as depicted in Figure 1, denoted Level 0.

All computations are performed using our in-house developed C++ library HyENA. The dense computations are solved using Eigen's `PartialPivLU` [17], the FMM computations

using AHMED's `FGMRes` [19]. Further, the M2L-operators are compressed using the Singular Value Decomposition of LAPACK, and OpenMP is used for parallelization. All computation parameters are listed in Table 2.

The mean error of the computation is determined using (5). It is evaluated for a set of 27 points in the interior of the box $[-\mathbf{0.01}, \mathbf{0.01}] \subset \mathbb{R}^3$ in the center of the computation domain. We compare the inner error of the FM-BEM to the dense calclulation to find the required minimal interpolation order, which turns out to be $p = 4$ for Level 3. The errors are listed in Tables 3 (SLP) and 4 (DLP). We note that for the finest mesh (Level 6), where a comparision with a dense computation is not possible, the order of convergence (eoc) of the SLP approach deviates from its assumed value of 2. The convergence of the coarser levels exceeds that value, which might explain this deviation.

	SLP					DLP				
Mesh level	N	L	p	ϵ_{SVD}	N_{it}	M	L	p	ϵ_{SVD}	N_{it}
3	3072	2	4	10^{-4}	30	1544	2	4	10^{-4}	15
4	12288	3	5	10^{-5}	38	6152	3	5	10^{-5}	18
5	49152	4	6	10^{-6}	48	24584	4	6	10^{-6}	19
6	196608	5	7	10^{-7}	61	98312	5	7	10^{-7}	20

Table 2: Indirect SLP, DLP: computation parameters

	displacement				pressure			
Mesh level	ϵ_{dense}	eoc	ϵ_{FM}	eoc	ϵ_{dense}	eoc	ϵ_{FM}	eoc
3	2.306E-4		2.358E-4		4.388E-4		4.350E-4	
4	4.018E-5	2.52	3.957E-5	2.57	7.904E-5	2.47	7.859E-5	2.47
5	7.363E-6	2.45	7.126E-6	2.47	1.458E-5	2.44	1.405E-5	2.48
6	-	-	3.193E-6	1.16	-	-	5.063E-6	1.47

Table 3: Indirect SLP: mean relative inner error

	displacement				pressure			
Mesh level	ϵ_{dense}	eoc	ϵ_{FM}	eoc	ϵ_{dense}	eoc	ϵ_{FM}	eoc
3	9.224E-4		9.487E-4		4.333E-4		4.417E-4	
4	2.311E-4	2.00	2.376E-4	2.00	9.981E-5	2.12	1.038E-4	2.09
5	6.270E-5	1.88	6.270E-5	1.92	2.228E-5	2.16	2.387E-5	2.12
6	-	-	1.495E-5	2.07	-	-	1.113E-5	1.10

Table 4: Indirect DLP: mean relative inner error

Further, we analyze the memory requirement of the FM-BEM. It is calculated by the total size of all FMM operators. Comparision with the dense operators yields the compression of the FM-BEM. Table 5 lists the memory requirement, its scaling with respect to refinement and the compression.

The classical FMM scales linearly in N/M for constant interpolation order p. To preserve the convergence of the inner error the interpolation order p needs to be increased for each refinement

	SLP					DLP				
Mesh level	M_{dense}	scal.	M_{FM}	scal.	compr.	M_{dense}	scal.	M_{FM}	scal.	compr.
3	0.15		0.17		115%	0.04		0.07		186%
4	2.42	2.00	1.22	1.41	51%	0.61	1.99	1.12	1.99	185%
5	38.65	2.00	4.65	0.96	12%	9.67	2.00	4.64	1.02	48%
6	618.48	2.00	15.25	0.86	2%	154.64	2.00	14.13	0.80	9%

Table 5: Indirect SLP, DLP: memory requirement M[GB] and compression

and, hence, this scaling cannot be achieved. Still, the memory consumption of the FM-BEM scales better than the dense computation and is close to 1. The measured scalings below 1 arise from the compression of the M2L-operators using SVD. The indirect SLP collocation FM-BEM requires less memory than the dense case from refinement level 4 upwards, with a compression of 51% and better. The indirect DLP collocation FM-BEM achieves a compression below 1 for levels 5 and higher. The validity of the FM-BEM is confirmed by the convergence of the inner error.

5 CONCLUSION

The presented FM-BEM circumvents three major problems of the standard BEM for poroelasticity: First, using the Chebyshev interpolation-based FMM, the number of evaluations of the computationally expensive fundamental solution $\mathbf{U}^*(\mathbf{x}, \mathbf{y})$ is significantly reduced. Second, for large problems, the storage requirement of the FM-BEM operators is much smaller than for the dense operators. A significant compression, though, was only achieved using the SVD to compress the M2L-operators.

Third, the different physical natures of the unknowns lead to system matrices whose entries vary over several orders of magnitude. The Chebyshev interpolation is insensitive to such varying magnitudes, and numerical tests show that even for the untransformed material, this most simple interpolation yields similar relative errors for all four matrix blocks. The use of an iterative solver, however, requires the use of the transformed material data as described in Section 2. Finally, the accuracy of the presented FM-BEM is mirrored by the convergence of the mean inner error.

REFERENCES

[1] O. Steinbach, *Numerical Approximation Methods for Elliptic Boundary Value Problems*. Springer New York, 2008.

[2] C. Lubich, Convolution quadrature and discretized operational calculus I. *Numer, Math.*, **52**, 129-145, 1988.

[3] C. Lubich, Convolution quadrature and discretized operational calculus II. *Numer, Math.*, **52**, 413-425, 1988.

[4] M. Schanz, Poroelastodynamics: Linear Models, Analytical Solutions, and Numerical Methods. *App. Mech. Rev.*, **62**, 030803-1 - 030803-15, 2009.

[5] A. H.-D. Cheng, T. Badmus, and D. Beskos, Integral Equations for Dynamic Poroelasticity in Frequency Domain With BEM Solution. *J. Eng. Mech.*, **117**, 1136-1157, 1991.

[6] J. Domínguez, Boundary Element Approach for Dynamic Poroelastic Problems. *Int. J. Numer. Methods Eng.*, **35**, 307-324, 1992.

[7] M. Schanz, *Wave Propagation in Viscoelastic and Poroelastic Continua. A Boundary Element Approach*, volume 2 of *Lecture notes in applied mechanics*. Springer, 2001.

[8] M. Messner and M. Schanz, A Symmetric Galerkin Boundary Element Method for 3d Linear Poroelasticity. *Acta Mech.*, **223**, 1751-1768, 2012.

[9] M. Bebendorf, Approximation of boundary element matrices. *Numer. Math.*, **86**, 565-589, 2000.

[10] W. Fong and E. Darve, The black-box fast multipole method. *J. Comp. Phys.*, **228**, 8712-8725, 2009.

[11] L. Greengard and V. Rokhlin, A fast algorithm for particle simulations. *J. Comp. Phys.*, **73**, 325-348, 1987.

[12] N. Nishimura, Fast multipole accelerated boundary integral equation methods. *Appl. Mech. Rev.*, **55**, 299-324, 2002.

[13] H. Fujiwara, The fast multipole method for solving integral equations of three-dimensional topography and basin problems, *Geophys. J. Int.* **140**, 198-210, 2000.

[14] S. Chaillat, M. Bonnet and J.-F. Semblat, A multi-level fast multipole BEM for 3-d elastodynamics in the frequency domain. *Comput.Methods in Appl.Mech.Eng.* **197** 4233-4249, 2008.

[15] M. A. Biot, Theory of propagation of elastic waves in a fluid-saturated porous solid I: Low-frequency range. *J. Acoust. Soc. Am.*, **28**, 168-178, 1956.

[16] Y. K. Kim and H. B. Kingsbury, Dynamic Characterization of Poroelastic Materials. *Exp. Mech.*, **19**, 252-258, 1979.

[17] G. Guennebaud and B. Jacob, Eigen v3. http://eigen.tuxfamily.org, 2010.

[18] M. Schanz and L. Kielhorn, Dimensionless Variables in a Poroelastodynamic Time Domain. *Building Research Journal* **53**, 175-189, 2005.

[19] M. Bebendorf. Another software library on hierarchical matrices for elliptic differential equations (AHMED). http://bebendorf.ins.uni-bonn.de/AHMED.html, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ENERGY STABLE HIGH ORDER FINITE DIFFERENCE METHODS ON STAGGERED GRIDS: AN INITIAL INVESTIGATION

Ossian O'Reilly[1,2], Tomas Lundquist[2], and Jan Nordström[2]

[1]Department of Geophysics
Stanford University, CA 94305-2215
e-mail: ooreilly@stanford.edu

[2] Department of Mathematics, Division of Computational Mathematics
Linköping University, SE-581 83 Linköping, Sweden
e-mail: {tomas.lundquist,jan.nordstrom}@liu.se

Keywords: Staggered Grids, High Order Finite Difference Methods, Summation-By-Parts, Weak Boundary Conditions , Energy stability, Wave Equations.

Abstract. *We consider wave equations in first-order form and derive provably stable, high order finite difference operators on staggered grids. This is the first time that stability has been proven for initial boundary value problems for wave equations on staggered grids. The staggered grid operators are in summation-by-parts form and when combined with weak boundary conditions, lead to an energy stable scheme. Numerical computations for the two dimensional acoustic wave equation in Cartesian geometries corroborate the theoretical developments.*

1 INTRODUCTION

High order finite difference methods are highly effective for wave propagation over long distances. In particular for electromagnetic and elastic wave propagation, staggered spatial schemes perform very well [5]. Arguably the most significant challenge in a staggered grid finite difference formulation is the implementation of boundary and interface conditions at curved boundaries and material discontinuities [8]. The curved boundary is often approximated using a stair-cased grid approximation, which results in at most first-order accuracy. An alternative strategy is to conform to the boundary geometry using a curvilinear coordinate transformation. However, unless the resulting spatial discretization is energy conserving, instabilities can develop.

The curvilinear coordinate transformation technique works exceptionally well for high order finite difference methods on nodal grids that combine summation-by-parts (SBP) operators [3, 6, 11] and weak boundary conditions using the simultaneous approximation term (SAT) method [2]. The SBP-SAT approach leads to a provably stable multiblock scheme, even for highly skewed grids (provided that the Jacobian of the coordinate transformation is not ill-conditioned). So far, the SBP-SAT approach has only been applied on nodal grids. Thus, the motivation for this work is to develop a provably stable curvilinear multiblock scheme on staggered grids. As a first step toward this goal, we formulate the SBP-SAT approach on staggered grids in Cartesian geometries and construct corresponding SBP operators of the first derivative. To the best of our knowledge, this is the first time that stability has been proven for a high order finite difference method on a staggered grid for initial boundary value problems for wave equations.

2 DEFINITIONS

We begin by discretizing the interval $[\alpha\ \beta] \in \mathbb{R}$ using a nodal grid $\mathbf{x}_+$ and a cell-centered grid $\mathbf{x}_-$ (see e.g, [7]) . The grids are given by $\mathbf{x}_+ = (x_0,\ x_1,\ \ldots, x_N)^T \in \mathbb{R}^{N+1}$ and $\mathbf{x}_- = \left(x_0,\ x_{1/2},\ x_{3/2},\ \ldots,\ x_{N-1/2},\ x_N\right)^T \in \mathbb{R}^{N+2}$, where

$$x_j = \alpha + j\Delta x,\ 0 \leq j \leq N,\ x_{j-1/2} = \alpha + (j-1/2)\Delta x,\ 1 \leq j \leq N,$$

and $\Delta x = (\beta - \alpha)/N$ is the grid spacing (Figure 1). Note that both grids share the same boundary points.

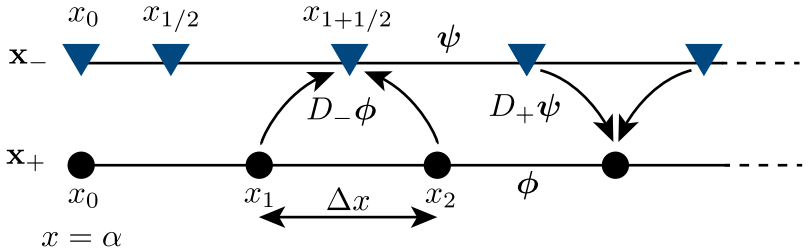


Figure 1: Nodal grid $\mathbf{x}_+$ and cell-centered grid $\mathbf{x}_-$. Both grids share the same boundary points. The difference operator D_- acts on a grid function ϕ defined on the nodal grid $\mathbf{x}_+$ and stores the result on the cell-centered grid $\mathbf{x}_-$, D_+ acts on a grid function ψ defined on the cell-centered grid $\mathbf{x}_-$ and stores the result on the nodal grid $\mathbf{x}_+$.

SBP staggered grid operators employ a high order, central finite difference approximation at each interior grid point, and one-sided difference approximations near the boundary. For

example, 2nd-order accuracy in the interior is given by the standard central approximation

$$\left.\frac{d\psi}{dx}\right|_{x_i} \approx \frac{\psi_{i+1/2}-\psi_{i-1/2}}{\Delta x},\ \left.\frac{d\phi}{dx}\right|_{x_{i-1/2}} \approx \frac{\phi_i-\phi_{i-1}}{\Delta x},$$

for smooth grid functions $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ defined on $\mathbf{x}_-$ and $\mathbf{x}_+$, respectively. The SBP staggered grid operators are defined by

$$D_+ = P_+^{-1}Q_+ \in \mathbb{R}^{(N+1)\times(N+2)},\quad D_- = P_-^{-1}Q_- \in \mathbb{R}^{(N+2)\times(N+1)},$$

where $P_+ \in \mathbb{R}^{(N+1)\times(N+1)}$, $Q_+ \in \mathbb{R}^{(N+1)\times(N+2)}$, $P_- \in \mathbb{R}^{(N+2)\times(N+2)}$, $Q_- \in \mathbb{R}^{(N+2)\times(N+1)}$. Here, D_+ acts on grid functions defined on $\mathbf{x}_-$ and D_- acts on grid functions defined on $\mathbf{x}_+$. The way that D_+, D_- acts on $\mathbf{x}_-$, $\mathbf{x}_+$ and store the result is schematically shown in Figure 1, for the 2nd-order accurate case. For an order of accuracy q we have the relations

$$D_+\boldsymbol{\psi}^k = k\boldsymbol{\phi}^{k-1},\ D_-\boldsymbol{\phi}^k = k\boldsymbol{\psi}^{k-1},\ k=0,1,\ldots,q,$$

where $\boldsymbol{\phi}^k$ denotes the kth-monomial (each component of the vector is raised to the kth-power). In the interior the accuracy is $q=2s$ and for points near the boundary it is $q=s$. In addition, the SBP staggered grid operators satisfy:

(i) P_+ and P_- are symmetric and positive definite matrices defining the respective discrete L_2 norms

$$\|\boldsymbol{\phi}\|_{h+}^2 = \boldsymbol{\phi}^T P_+ \boldsymbol{\phi} \approx \int_\alpha^\beta \phi^2 dx,\ \|\boldsymbol{\psi}\|_{h-}^2 = \boldsymbol{\psi}^T P_- \boldsymbol{\psi} \approx \int_\alpha^\beta \psi^2 dx.$$

(ii) The matrices Q_+ and Q_- satisfy the SBP property

$$Q_+ + Q_-^T = B_+ = B_-^T = \mathbf{e}_{N+}\mathbf{e}_{N-}^T - \mathbf{e}_{0+}\mathbf{e}_{0-}^T \in \mathbb{R}^{(N+1)\times(N+2)}. \tag{1}$$

In (1), the vectors $\mathbf{e}_{0\pm}$,$\mathbf{e}_{N\pm}$ are of appropriate sizes, and gathers the value at each boundary point, i.e.,

$$\mathbf{e}_{0+}^T\mathbf{x}_+ = \mathbf{e}_{0-}^T\mathbf{x}_- = \alpha,\ \mathbf{e}_{N+}^T\mathbf{x}_+ = \mathbf{e}_{N-}^T\mathbf{x}_- = \beta. \tag{2}$$

For example, a pair of 2nd-order SBP staggered grid operators are given by

$$Q_+ = \begin{pmatrix} -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & & \\ -\frac{1}{2} & -\frac{1}{4} & \frac{3}{4} & & \\ & & -1 & 1 & \\ & & & \ddots & \ddots \end{pmatrix},\ Q_- = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} & & & \\ -\frac{1}{4} & \frac{1}{4} & & & \\ -\frac{1}{4} & -\frac{3}{4} & & & \\ & & -1 & 1 & \\ & & & \ddots & \ddots \end{pmatrix}$$

and

$$P_+ = \text{diag}\left(\frac{1}{2},\ 1, 1, \ldots\right)\Delta x,\ P_- = \text{diag}\left(\frac{1}{2},\ \frac{1}{4},\ \frac{5}{4},\ 1, \ldots\right)\Delta x.$$

3 ANALYSIS

As a prototype problem, we consider the scalar wave equation in one dimension

$$u_t + Au_x = 0, \quad \alpha \leq x \leq \beta,\ t \geq 0, \quad u = \begin{pmatrix} p \\ v \end{pmatrix},\ A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3}$$

subject to appropriate initial and boundary conditions. We apply the energy method [1] to (3), which leads to

$$\frac{d}{dt}\|u\|^2 = -u^T Au\big|_\alpha^\beta = -2pv\big|_\alpha^\beta, \tag{4}$$

where $\|u\| = \int_\alpha^\beta u^T u dx$. By considering the boundary at $x = \alpha$ only and diagonalizing

$$A = X\Lambda X^T, X = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix},\ \Lambda = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \tag{5}$$

we obtain

$$\frac{d}{dt}\|u\|^2 = w^T\Lambda w\big|_{x=\alpha} = (w^+)^2\big|_{x=\alpha} - (w^-)^2\big|_{x=\alpha},$$

where $w = (w^-, w^+)$, $w^\pm = (X^\pm)^T\mathbf{u}$ are the characteristic variables, and $\Lambda = \text{diag}(\Lambda^-, \Lambda^+)$ contain the positive and negative eigenvalues and the corresponding eigenvectors $X = (X^-, X^+)$, respectively. The linear homogeneous boundary condition

$$w^+\big|_{x=\alpha} - rw^-\big|_{x=\alpha} = 0 \tag{6}$$

leads to

$$\frac{d}{dt}\|u\|^2 = w^T\Lambda w = (r^2-1)(w^-)^2\big|_{x=\alpha}, \tag{7}$$

and we have an estimate for $|r| \leq 1$.

3.1 The semi-discrete approximation

An SBP-SAT staggered grid approximation of (3) is given by

$$\mathbf{u}_t + \tilde{A}\mathbf{u} = P^{-1}E_0\Sigma L E_0^T\mathbf{u}, \tag{8}$$

where

$$\mathbf{u} = \begin{pmatrix} \mathbf{p} \\ \mathbf{v} \end{pmatrix},\ \tilde{A} = \begin{pmatrix} 0 & D_- \\ D_+ & 0 \end{pmatrix},\ E_0 = \begin{pmatrix} \mathbf{e}_{0-} & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_{0+} \end{pmatrix},\ P = \begin{pmatrix} P_- & 0 \\ 0 & P_+ \end{pmatrix},$$

and $\mathbf{p} = \big(p_0,\ p_{1/2},\ \ldots,\ p_{N-1/2},\ p_N\big)^T$ is stored on $\mathbf{x}_-$, $\mathbf{v} = (v_0,\ v_1,\ \ldots,\ v_N)^T$ is stored on $\mathbf{x}_+$. The boundary condition is weakly imposed at $x = \alpha$ with the aid of E_0^T that gathers the solution on the boundary, i.e. $\mathbf{u}_0 = E_0^T\mathbf{u} = (p_0,\ v_0)$. The boundary operator $L \in \mathbb{R}^{1\times 2}$ is given by (6), that is

$$L\mathbf{u}_0 = \big((X^+)^T - r(X^-)^T\big)\,\mathbf{u}_0 = w_0^+ - rw_0^-. \tag{9}$$

The penalty matrix $\Sigma \in \mathbb{R}^{2\times 1}$ is determined by applying the discrete energy method (see [12] for a general outline of the procedure). We have

$$\begin{aligned}\frac{d}{dt}\|\mathbf{u}\|_P^2 &= -2\mathbf{p}^T Q_- \mathbf{v} - 2\mathbf{v}^T Q_+ \mathbf{p} + \mathbf{u}_0^T(\Sigma L + L^T\Sigma^T)\mathbf{u}_0 \\ &= -2\mathbf{p}^T\left(Q_+^T + Q_-\right)\mathbf{v} + \mathbf{u}_0^T(\Sigma L + L^T\Sigma^T)\mathbf{u}_0 \\ &= -2\left(p_N v_N - p_0 v_0\right) + \mathbf{u}_0^T(\Sigma L + L^T\Sigma^T)\mathbf{u}_0, \end{aligned} \tag{10}$$

where $\|\mathbf{u}\|_P^2 = \mathbf{u}^T P \mathbf{u}$. Note that the SBP property (1) was used in the last step. We consider only the boundary $x = \alpha$ and write (10) in quadratic form

$$\frac{d}{dt}\|\mathbf{u}\|_P^2 = \mathbf{u}_0^T A \mathbf{u}_0 + \mathbf{u}_0^T(\Sigma L + L^T\Sigma^T)\mathbf{u}_0.$$

By using the diagonalization $A = X\Lambda X^T$ given in (5) and applying (9), we find

$$\frac{d}{dt}\|\mathbf{u}\|_P^2 = \mathbf{w}_0^T\Lambda\mathbf{w}_0 + \mathbf{w}_0^T X^T\Sigma(w_0^+ - r w_0^-) + (w_0^+ - r w_0^-)\Sigma^T X\mathbf{w}_0, \tag{11}$$

where $\mathbf{w}_0 = (w_0^-,\ w_0^+) = X^T\mathbf{u}_0$. Finally, the choice

$$\Sigma = -X^+\Lambda^+ \tag{12}$$

leads to

$$\frac{d}{dt}\|\mathbf{u}\|_P^2 = (r^2-1)(w_0^-)^2 - (w_0^+ - r w_0^-)^2 \leq 0, \quad |r| \leq 1.$$

The result is similar to (7), but also includes numerical dissipation that vanishes with grid refinement.

Remark 1. *The penalty matrix and boundary operator in (8) are given by*

$$\Sigma = -\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix},\ L = \frac{1}{\sqrt{2}}\begin{pmatrix} 1+r \\ 1-r \end{pmatrix}^T,$$

where (5), (9), and (12) have been used.

4 THE TWO-DIMENSIONAL WAVE EQUATION

We now consider the wave equation in two dimensions

$$Mu_t + Au_x + Bu_y = 0, \quad (x,y) \in \Omega,\ t \geq 0, \tag{13}$$

where

$$u = \begin{pmatrix} p \\ v_1 \\ v_2 \end{pmatrix},\ M = \begin{pmatrix} K^{-1} & 0 & 0 \\ 0 & \rho & 0 \\ 0 & 0 & \rho \end{pmatrix},\ A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{14}$$

The system (13) describes the propagation of acoustic waves with pressure $p = p(x,y,t)$, and velocity vector $v = v(x,y,t) = (v_1,\ v_2)$ in a medium characterized by the fluid density

$\rho(x,y) > 0$ and the bulk modulus $K(x,y) > 0$. The same set of equations also describes electromagnetic transverse wave propagation [9]. The energy method applied to (13) leads to

$$\frac{1}{2}\frac{d}{dt}\|u\|_M^2 = -\oint_{\partial\Omega} pv \cdot \hat{n} ds, \tag{15}$$

where $\|u\|_M^2 = \int_\Omega u^T M u dx dy$, $\hat{n}$ is the outward-pointing unit normal associated with the boundary $\partial\Omega$, and ds is the infinitesimal line element. The number and form of the boundary conditions can be determined by following the procedure in section 3. For instance, the boundary condition

$$p - Z\left(v \cdot \hat{n}\right) = r\left[p + Z\left(v \cdot \hat{n}\right)\right], \quad (x,y) \in \partial\Omega, \tag{16}$$

where $Z = \sqrt{\rho K}$, results in an estimate if $|r| \leq 1$.

4.1 The semi-discrete approximation

An SBP-SAT staggered grid approximation of (13) is given by

$$\tilde{M}\mathbf{u}_t + \tilde{A}\mathbf{u} + \tilde{B}\mathbf{u} = \mathbf{BT}_T + \mathbf{BT}_B + \mathbf{BT}_L + \mathbf{BT}_R. \tag{17}$$

In (17), we have introduced

$$\mathbf{u} = \begin{pmatrix} \mathbf{p} \\ \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix}, \ \tilde{M} = \begin{pmatrix} K^{-1} & 0 & 0 \\ 0 & \rho_1 & 0 \\ 0 & 0 & \rho_2 \end{pmatrix},$$

$$\tilde{A} = \begin{pmatrix} 0 & D_{x-} \otimes I_{y-} & 0 \\ D_{x+} \otimes I_{y-} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ \tilde{B} = \begin{pmatrix} 0 & 0 & I_{x-} \otimes D_{y-} \\ 0 & 0 & 0 \\ I_{x-} \otimes D_{y+} & 0 & 0 \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product. Here we have chosen to store p at the center of the grid cells (Figure 2). The diagonal matrices K^{-1}, ρ_1, ρ_2 hold the material properties and coincide with the location of p, v_1, v_2, respectively. Furthermore, $I_{x\pm}$ and $I_{y\pm}$ are identity matrices of appropriate sizes.

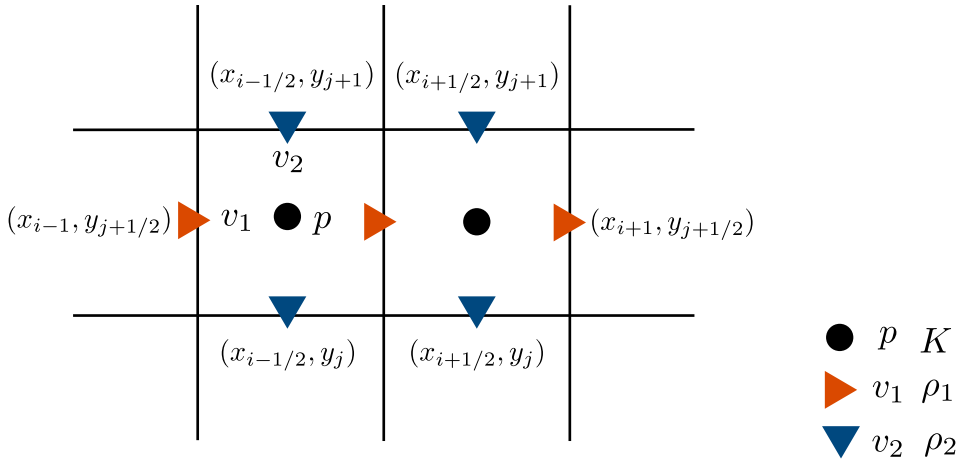


Figure 2: Staggered grid layout for the acoustic wave equation (13).

The boundary condition (16) on the top boundary is weakly imposed by

$$\begin{aligned} \mathbf{BT}_T = P^{-1} E_T P_T \Sigma_T L_T E_T^T \mathbf{u}, \qquad P = \begin{pmatrix} P_{x-} \otimes P_{y-} & 0 & 0 \\ 0 & P_{x+} \otimes P_{y-} & 0 \\ 0 & 0 & P_{x-} \otimes P_{y+} \end{pmatrix}, \\ P_T = \begin{pmatrix} P_{x-} & 0 & 0 \\ 0 & P_{x+} & 0 \\ 0 & 0 & P_{x-} \end{pmatrix}, \quad E_T = \begin{pmatrix} I_{x-} \otimes \mathbf{e}_{y_N-} & 0 & 0 \\ 0 & I_{x+} \otimes \mathbf{e}_{y_N-} & 0 \\ 0 & 0 & I_{x-} \otimes \mathbf{e}_{y_N+} \end{pmatrix}. \end{aligned} \tag{18}$$

The penalty matrix and boundary operator related to (18) are given by

$$\Sigma = \frac{1}{2} \begin{pmatrix} -Z^{-1} \\ \hat{n}_1 \\ \hat{n}_2 \end{pmatrix}, \; L(r) = \begin{pmatrix} 1-r \\ -Z\hat{n}_1 \left(1+r\right) \\ -Z\hat{n}_2 \left(1+r\right) \end{pmatrix}^T, \tag{19}$$

where Σ_T and L_T in (18), are obtained by inserting the unit normal on the boundary $\hat{n} = (0, \; 1)^T$ into (19).

5 NUMERICAL EXPERIMENTS

We test the accuracy of the SBP staggered scheme by using a manufactured solution. The SBP staggered scheme are compared against SBP nodal schemes of the same order for wave propagation over long times. Finally, we demonstrate that the SBP staggered scheme can accurately propagate waves generated by a singular source term acting on the boundary.

In each numerical experiment there is a fundamental wavelength λ that must be resolved. We use λ and the wave speed $c = \sqrt{K/\rho}$ to nondimensionalize (13), assuming constant material properties ($\rho = $ const., $K = $ const.). The dimensionless temporal and spatial scales are given by

$$\bar{t} = kct, \; \bar{x} = kx, \; \bar{y} = ky,$$

where $k = 2\pi/\lambda$. We also nondimensionalize pressure and velocity using

$$\bar{p} = \frac{p}{\rho c V}, \; \bar{v}_1 = \frac{v_1}{V}, \; \bar{v}_2 = \frac{v_2}{V},$$

where V is a reference amplitude. The wave equation (13) in dimensionless form becomes

$$\bar{u}_{\bar{t}} + \bar{A}\bar{u}_{\bar{x}} + \bar{B}\bar{u}_{\bar{y}} = 0, \; \bar{u} \; = \begin{pmatrix} \bar{p} \\ \bar{v}_1 \\ \bar{v}_2 \end{pmatrix}, \; \bar{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; \bar{B} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{20}$$

Throughout the remainder of the paper we will exclusively use the nondimensional form (20). For notational convenience we will drop the superscript denoting nondimensional variables. In the experiments, we will use the same grid spacing in each coordinate direction $\Delta x = \Delta y$, and advance in time using a 4th-order Runge-Kutta scheme and a sufficiently small time step to make the temporal errors negligible. The error will be measured as $\|\mathbf{u}^{(\Delta x_1)} - \mathbf{u}^*\|_h^2$ where $\mathbf{u}^{(\Delta x)}$ is the numerical solution obtained using the grid spacing Δx and $\mathbf{u}^*$ is computed from an analytic solution, sampled at the grid points.

5.1 Accuracy

To assess the accuracy of the SBP staggered grid operators, we introduce the manufactured solution

$$\begin{aligned} p(x,y,t) &= \sin(x+y+\sqrt{2}t), \\ v_1(x,y,t) &= -\frac{1}{\sqrt{2}}\sin(x+y+\sqrt{2}t), \\ v_2(x,y,t) &= -\frac{1}{\sqrt{2}}\sin(x+y+\sqrt{2}t) \end{aligned} \tag{21}$$

The computational domain is the square $\Omega = (0,\ 2\pi m) \times (0,\ 2\pi m)$, where $m = 10$ (number of wavelengths). The numerical boundary condition is imposed by setting $r = 0$ in (16) and using the manufactured solution (21) as boundary data. The numerical solution is advanced in time until $t = 2\pi m$ and is resolved using $2\pi/\Delta x = 6.4$ grid points per wavelength on the coarsest grid. For numerical schemes constructed using nodal SBP operators the convergence rate is expected to be $s+1$, where $2s$ is the interior accuracy [10]. Figure 3 shows that the SBP staggered scheme converges at this rate or higher.

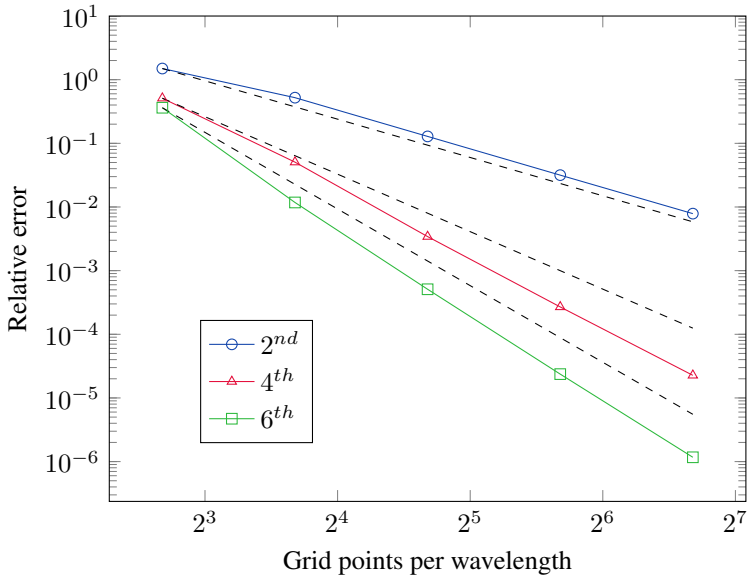


Figure 3: Error in the numerical solution computed using SBP staggered grid operators for the test problem (21). The dashed lines show reference convergence rates $s = 2, 3, 4$, respectively.

5.2 Dispersion error

To assess the dispersion error of the SBP staggered grid operators, we introduce the manufactured solution

$$\begin{aligned} p(x,y,t) &= \sin(x)\sin(y)\cos(\sqrt{2}t), \\ v_1(x,y,t) &= -\frac{1}{\sqrt{2}}\cos(x)\sin(y)\sin(\sqrt{2}t), \\ v_2(x,y,t) &= -\frac{1}{\sqrt{2}}\sin(x)\cos(y)\sin(\sqrt{2}t), \end{aligned} \tag{22}$$

defined on the computational domain $\Omega = (0,\ 2\pi m) \times (0,\ 2\pi m)$, where $m = 16$ (the number of wavelengths). The manufactured solution (22) satisfies the boundary condition $p = 0$ on all of the boundaries. The numerical boundary condition is imposed by setting $r = -1$ in (19). For this boundary condition the continuous energy is conserved. Therefore solution is trapped inside the computational domain, causing the error to grow in time. This test will also confirm that the implementation is stable by integrating over long times.

We compare the performance of the 6th-order SBP staggered and nodal grid operators using $2\pi/\Delta x \approx 16$ grid points per wavelength, and advance until the final time $t \approx 10^4$, which requires nearly 5×10^4 time steps. The relative error growth in time of the SBP staggered and nodal schemes is shown in Figure 4. In a practical calculation, it is useful to define an error threshold that denotes the maximum relative error that can be tolerated. If the maximum error is $\epsilon = 0.05$, then the SBP staggered and nodal schemes reach this threshold after about $t = 4775$ and $t = 616$, respectively. This test clearly demonstrates the superior accuracy of the SBP staggered grid operators.

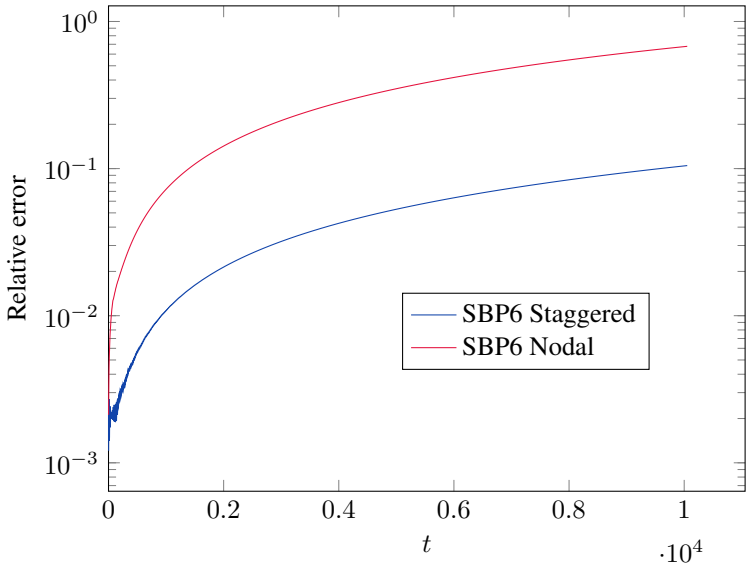


Figure 4: Error growth in time for the test problem (22) using 16 grid points per wavelength.

5.3 An additional capability

A staggered grid difference method offers the ability to discretize a singular source term without introducing any spurious oscillations. We will confirm that the new SBP staggered grid operators also possess this property.

On the top boundary, we specify the boundary condition $p = g(t)\delta(x - x_*)$, where $g(t)$ is the Ricker wavelet function

$$g(t) = (1 - 2\pi^2 f_0^2 (t - t_0)^2) e^{-\pi^2 f_0^2 (t - t_0)^2}, \tag{23}$$

and $\delta(x)$ is the Dirac delta function. The Dirac delta function is defined by

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_*)dx = f(x_*), \tag{24}$$

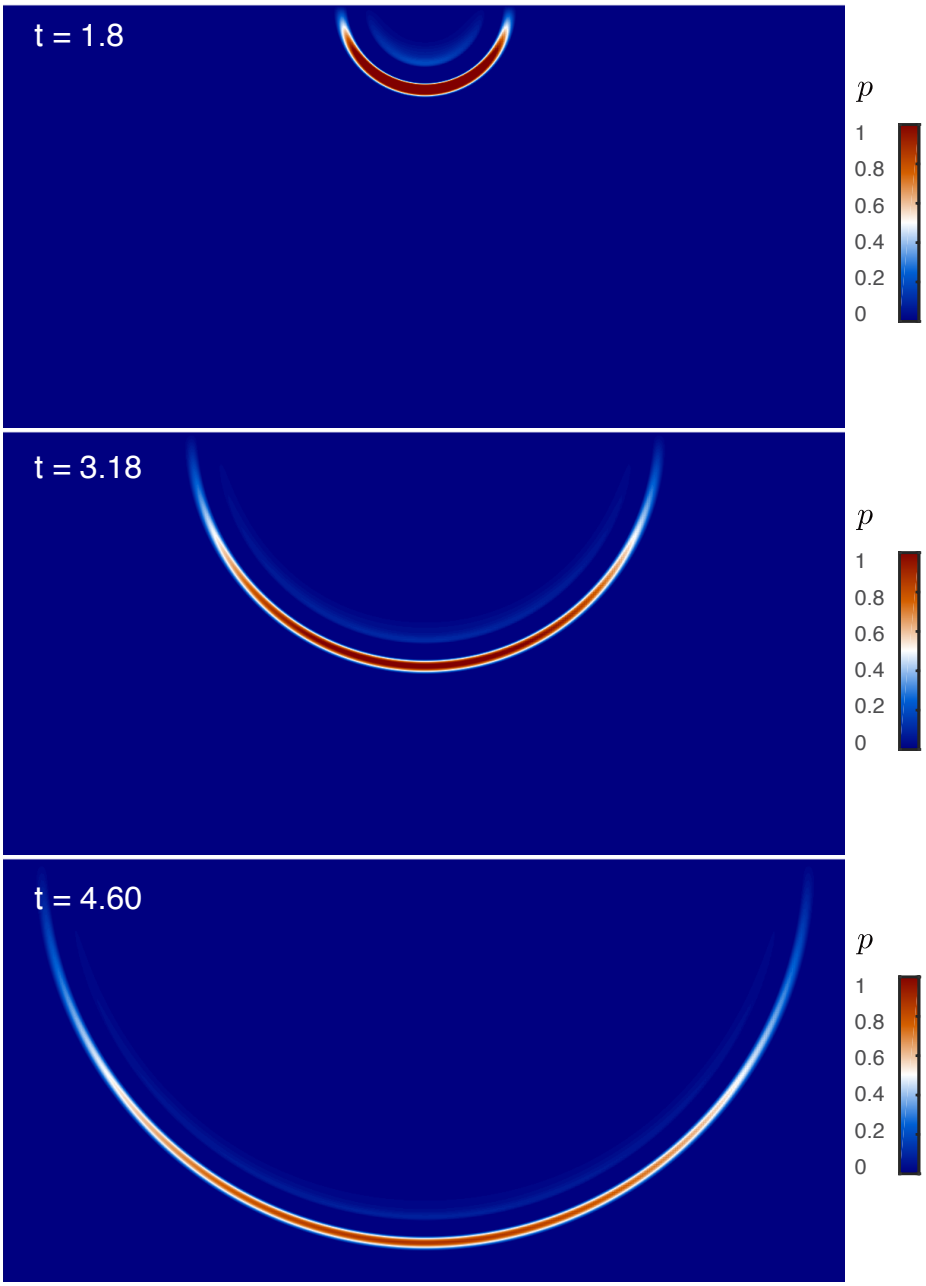


Figure 5: The normalized pressure wave field generated by a singular source term placed in the middle of the top boundary. The simulation uses 6 grid points per minimum wavelength and the 6th-order SBP staggered grid operators. Note that only positive values of p are shown.

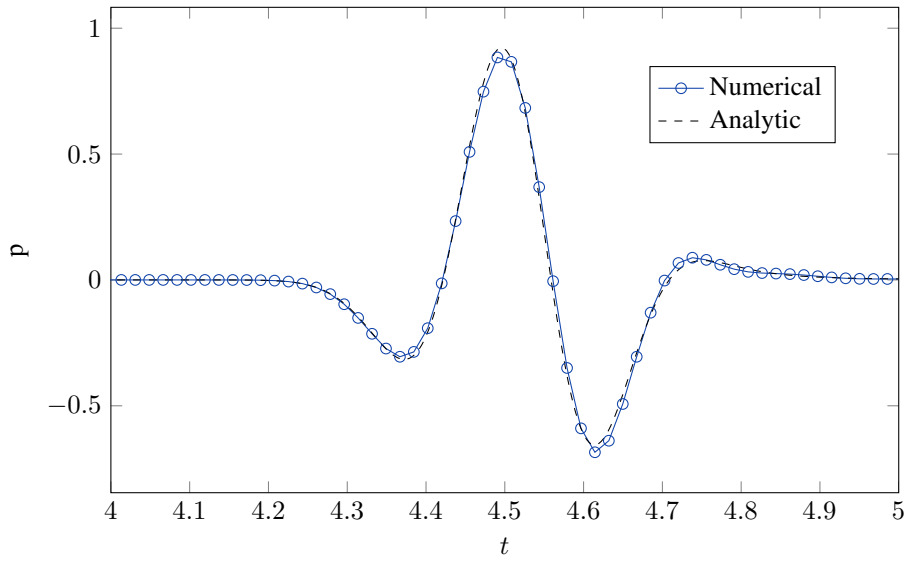


Figure 6: The numerical solution obtained at a fixed point in space 26 minimum wavelengths away from the source. The simulation uses 6 grid points per minimum wavelength and the 6th-order SBP staggered grid operators.

for some function $f(x)$. We approximate (24) by

$$\sum_{j=0}^{N} d_j f(x_j) \approx f(x_*), \tag{25}$$

where the sum is taken over N grid points and the weights d_j at each grid point x_j are determined such that (25) is an accurate approximation of (24). If the source location x_* coincides with a grid point x_l, then the following approximation is commonly used

$$d_j = \begin{cases} \frac{1}{\Delta x}, & j = l, \\ 0, & j \neq l. \end{cases}$$

Otherwise, if x_* is not located exactly on a grid point, then the approximation (26) will be 1st-order accurate. We store the discretization of the delta function at the cell-centered grid points along the top boundary in the vector $\mathbf{d} = \left(d_0,\, d_{1/2},\, \ldots,\, d_{N-1/2},\, d_N\right)^T$ and set the source location to the grid point in the middle of the boundary. The boundary condition is implemented by

$$\mathbf{B}\mathbf{T}_T = P^{-1} E_T P_T \Sigma_T L_T (r = -1) \left(E_T^T \mathbf{u} - g(t)\mathbf{e}\right),\ \mathbf{e} = (\mathbf{d},\, \mathbf{0},\, \mathbf{0})^T,$$

where the dimensionless penalty matrix Σ_T and boundary operator L_T are determined by (19).

We use the 6th-order SBP staggered grid operators with $2\pi/\Delta x \approx 6$ grid points per minimum wavelength. The minimum wavelength λ_{min} is defined by the maximum frequency f_{max} above which the spectral amplitude becomes less than 5% of the maximum value associated with the fundamental frequency f_0 of the Ricker wavelet (23) [4]. The maximum frequency is estimated as $f_{max} \approx 2.5 f_0$. Thus, $\lambda_{min} = 1/f_{max}$ since the wave equation has been nondimensionalized. Figure 5 shows the pressure wave field at different instances in time. We also compare the numerical solution against an analytic solution at a fixed point 26 minimum wavelengths away from the source (Figure 6). The numerical solution is in very good agreement with the analytic solution. The maximum error in amplitude is $\approx 4\%$.

6 CONCLUSIONS

We have extended the SBP-SAT methodology to staggered grids. The general procedure was outlined for both the wave equation in one dimension and two dimensions. The accuracy of the scheme was investigated using a smooth analytic solution. A test with a singular source term on the boundary was used to show that the SBP boundary modification did not destroy the excellent dispersion properties of staggered grid difference methods. Due to the generality of the SBP-SAT methodology, the work can easily be extended to other types of wave equations.

Acknowledgement

We thank Eric M. Dunham for many helpful discussions. Eric M. Dunham derived the analytic solution used in the final numerical experiment.

REFERENCES

[1] Gustafsson, B. and Kreiss, H.O. and Oliger, J, *Time dependent problems and difference methods*. Wiley-Interscience, 1995.

[2] M.H. Carpenter,D. Gottlieb, and S. Abarbanel, Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes, *Journal of Computational Physics*, **111**, 220, 1994.

[3] B.Strand, Summation by parts for finite difference approximations for d/dx *Journal of Computational Physics*,**110**,47, 1994.

[4] Dimitri Komatitsch and Jean-Pierre Vilotte, The spectral element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures, *Bulletin of the seismological society of America*, **88-2**,368-392, 1998,

[5] Zingg, David W, Comparison of high-accuracy finite-difference methods for linear wave propagation, *SIAM Journal on Scientific Computing*, **22-2**,476-502, 2000.

[6] Nordström, Jan and Carpenter, Mark H, High-order finite difference methods, multidimensional linear problems, and curvilinear coordinates, *Journal of Computational Physics*, **173-1**, 149-174, 2001.

[7] Nordström, Jan and Björck, Martin, Finite volume approximations and strict stability for hyperbolic problems, *Applied numerical mathematics*, **38-3**,237-255, 2001.

[8] Hesthaven, Jan S, High-order accurate methods in time-domain computational electromagnetics: A review, *Advances in imaging and electron physics*, **127**, 59-125, 2003.

[9] Nordström, Jan and Gustafsson, Rikard, High order finite difference approximations of electromagnetic wave propagation close to material discontinuities, *Journal of Scientific Computing*, **18-2**, 215-234, 2003.

[10] Svärd, Magnus and Nordström, Jan On the order of accuracy for difference approximations of initial-boundary value problems, *Journal of Computational Physics*, **218-1**, 333-352, 2006.

[11] Svärd, Magnus and Nordström, Jan, Review of summation-by-parts schemes for initial–boundary-value problems, *Journal of Computational Physics*, **268**,17-38,2014.

[12] Nordström, Jan, Well Posed Problems and Boundary Conditions in Computational Fluid Dynamics, *22nd AIAA Computational Fluid Dynamics Conference, at Dallas, Texas, USA*, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ALTERNATIVE TO RETURN-MAPPING ALGORITHM FOR COMPUTING PLASTIC STRAIN. APPLICATION TO DILATANT MATERIALS

Siegfried Maiolino[1,2]

[1]Cerema/DTerCE/DLL/RRMS–ER29
25, avenue Mitterrand
69674 Bron cedex France
e-mail: siegfried.maiolino@cerema.fr

[2] Laboratoire de Mécanique des Solides –UMR 7649
École Polytechnique
91128 Palaiseau cedex France

Keywords: Computing Methods, Elastoplasticity, Finite Element Methods, Numerical Methods, Geomechanics, Return Mapping

Abstract. *Computing plastic strain is a crucial issue in finite element methods. This problem is also known as closest point projection. The radial return used for circular models reduces the computations to literal expressions. But in geomechanics, the deviatoric shape of yield functions is generally non circular, so that return mapping algorithm becomes cumbersome and time consuming.*

Works that will be presented rather focus on a geometric based methods. It will be demonstrated that the numerical problem of closest point projection of the trial stress on the yield surface is equivalent to a geometrical bounded problem. Whereas this property is intuitive, the tools ensuring a straightforward equivalence between the two problems were to be developed.

We identify the geometric problem associated to the problem of the closest point projection in the deviatoric plane. The geometric problem is independent from the mechanical one, and can be solved with trigonometric and geometric laws. Those laws are integrated in a general algorithm to compute plastic strain, taking account of associated and non associated dilatancy for the computation of volumic plastic strains

1 INTRODUCTION

The Mohr'envelope of many porous media - soils, rocks, bones, compacted powder, show their dependence to mean stress, and also differences in strength between triaxial extension and traxial compression. Criteria like Coulomb or Hœk-Brown [3] take into account this dependence but present corners, whereas circular yield functions like Drucker-Prager don't. Experimental results using true triaxial tests prove that geomaterials present a triangular deviatoric shape with rounded corners [12]. Taking into account this particular shape in a smooth criterion involve using the third invariant. Various yield functions had been proposed, for soils [6, 11], concrete [16] or rocks [8].

The radial return [15] used for circular models reduces the computations to literal expressions [5]. The main drawback of non circular models is that return mapping algorithm becomes complex, expensive [13] and time consuming, even if efficiency is increased with spectral decomposition techniques [1, 2]. Works that will be presented rather focus on a geometric based methods : in order to bypass the computational costs of return mapping algorithm, we will focus on simpler equivalent geometric problems.

2 POLAR DECOMPOSITION OF THE YIELD SURFACE

Traction stresses are positive, and the principal stresses ordered as follow : $\sigma_I \geq \sigma_{II} \geq \sigma_{III}$

2.1 Geometric parametrage

For a given mean stress $\left(\sigma_m = \text{Tr}\underline{\underline{\sigma}}/3\right)$, the yield surface can be reduced to its cross-sectional shape on the deviatoric plane, or π plane. A yield surface ($f\left(\underline{\underline{\sigma}}\right) = 0$) can be represented in a unique manner by the mean stress and the deviatoric invariants $\left(J_2 = \frac{1}{2}\text{Tr}\left(\underline{\underline{s}}^2\right),\ J_3 = \frac{1}{3}\text{Tr}\left(\underline{\underline{s}}^3\right)\right.$, with $\left.\underline{\underline{s}} = \underline{\underline{\sigma}} - \sigma_m\underline{\underline{\mathbb{1}}}\right)$, but it is more practical to replace the third invariant by the Lode angle θ, to work in the π plane (deviatoric plane).

$$-\frac{\pi}{6} \leq \theta = \frac{1}{3}\arcsin\left(\frac{-3\sqrt{3}}{2}\frac{J_3}{\sqrt{J_2}^3}\right) \leq \frac{\pi}{6} \tag{1}$$

The set $\left(\sqrt{J_2}, \theta\right)$ define polar coordinates on one sixth of the deviatoric plane, which is sufficient for an isotropic criterion. Zienkiewicz and Pande [17], using the fact that a yield surface can be reduced to its polar expression, provided tools to study the regularity, the sensitivity to the extension and the convexity of a criterion starting from the shape function $g_p\left(\theta\right)$

$$\sqrt{J_2} = \sigma^+ g_p\left(\theta\right) \tag{2}$$

The deviatoric radius : $\sigma^+(\sigma_m) = \sqrt{J_2}_{/\theta=\frac{\pi}{6}}$, gives the yield function in the meridional plane $\left(\sigma_m, \sqrt{J_2}\right)$, for $\theta = \frac{\pi}{6}$. This value of the Lode angle corresponds to a classical triaxial test, or compression triaxial test ($\sigma_I = \sigma_{II} > \sigma_{III}$). The function $g_p\left(\theta\right)$ is the shape function of the yield surface in the deviatoric plane. It is normalized $\left(g_p\left(\frac{\pi}{6}\right) = 1\right)$ and gives directly the value of the extension ratio $g_p\left(-\frac{\pi}{6}\right) = L_S$ which is more detailed in the following section.

2.2 Characteristic function of a material

The deviatoric radius σ^+ can be easily deduced from triaxial compression tests. Whether the shape is straight or parabolic, the deviatoric radius used can be the Coulomb or Hœk-Brown.

The extension ratio L_S has a physical meaning and can be determined from experiment. The condition $\theta = -\frac{\pi}{6}$ corresponds to extension triaxial tests ($\sigma_I > \sigma_{II} = \sigma_{III}$), which can be performed with the same triaxial cell as compression triaxial test.

$$L_S = \frac{\sqrt{J_2}\left(\theta = -\frac{\pi}{6}\right)}{\sqrt{J_2}\left(\theta = \frac{\pi}{6}\right)} = \frac{(\sigma_I - \sigma_{III})\,(extension)}{(\sigma_I - \sigma_{III})\,(compression)} \tag{3}$$

Physically, this means that for the same mean stress, the yield value of $\sqrt{J_2}$ would be lower in extension than in compression. The value of L_S is directly linked to the deviatoric shape of a yield surface. While this value can be independent from the mean stress (Coulomb), some rocks present a shape of their yield surface changing from triangular (low confinement) to circular (high confinement) [4], i.e, L_S increases from 0.5 to 1.

2.3 Introduction of the orthoradial tensor

We consider for stresses and strains (i.e. symetric second order tensors) the following scalar product. For two tensors, $\underline{\underline{T_1}}$ and $\underline{\underline{T_2}}$:

$$\underline{\underline{T_1}} \cdot \underline{\underline{T_2}} = \underline{\underline{T_1}} : \underline{\underline{T_2}} = \mathrm{Tr}\underline{\underline{T_1 T_2}} \tag{4}$$

Hence defining the following norm (Frobenius norm) for a symetric second order tensor $\underline{\underline{T}}$:

$$\| \underline{\underline{T}} \| = \sqrt{\underline{\underline{T}} : \underline{\underline{T}}} \tag{5}$$

We introduce the orthoradial tensor, $\underline{\underline{v}}$:

$$\underline{\underline{v}} = 3\frac{\sqrt{3}}{2}\frac{1}{J_2}\underline{\underline{s}}^2 - \sqrt{3}\underline{\underline{\mathbb{1}}} - \frac{9\sqrt{3}J_3}{4{J_2}^2}\underline{\underline{s}} \tag{6}$$

This tensor is orthoradial as $\underline{\underline{v}} \cdot \underline{\underline{\mathbb{1}}} = \underline{\underline{v}} \cdot \underline{\underline{s}} = 0$, hence those three tensors, $\underline{\underline{\mathbb{1}}}$, $\underline{\underline{s}}$, $\underline{\underline{v}}$ constitute an orthogonal basis of symetric second order tensors, for the Frobenius scalar product.

We can then easily decompose the derivatives of the yield function along this orthogonal basis, as the expression of the gradient of invariants can easily be expressed. It is necessary to introduce the orthogonal tensor, as the gradient of the third invariant cannot be expressed using only the hydrostatic tensor or the deviatoric tensor. Expressions of the three gradients of invariants using the orthogonal basis are the following :

$$\frac{\partial I_1}{\partial \underline{\underline{\sigma}}} = \underline{\underline{\mathbb{1}}} \tag{7}$$

$$\frac{\partial J_2}{\partial \underline{\underline{\sigma}}} = \underline{\underline{s}} \tag{8}$$

$$\frac{\partial J_3}{\partial \underline{\underline{\sigma}}} = \underline{\underline{s}}^2 - \frac{2J_2}{3}\underline{\underline{\mathbb{1}}} = \frac{3J_3}{2J_2}\underline{\underline{s}} + \frac{2J_2}{3\sqrt{3}}\underline{\underline{v}} \tag{9}$$

Hence the gradient of any yield surface can be orthogonally decomposed.

$$\frac{\partial f}{\partial \underline{\underline{\sigma}}} = f_u\underline{\underline{\mathbb{1}}} + f_s\underline{\underline{s}} + f_v\underline{\underline{v}} \tag{10}$$

We can observe that the deviatoric part of the gradient, $\frac{\partial f}{\partial \underline{\underline{\sigma}}}$ can be split in two orthogonal component, a radial $f_s\underline{\underline{s}}$, and a orthoradial, $f_v\underline{\underline{v}}$. This later part is null for criteria independent from the third invariant.

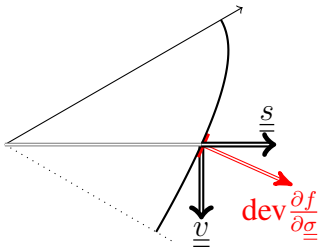


Figure 1: Otrthogonal decomposition of the deviatoric part of the gradient

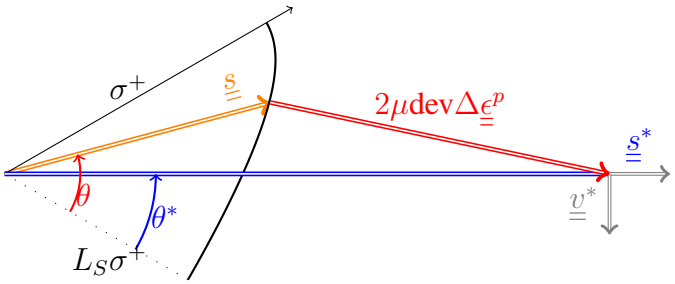


Figure 2: Physical problem in the deviatoric plane

2.4 Consequences for return mapping algorithm

Let us consider at an integration point, the increment from step n to $n+1$. We want to calculate the plastic strain, if the trial stress $\underline{\underline{\sigma}}^*$ doesn't satisfy the yield condition. We want to implicitly solve the relation that gives the stress.

$$\underline{\underline{\sigma}}_{n+1} = \underline{\underline{\sigma}} = \underline{\underline{\sigma}}^* - \underline{\underline{\underline{\underline{L}}}}\Delta\underline{\underline{\epsilon}}^p \tag{11}$$

Where $\underline{\underline{\underline{\underline{L}}}}$ is the elasticity tensor. We can split this relation between two orthogonal components : an hydrostatic and a deviatoric.

$$\sigma_m - \sigma_m^* = -K\text{Tr}\Delta\underline{\underline{\epsilon}}^p \tag{12}$$

$$\underline{\underline{s}} - \underline{\underline{s}}^* = -2\mu\text{dev}\Delta\underline{\underline{\epsilon}}^p \tag{13}$$

Where K is the bulk modulus and μ the shear modulus.

We can notice that the hydrostatic part (12) is purely scalar and that the main difficulties come from the deviatoric part(13).

3 GEOMETRIC EQUIVALENCE OF CLOSEST POINT PROJECTION

We introduce the following quantity [9, 10]:

$$\rho = \frac{\sqrt{J_2}^*}{\sigma^+(\sigma_m^* + \Delta\sigma_m)} \tag{14}$$

The closest point projection of $\underline{\underline{s}}^*$ on the trace of yield surface (Figure 2) is equivalent to the following geometric problem(Figure 3): find the closest point projection (polar coordinates:

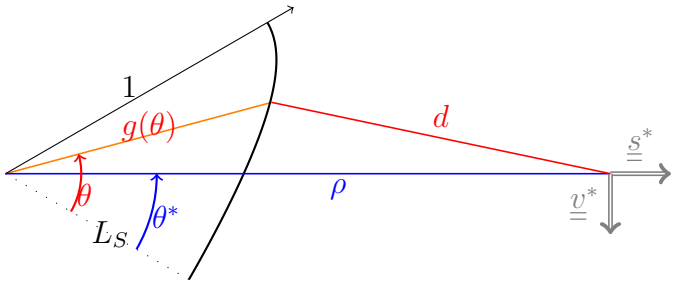


Figure 3: Geometric problem in polar coordinates

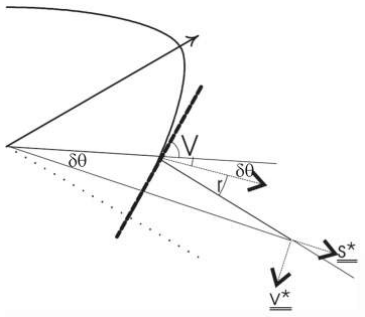


Figure 4: Angular relations: $V + r + \delta\theta = \frac{\pi}{2}, (\delta\theta = \theta - \theta^*)$

$(\theta, g_p(\theta))$) of the point (ρ, θ^*) on the curve defined by the shape function $g_p(\theta)$. The function $d(\theta)$ reaches its minimum at this point.

$$d(\theta)^2 = g_p^2(\theta) + \rho^2 - 2g_p(\theta)\,\rho\cos(\theta - \theta^*) \tag{15}$$

The norm of the deviatoric plastic strain is given directly : $\| \text{dev}\Delta\underline{\underline{\epsilon}}^p \| = \frac{d(\theta)\sigma^+}{\sqrt{2}\mu}$, and $\text{dev}\Delta\underline{\underline{\epsilon}}^p$ can be written in the local base associated to the trial stress

$$\text{dev}\Delta\underline{\underline{\epsilon}}^p = \| \text{dev}\Delta\underline{\underline{\epsilon}}^p \| \left(\frac{\cos r}{\sqrt{2}\sqrt{J_2^*}}\underline{\underline{s}}^* + \frac{\sin r}{\| \underline{\underline{v}}^* \|}\underline{\underline{v}}^* \right) \tag{16}$$

The value of the r angle can be deduced from trigonometric considerations (figure 4), V being the angle between the tangent to a polar curve and the radial axis :

$$\tan V = \frac{g}{g'} \tag{17}$$

4 COMPUTATION OF DEVIATORIC AND HYDROSTATIC PLASTIC STRAIN

The geometric equivalence allows to get the plastic strain if $\Delta\sigma_m$ is known. This quantity is calculated using an iterative method. Initially, $\Delta\sigma_m^0 = 0$, and the stopping condition c is satisfied when the ratio between hydrostatic and deviatoric parts is equal to the dilatancy angle δ.

$$c(\Delta\sigma_m, \| \text{dev}\Delta\underline{\underline{\epsilon}}^p \|, \tan\delta) = \| \text{dev}\Delta\underline{\underline{\epsilon}}^p \| + \frac{\Delta\sigma_m}{\sqrt{3}K\tan\delta} < \varepsilon \tag{18}$$

We use the following definition of the dilatancy angle :

$$\tan\delta = -\frac{\mu_l}{\sqrt{2}}\frac{\partial\epsilon_V^p}{\partial\epsilon_d^p} \tag{19}$$

Where ϵ_V^p is the volumic plastic strain, and $\mu_l = \tan 3\theta$ the factor initially defined by Lode [7]. Using the Lode factor allows to use a formulation that can be used for triaxial compression test as well as extenions test [14].

The dilatancy angle can be deduced from the expression of the yield function (associated potential) or from a non associated potential. Whereas it is not easy to identify the potential, without extensive true trixial tests, the dilatancy angle can be easily identified using classical triaxial compression tests.

If $\delta > 0$ material is said to be dilatant. The alternative return mapping algorithm can be expressed as follow, at a given integration point, for a dilatant material

1. Compute $\underline{\underline{\sigma}}^* = \underline{\underline{\sigma}}^0 + \underline{\underline{\underline{\underline{L}}}} \left(\underline{\underline{\epsilon}}_{n+1} - \underline{\underline{\epsilon}}_n^p \right)$
2. Check $f(\underline{\underline{\sigma}}^*) > 0$? No set $\underline{\underline{\sigma}}_{n+1} = \underline{\underline{\sigma}}^*$ and exit.
3. Yes : set i = 0 and $\Delta\sigma_m^0 = 0$
4. Set $\rho^i = \frac{\sqrt{J_2^*}}{\sigma^+(\sigma_m^* + \Delta\sigma_m^i)}$ and if L_S depends of mean stress : $L_S^i = L_S\left(\sigma_m^* + \Delta\sigma_m^i\right)$
5. Compute θ^i, $d\left(\theta^i\right)$, $\sqrt{J_2}^i$ and $\parallel \text{dev}\Delta\underline{\underline{\epsilon}}^p \parallel^i$
6. Evaluate $\tan\delta^i = \tan\delta\left(\sigma_m^* + \Delta\sigma_m^i, J_2^i, J_3^i\right)$
7. Evaluate stopping criterion $\mid c\left(\sigma_m^* + \Delta\sigma_m^i, \parallel \text{dev}\Delta\underline{\underline{\epsilon}}^p \parallel^i, \tan\delta^i\right) \mid < \epsilon$. If Yes compute angle r and tensor $\Delta\underline{\underline{\epsilon}}^p$. Update $\underline{\underline{\sigma}}_{n+1} = \underline{\underline{\sigma}}^* - \underline{\underline{\underline{\underline{L}}}}\Delta\underline{\underline{\epsilon}}^p$ and exit.
8. If No, set $i = i + 1$ and $\Delta\sigma_m^i = \Delta\sigma_m^{i-1} + \Delta^2\sigma_m$, then loop to step 4.

The evaluation of $\Delta^2\sigma_m$ depends of the nature of yield function f and δ

5 CONCLUSIONS

We have shown that for J_3 dependant yield function, the problem of closest point projection is equivalent to a pure geometric problem in polar coordinates. For different values of (ρ, θ^*), solutions can be computed and values of $d\left(\theta\right)\sin 3\theta$ and $g_p\left(\theta\right)$ are saved, allowing to short-cut computational costs of return mapping in the deviatoric plane. For the hydrostatic part, we propose a general algorithm that allows to take account of the dilatancy in the computation of plastic strain. Next step will be to define the stability conditions for negative dilatancy (i.e. contractant material). The last wil be to extend the process to non associated plasticity in the deviatoric plane. References and experimental studies of non associated flow rules in the deviatoric part are scarce, so many cases must be considered.

References

[1] R. Borja, K. Sama, and P. Sanz. On the numerical integration of three invariant elasto-plastic constitutive models. *Comput. Methods Appl. Mech. Engrg*, 192(9-10):1227–1258, 2003.

[2] C. Foster, R. Regueiro, A. Fossum, and R. Borja. Implicit numerical integration of a three-invariant, isotropic/kinematic hardening cap plasticity model for geomaterials. *Comput. Methods Appl. Mech. Engrg.*, 194(50-52):5109–5138, 2005.

[3] E. Hoek and E. Brown. Empirical strength criterion for rock masses. *J. Geotech. Engng DIV., ASCE*, 106(GT9):1013–1035, 1980.

[4] M. Kim and P. Lade. Modelling rock strength in three dimensions. *Int. Journ. Rock Mech. Min. Sci. Abstracts*, 21(1):21–33, 1984.

[5] R. D. Krieg and S. M. Key. Implementation of a Time Dependant Plasticity Theory into Structural Computer Programs. *Constitutive Equations in Viscoplasticity : Computational and Engineering Aspects*, 20:125–137, 1976.

[6] P. V. Lade. Elasto-plastic stress-strain theory for cohesionless soil with curved yield surfaces. *Int. Journ. Solids Structures*, 13:1019–1035, 1977.

[7] W. Lode. Versuche über den Einfluß der mittleren Hauptspannung auf das Fließen der Metalle Eisen, Kupfer und Nickel. *Zeitschrift für Physik*, 36:913–939, 1926.

[8] S. Maïolino. Proposition of a general yield function in geomechanics. *Comptes Rendus Mécanique*, 333:279–284, 2005.

[9] S. Maïolino. *Fonction de charge générale en géomécanique : application aux travaux souterrains*. PhD thesis, École Polytechnique, 2006. General yield function in geomechanics : application to tunneling (in French).

[10] S. Maïolino. Numerical abacuses method based on the equivalence between the closest point projection and a bounded geometric problem. In *8th.World Congress on Computational Mechanics (WCCM8) , 5th European Congress on Computational Methods in Applied Sciences and Engineeering (ECCOMAS 2008), Venice*, July 2008.

[11] H. Matsuoka and T. Nakai. Stress-deformation and strength characteristics of soil under three different principal stresses. In *Proc. JSCE*, volume 232, pages 59–70, 1974.

[12] P. Michelis. True triaxial yelding and hardening of rock. *J. Geotech. Engng DIV., ASCE*, 113(6):616–635, 1987.

[13] J. Simo and T. Hughes. *Computational Inelasticity*. 1998.

[14] L. Thorel. *Plasticité et endommagement des roches ductiles - application au sel gemme.* PhD thesis, 1994.

[15] J. L. Wilkins. Calculation of Elastic-plastic Flow. *Methods of Compuational Physics*, 8, 1964.

[16] K. William and E. Warnke. Constitutive models for the triaxial behavior of concrete. In *International Association of Bridge and Structural Engineering (IABSE) Seminar on "Concrete Structures Subjected to Triaxial Stresses", Bergamo*, volume 19, pages 1–30, 1975.

[17] O. C. Zienkiewicz and G. N. Pande. Some useful forms of isotropic yeld surfaces for soil and rock mechanics. In *Numerical methods in soil and rock mechanics, Karlsruhe*, pages 3–16, september 1975.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DISCONTINUOUS GALERKIN METHOD WITH REDUCED INTEGRATION SCHEME FOR THE BOUNDARY TERMS IN ALMOST INCOMPRESSIBLE LINEAR ELASTICITY

Hamid R. Bayat, Stephan Wulfinghoff, and Stefanie Reese

Institute of Applied Mechanics (IFAM), RWTH Aachen University
Mies-van-der-Rohe St. 1, 52074 Aachen, Germany

e-mail: hamid.reza.bayat@rwth-aachen.de
stephan.wulfinghoff@rwth-aachen.de
stefanie.reese@rwth-aachen.de

Keywords: Discontinuous Galerkin, Reduced Integration, Linear Elasticity, Incompressibility, Locking.

Abstract.

A problem of incompressibility is explored in this paper with application of a discontinuous Galerkin (dG) method with reduced integration. We apply two reduced integration schemes, namely fully and mixed reduced integrations for the Incomplete Interior Penalty Galerkin (IIPG) class of the dG method. The numerical results show convergent solutions with respect to a sufficiently large value of the penalty term and number of elements. Additionally, a comparison between the standard continuous Galerkin (cG) method and the dG method are established to compare and contrast the behavior. Finally, the dG method shows a faster convergence with respect to the number of elements.

1 INTRODUCTION

Recent advancements in the field of computational engineering have enabled scientists to define novel discretization schemes in solid mechanics. One of these schemes, discussed in this paper is the discontinuous Galerkin (dG) method. Unlike many standard Galerkin methods, the dG method is a non-conforming finite element method. Using the dG method, one obtains discontinuities between the interior element boundaries. This is achieved through a weak enforcement of the continuity on displacement. To this end, one must modify the weak form in a way that the integration by parts is applied not only on the domain boundaries, but also additionally, on the subdomain boundaries [1].

The discontinuous Galerkin method was initially introduced as a new classification in finite element methods by Reed and Hill [2] to solve a problem of a nuclear transport partial differential equation in 1971. Later, Baker [3] applied a dG method with some modifications for elliptic problems - refer to [4] and [5] for more details. Nitsche [6] contributed to dG a penalty term on the internal subdomain boundaries in order to stabilize the solution.

Although, dG was initially used in fluid mechanics [7], it found its way to solid mechanics as well (e.g. [8]) to remedy frequently encountered problems. An application of this method is seen in elliptic problems like incompressibility, which results in the well-known volumetric locking phenomenon or in some shell (plate) constructions with shear locking problems [8]. Hansbo und Larson [9] investigated the locking-free behavior of the dG method for (near) incompressibility with triangle meshes. Another application of the dG method in solid mechanics can be found in the work of Mergheim et al. [10]. They applied dG elements in the prefailure regime to avoid stress oscillations just before the failure.

In this paper, we investigated the reduced integration method for certain boundary terms to make this method more efficient and reduce the time of calculation.

The present work is organized as follows: the first chapter introduces the governing equations of the dG method. Then the numerical integration scheme and the reductions, which are applied in order to decrease the calculation time, are clarified. Finally, a benchmark example is simulated to evaluate the method and investigate the differences between the continuous and discontinuous Galerkin methods.

2 GOVERNING EQUATIONS OF DG

The dG method has different variations depending on the extended terms of its weak form. In our case, we apply the Incomplete Interior Penalty Galerkin (IIPG). This method is non-symmetric due the absence of the symmetric term of dG method and contains the stabilization term, namely penalty term.

2.1 Strong form

The strong form of the equilibrium of the forces in standard FEM and the boundary conditions are given by

$$\begin{aligned} -div(\boldsymbol{\sigma}) &= \boldsymbol{f} \quad \text{in} \quad \mathcal{B}, \\ \boldsymbol{u} &= \boldsymbol{u}^p \quad \text{on} \quad \partial\mathcal{B}_u, \\ \boldsymbol{\sigma}\,\boldsymbol{n}_e &= \boldsymbol{t}^p \quad \text{on} \quad \partial\mathcal{B}_t. \end{aligned} \tag{1}$$

Where $\boldsymbol{\sigma} = \boldsymbol{C}:\boldsymbol{\varepsilon}$ is the Cauchy stress tensor with $\boldsymbol{C}$ as the forth order stiffness tensor, e.g. elasticity module E and Poisson's ratio ν and $\boldsymbol{\varepsilon}$ as the symmetric strain tensor, which is in fact the symmetric part of the gradient of the displacement field $\boldsymbol{u}$. Additionally, $\boldsymbol{f}$ represents the body forces, $\boldsymbol{u}^p$ and $\boldsymbol{t}^p$ are the prescribed displacement and prescribed traction on the Drichlet boundary $\partial\mathcal{B}_u$ and Neumann boundary $\partial\mathcal{B}_t$, respectively.

In the dG method we additionally introduce discontinuities along the internal boundaries Γ of the body $\mathcal{B}$. This divides the body into to $\mathcal{B}^+$ and $\mathcal{B}^-$ parts, with the normal vector $\hat{\boldsymbol{n}}$ directing from the negative side to the positive side, see Figure 1.

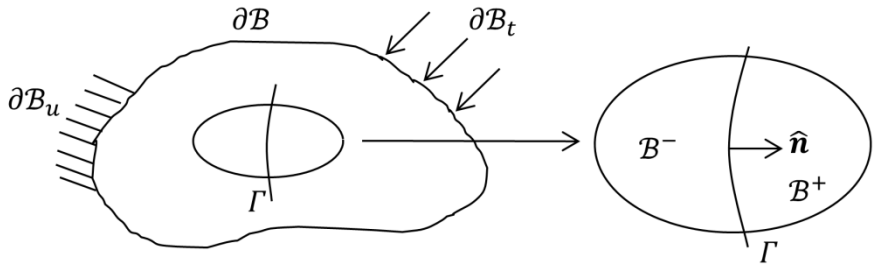


Figure 1: Discontinuity Γ within the body $\mathcal{B}$.

The jump $[\![\blacksquare]\!]$ and average $\{\blacksquare\}$ of quantities are defined as follows:

$$[\![\boldsymbol{u}]\!] = \boldsymbol{u}^+|_\Gamma - \boldsymbol{u}^-|_\Gamma,$$

$$\{\boldsymbol{u}\} = \frac{1}{2}\,(\boldsymbol{u}^+|_\Gamma + \boldsymbol{u}^-|_\Gamma). \tag{2}$$

Introducing the dG method to the strong form, we need to impose extra conditions on the internal boundaries Γ. On these boundaries, the continuity of the displacements field and the traction vector is prescribed by the exact solution:

$$[\![\boldsymbol{u}]\!] = 0$$

$$[\![\boldsymbol{\sigma}\,\hat{\boldsymbol{n}}]\!] = 0. \tag{3}$$

2.2 Weak form

The weak form of dG is obtained by integration by parts on the internal subdomains. Nonetheless, the penalty term must be added to stabilize the solution [6]:

$$\int_{\mathcal{B}^+\cup\mathcal{B}^-} \boldsymbol{\sigma}:\delta\boldsymbol{\varepsilon}\; dV + \int_\Gamma [\![\delta\boldsymbol{u}]\!]\cdot\{\boldsymbol{\sigma}\}\,\hat{\boldsymbol{n}}\; d\Gamma + \int_\Gamma \theta\,[\![\delta\boldsymbol{u}]\!]\cdot[\![\boldsymbol{u}]\!]\, d\Gamma = \int_{\mathcal{B}^+\cup\mathcal{B}^-} \boldsymbol{f}\cdot\delta\boldsymbol{u}\; dV + \int_{\partial\mathcal{B}_t} \boldsymbol{t}\cdot\delta\boldsymbol{u}\; dA \tag{4}$$

Where θ is the penalty parameter. As discussed in equation (3), the jump of displacements $[\![\boldsymbol{u}]\!]$ is equal to zero in the exact solution.

3 NUMERICAL INTEGRATION SCHEME

The dG element includes the information of two adjacent elements, i.e. displacements, strains and stresses. These quantities which appear in the integrands of the weak form (4) are numerically evaluated in two Gauss points on the discontinuity Γ with the means of Gaussian quadrature. Figure 2 illustrates where the Gauss points are located.

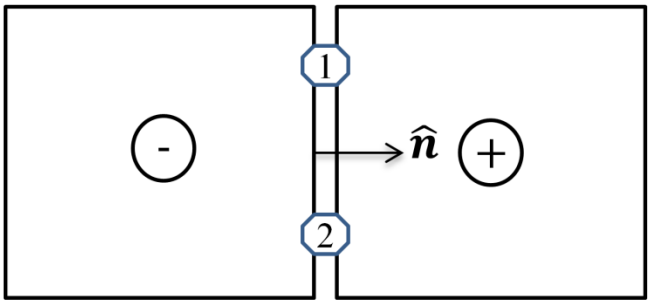


Figure 2. Gauss points 1 and 2 on discontinuity Γ.

The second term and the third term on the left hand side of the weak form (equation (4)), namely the dG term and the penalty term, respectively, are computed in three different ways. First, we evaluate these terms on both Gauss points 1 and 2 as in Figure 2. Then, a fully reduced integration is applied to decrease the Gauss points to one in the middle (Figure 3). Finally, a mixed integration scheme is applied, so that the dG term is evaluated only in the middle of Γ as in Figure 3 and the penalty term on both Gauss points like in Figure 2.

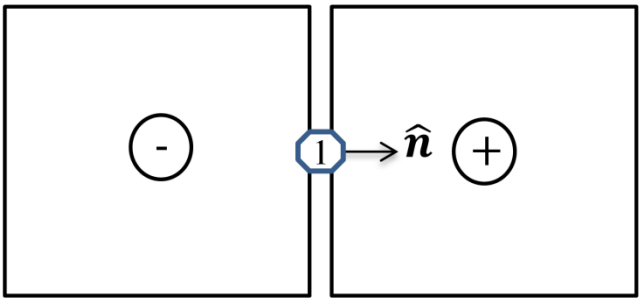


Figure 3. Gauss point 1 in the middle of discontinuity Γ.

4 NUMERICAL EXAMPLE

In this section, we investigate the performance of the dG method in comparison to the standard FEM in a numerical plane strain example with linear elastic isotropic material. The standard quadrilateral finite elements possess four nodes and four Gauss points for numerical integration. In addition, the shape functions are bilinear in both cG and dG elements.

A common benchmark problem [11] of Cook's membrane as shown in Figure 4 is studied. The left side is fixed in both directions and there is an in-plane shear load of $100\ N$ in vertical direction on the right side. The elastic modulus and Poisson's ratio of the material are given

by $E = 250\ MPa$ and $\nu = 0.4999$, correspondingly. The vertical displacement of the node P is to be investigated.

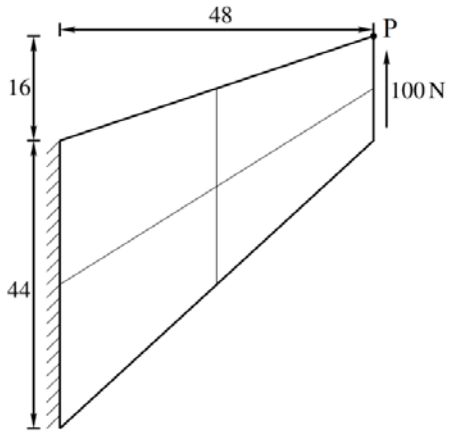


Figure 4. Geometry (in mm), boundary conditions, and loading of Cook's membrane.

Using the Finite Element Analysis Program FEAP to approximate the solution of this problem, we consider two cases of continuous Galerkin (cG) and dG methods. Nevertheless, the dG method was subcategorized into fully reduced and mixed integration schemes.

The results of the simulation depend on two factors, namely the penalty parameter and mesh refinement. Thus, the simulations are carried out for different values of the penalty value θ, varying from 5 to 20,000 and also different mesh sizes. Mesh refinement is done by simultaneous division of the neighboring sides.

Figure 6 and Figure 7 show that for a sufficiently large value of the penalty parameter θ, the results converge with respect to the number of elements. Unsurprisingly, the fully reduced integration scheme converges to wrong solution due to its instability caused by less number of integration points. It is thus more sensitive to the penalty value in comparison to the mixed integration scheme. Taking the θ value greater than 800, the fully reduced integration method converges from 64 element divisions independent of theta value (see Figure 6). Nonetheless, the mixed integration is almost independent of theta, when the number of divisions is higher than 128 (Figure 7).

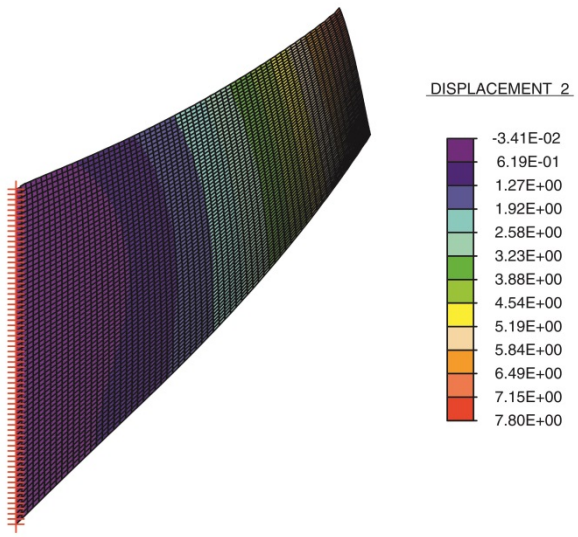


Figure 5. Vertical displacement contour for reduced integration with $\theta = 800$, $n = 64$

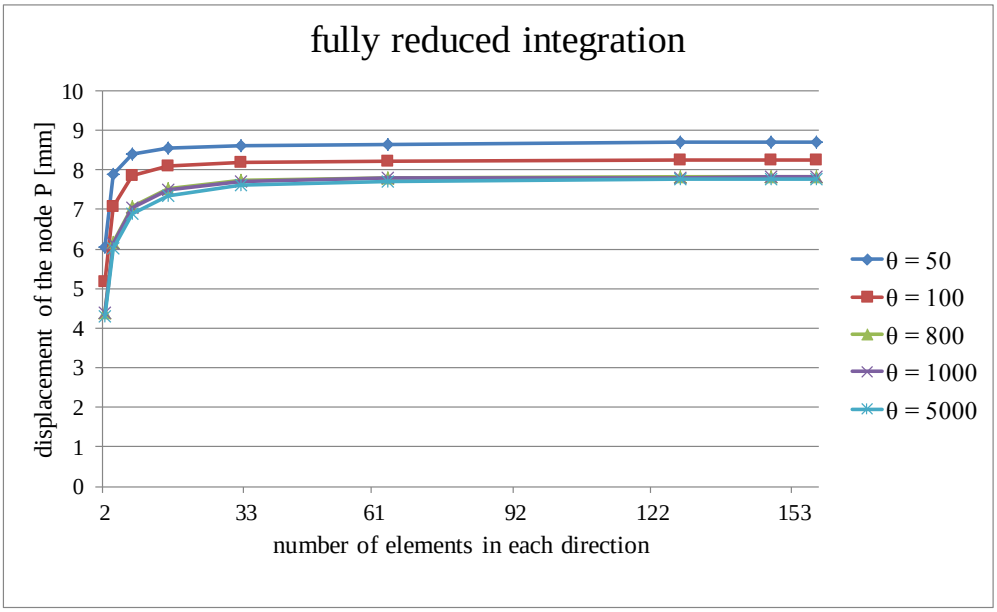


Figure 6. Vertical displacement of the node P for the fully reduced integration scheme

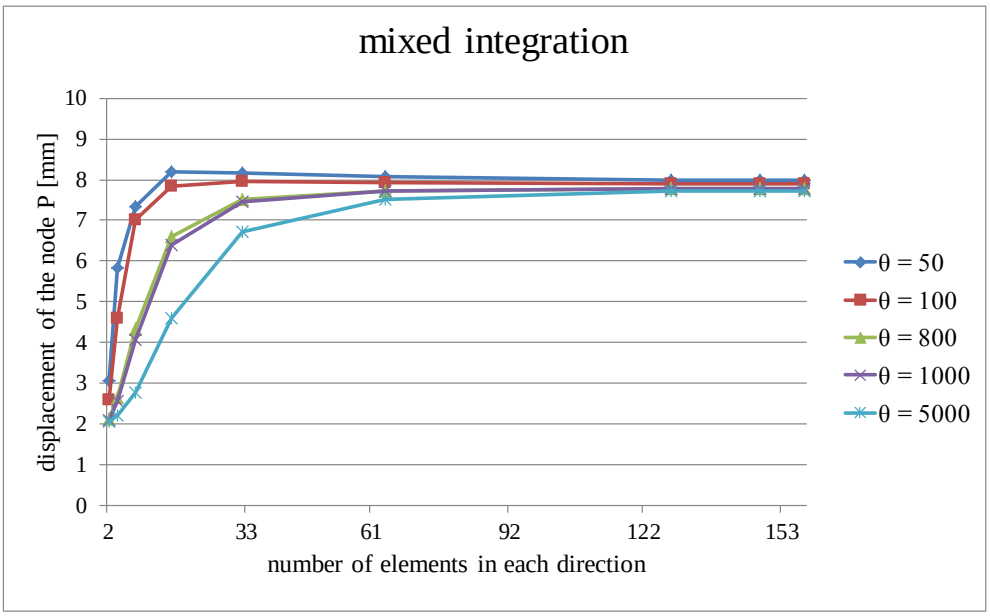


Figure 7. Vertical displacement of the node P for the mixed integration scheme

As it is seen in Figure 8, the dG method converges with much lower number of elements in comparison to the standard continuous Galerkin (cG) method.

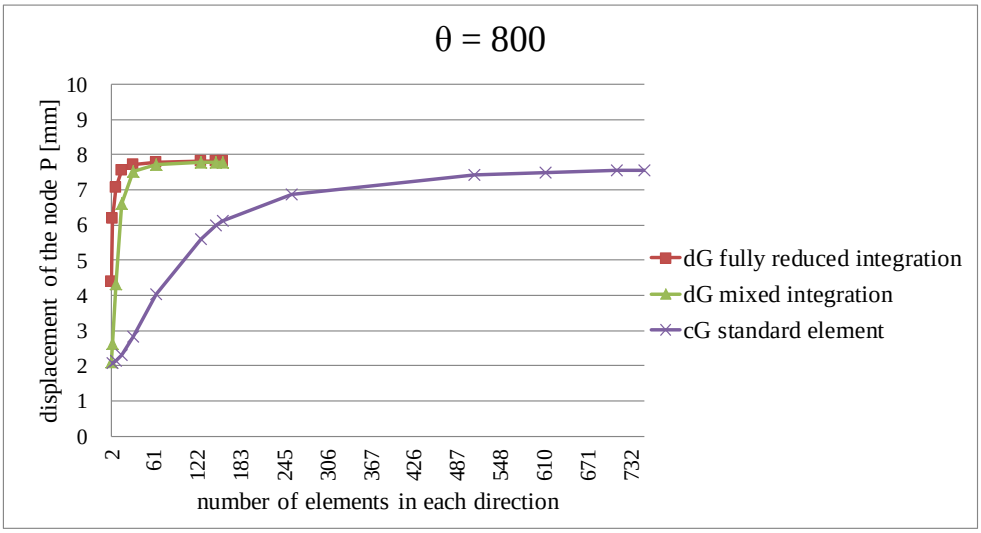


Figure 8. Nodal displacement comparison of cG and dG methods for $\theta = 800$ in vertical direction

5 CONCLUSIONS

In the present work, we utilized the dG method with reduced integration scheme. The results for different cases of the fully reduced and mixed integration methods with respect to IIPG dG terms were examined. Both schemes converged for a specific number of elements and specific value of penalty parameter. Although the dG method converged with a significantly lower number of elements compared to the cG method, one must still consider the major increase of the number of degrees of freedom in dG method, which will bring about a longer calculation time.

6 ACKNOWLEDGMENT

We gratefully appreciate the financial support of the German Science Foundation (DFG) in our project SPP 1748 – “Reliable Simulation Techniques in Solid Mechanics. Development of Non-standard Discretisation Methods, Mechanical and Mathematical Analysis”.

REFERENCES

[1] J. Mergheim, E. Kuh and P. Steinmann, "A finite element method for the computational modelling of cohesive cracks," *International Journal for Numerical Methods in Engineering,* vol. 63, no. 2, p. 276–289, 2005.

[2] W. Reed and T. Hill, "Triangular Mesh Methods for the Neutron Transport Equations," Proceedings of the American Nuclear Society, Los Alamos , 1973.

[3] G. A. Backer, "Finite element methods for elliptic equations using nonconforming elements," *Math. Comp. 31,* pp. 45-59, 1977.

[4] D. N. Arnold, "An Interior Penalty Finite Element Method with Discontinuous Elements," *SIAM Journal on Numerical Analysis,* vol. 19, pp. {742-760, 1982.

[5] M. F. Wheeler, "An Elliptic Collocation-Finite Element Method with Interior Penalties," *SIAM Journal on Numerical Analysis,* vol. 15, pp. 152-161, 1978.

[6] J. Nitsche, "Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind," *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg,* vol. 36, pp. 1865-8784, 2013.

[7] B. Cockburn, G. E. Karniadakis and C. W. Shu, "The Development of Discontinuous Galerkin Methods," *Lecture Notes in Computational Science and Engineering,* pp. 3-50, 2000.

[8] A. Lew, A. Eyck and R. Rangarajan, "Some Applications of Discontinuous Galerkin Methods in Solid Mechanics," *IUTAM Symposium on Theoretical, Computational and Modelling Aspects of Inelastic Media.,* pp. 167-189, 2008.

[9] P. Hansbo and M. G. Larson, ""Discontinuous Galerkin methods for incompressible and nearly incompressible elasticity by Nitsche's method," *Computer Methods in Applied Mechanics and Engineering,* vol. 191, pp. 1895 - 1908, 2002.

[10] J. Mergheim, E. Kuhl and P. Steinmann, "A hybrid discontinuous Galerkin/interface method for the computational modelling of failure," *Communications in Numerical Methods in Engineering,* vol. 20, pp. 511--519, 2004.

[11] B. P. "Lamichhane, B. D. Reddy and B. I. Wohlmuth, "Convergence in the incompressible limit of finite element approximations based on the Hu-Washizu formulation," *Numerische Mathematik,* vol. 104, pp. 151--175, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
J.F. Aguilar Madeira, Helder C. Rodrigues (eds.)
Crete Island, Greece, 5–10 June 2016

A METHOD FOR BI-DIRECTIONAL COUPLING OF STRUCTURE AND SYSTEM IN THE OPTIMIZATION OF MULTI-FUNCTIONAL COMPONENTS

Ajit Panesar*, Ian Ashcroft, Ricky Wildman and Richard Hague

Faculty of Engineering, University Park, University of Nottingham, NG7 2RD, UK
{*ajit.panesar, ian.ashcroft, ricky.wildman, richard.hague}@nottingham.ac.uk

Keywords: Optimization, Multi-functional Additive Manufacturing, Computational Methods

Abstract. *The driver for this research is the development of multi-material additive manufacturing processes that provides the potential for multi-functional parts to be manufactured in a single operation. To exploit the potential benefits of this emergent technology, new design, analysis and optimization methods are needed. This paper proposes a method in which a multifunctional part, consisting of a system, comprised of a number of connected functional components within a mechanical structure, can be optimized. The main contribution of this paper is the coupling strategy that enables the structural topology optimization (TO) of a part to be carried out in conjunction with the internal system design. This is achieved by accommodating the effects of system integration on the structural response of the part within TO. The method is demonstrated by performing a coupled optimization on a cantilever plate with integrated components and circuitry. The results demonstrate that the method is capable of designing an optimized multifunctional part in which both the structural and system requirements are considered.*

1 INTRODUCTION

A multifunctional part, by definition, has multiple uses, such as structural and electrical functions, for example, a structural health monitoring (SHM) part. Multifunctional designs could be realized using additive manufacturing (AM) multi-material processes, and allows for a new AM design paradigm. The manufacturing processes, such as multi-head ink jet printing, capable of producing these parts are still under development, with considerable ongoing research into materials and process configuration. A variety of techniques have been proposed, primarily using stereolithography and direct write/print technologies and the reader is directed to [1] for a history of work carried out in this area. The EPSRC Centre in Innovative Manufacturing in Additive Manufacturing at the University of Nottingham, UK, has the development of multi-functional 3D printing processes, specifically multi-material jetting, as one of its main aims. The Centre also focuses on developing design optimization strategies and methods to enable this multifunctional design paradigm.

The multi-material manufacturing capability expands the possible design freedom from purely design of single material boundary geometry to also include material composition and functionality through the volume of the part. The motivation for this work lies in the realization of the ultimate aim which is to be able to intelligently optimize the design of a multifunctional part, such as the concepts included in Figure 1. Such multifunctional AM (MFAM) designs require coupling of the embedded system optimization (i.e. intelligent placement of system components and the associated routing) with a topology optimization (TO) routine (i.e. structural optimization technique that iteratively improves the material layout within a given design space, for a given set of loads and boundary conditions [2][3]). This coupling, illustrated in Figure 2, in principle, should enable in a more compact, better integrated and capable design.

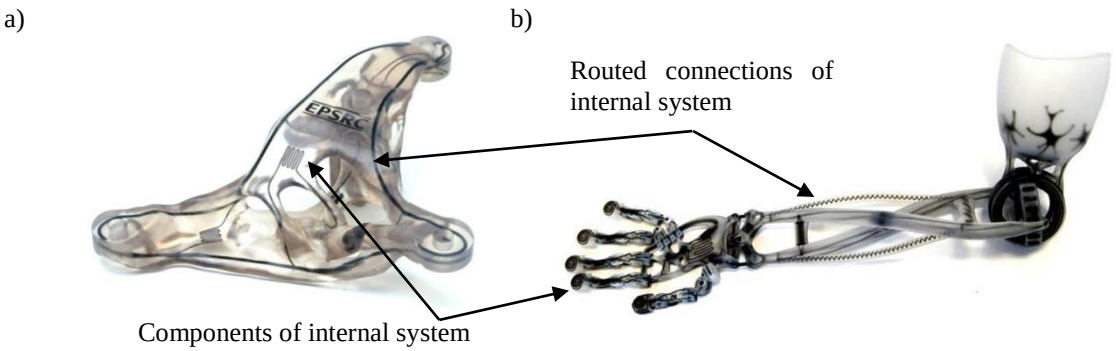


Figure 1: Multi-material jetted concept prototype - a) an example of a topologically optimized structural part with integrated internal system of placed components and the associated routing, b) a prosthetic arm with embedded systems and the associated connections between components [4].

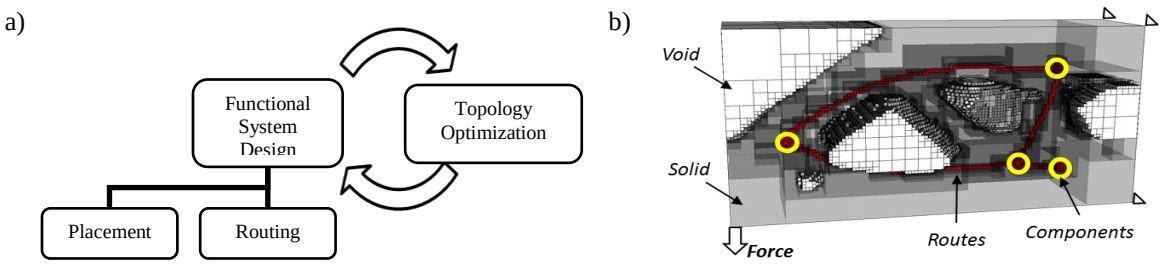


Figure 2: Coupling placement and routing optimization with structural topology optimization

The paper takes the following structure: firstly, some background information about the over-arching design framework is provided; secondly, the approach adopted for the design of functional systems is outlined; thirdly, the structure-system coupling strategy is detailed; and lastly, the appropriateness of this strategy is demonstrated by evaluating and discussing the results for example test cases.

2 BACKGROUND

The overall optimization based design framework [5] for MFAM, shown in Figure 3, has three primary strands within the structure and internal system coupling strategy: Firstly, the placement of components within the part, secondly, the routing between these components, and thirdly, accommodating the effect of integrating these components on the structural response of the part by modification of the structure using an optimization strategy. Completing the MFAM design strategy are the incorporation of design constraints and strategies employed for efficient computation of results.

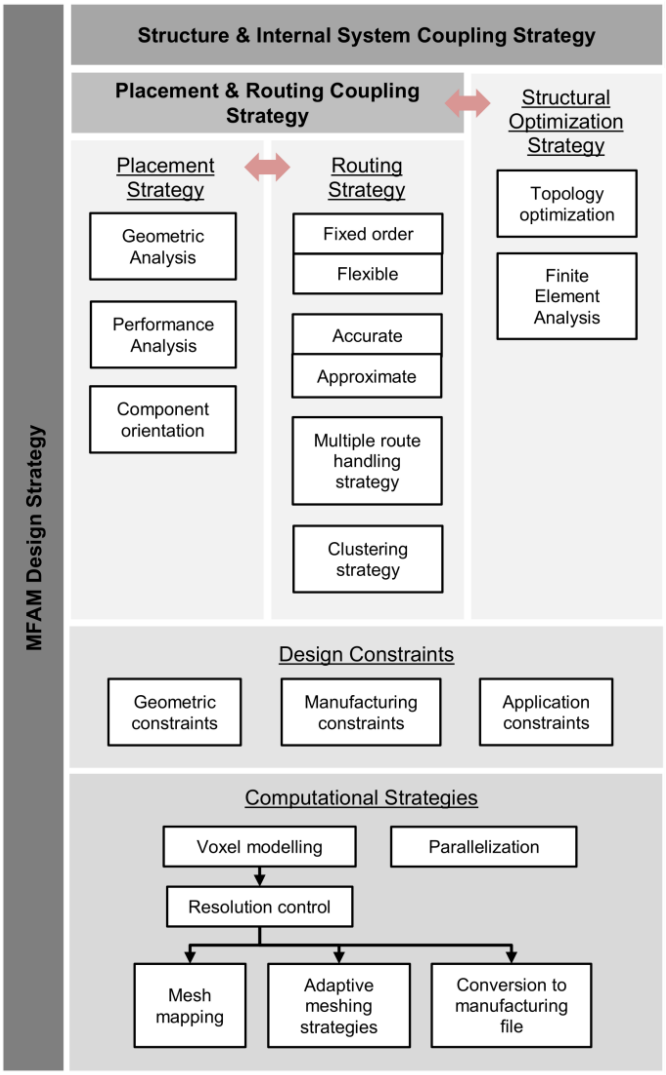


Figure 3: Overall design framework for MFAM

3 METHODOLOGY

3.1 Functional System Design

By functional system design/optimization, the authors mean the intelligent placement of components (based on some pre-determined, performance and/or geometry criterion) and generating the connections to form a circuit, commonly termed routing. Although in principle it would be best to perform placement and routing in one step as placement has significant repercussions on the routing but due to the nested dependencies these can be more efficiently (in terms of computational expense) tackled independently.

One of the key enablers, making the MFAM system design possible, is the skeletal information. This can be obtained through the process of skeletonization which is the general name given to a process which reduces the quantity of geometric information (i.e. dimensionality) required to represent a structure whilst preserving the essence of the topology. In 3D, this means a 2D medial surface and a 1D medial axis. A thinning algorithm, as detailed in [6][7], has been used to obtain the skeletal information of the part's topology. The reader is directed to the authors previous works [5] for details on how skeletal information is utilized for placement and routing within the context of MFAM design.

With regards to the system design considered herein, placement of the component is kept fixed (i.e. a set of pre-determined points) and the method of accurate routing as described in [5] is implemented. The system optimization problem therefore becomes a routing optimization problem where the aim is to improve the circuit efficiency by lowering resistance, which is proportional to the conductive track length. This is, in principle, achieved by identifying the shortest paths between components subject to design rules and constraints. By doing so, we also minimize the utilization of the conductive track material.

3.2 Coupling Strategy

The three primary strands of Figure 3 are used to formulate a coherent design procedure by devising suitable coupling strategies; specifically, the coupling between placement and routing, and the coupling between the structural optimization and the placement and routing. The first of these couplings could be achieved with a heuristic approach similar to those used for standard PCB or VLSI design [8], or a general purpose metaheuristic algorithm such ACO. Addressing the second of these couplings, is an approach that incorporates the effects of placement and routing methods via the finite element analysis (FEA) into a structural topology optimization (TO) algorithm.

Figure 4 shows the algorithm that couples TO routine (specifically, bi-directional evolutionary structural optimization (BESO) algorithm [3]) and a system optimization (specifically, placement of components and associated routing). This coupled optimization strategy is essential to fully exploit the design freedoms offered by MFAM. The main reason for the choice of BESO was the well-defined solid-void representation provided at every iteration within the TO which meant that the system optimization could be performed, for instance, at every iteration of TO.

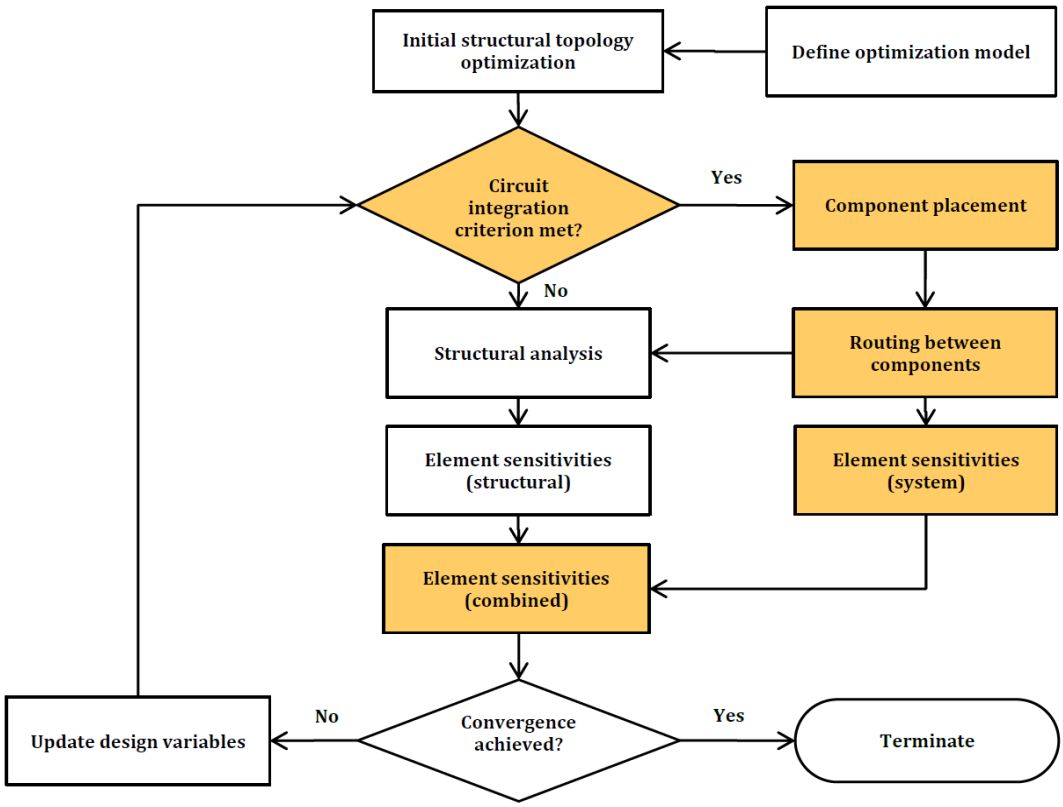


Figure 4: Flowchart showing the coupled optimization procedure

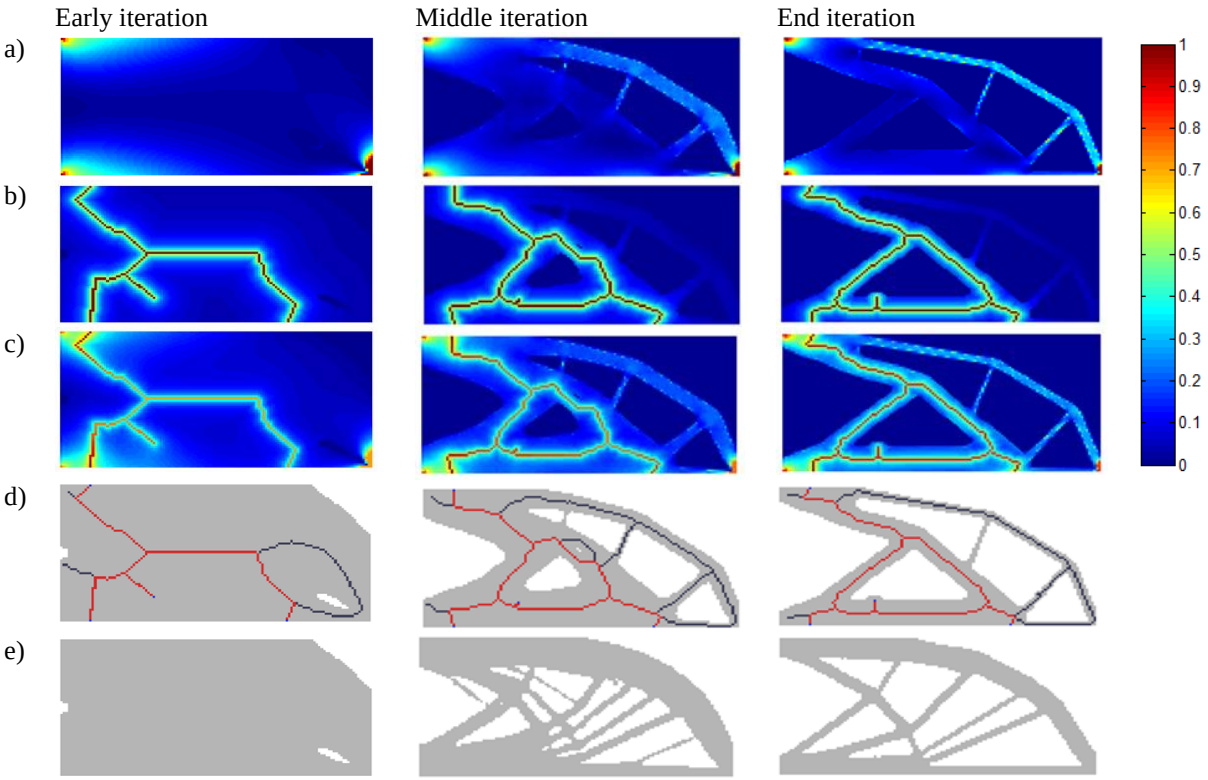


Figure 5: Effects of system integration included in the process of structural topology optimization - a) Sensitivities for structure, b) sensitivities for internal system, c) combined sensitivities, d) resulting coupled solution, and e) TO using just structural sensitivities for comparison.

This work is the natural evolution of the earlier works by the authors. The preliminary investigation of [9] (see Figure 2b) looked at integrating the system design into a structural TO algorithm such that the FEA conducted as part of TO accounted for updated material properties for regions where the components were placed and the routes were identified. Subsequently, the authors, extend this work to benefit from a bi-directional coupling between the TO and system design [10]. This is best illustrated by Figure 5 wherein we can observe the use of elemental sensitivities from both the structural and system aspect of our design to update the design variables for subsequent optimization runs. It is worth pointing out that in both these works, the routing was limited to what the authors refer to as approximate routing [5] (i.e. routes constrained to medial axis of the structure). This resulted in unstable evolution of the system configurations as the skeletal topology would drastically change due to the sudden dis-connectivity/disappearance in/of structural members. Therefore, with this work, the authors seek to test the robustness of the proposed coupling strategy (with improved heuristics) on problems employing accurate routing.

3.3 Heuristic sensitivity definition

The coupled optimization procedure of Figure 4 is built upon the principles of the revised BESO method [11][12]. The two key contributions made towards this coupled optimization algorithm are: a) heuristic definition for computing system associated elemental sensitivities and b) appropriate way of combining the system elemental sensitivities with that of the structural counterpart as shown by eq (1)

$$ {}^{C}\alpha_i = \frac{{}^{S}\alpha_i + \lambda_1 \times ({}^{R}\alpha_i)}{1 + \lambda_1} \tag{1} $$

Where, ${}^{S}\alpha_i$ represents the normalized structural elemental sensitivities (i.e. normalized strain energies) after thresholding the outliers (i.e. at point of loading and boundary conditions) and ${}^{R}\alpha_i$ represents the normalized system elemental sensitivities. λ_1 is a user defined weight to control the influence/bias of the system design on the overall coupled solution.

As this work focuses on examples where the placement location of components is predetermined/specified therefore the system associated elemental sensitivities can be determined exclusively from the routing aspect of system design using eq (2)

$$ {}^{R}\alpha_i = \frac{1}{1 + d_i} \tag{2} $$

where, d_i is the Euclidian distance between 'ith' element within the design domain and the closest point from it on the routed paths. Doing so, assigns a value of '1' to those elements which form a route and a lower value for elements that are further away from the routed paths.

Earlier works [10] (for example, see results of Figure 5) have indicated in the ${}^{R}\alpha_i$ values to be bounded and therefore ${}^{R}\alpha_i$ is set to zero for d_i greater than the filter radius (used for TO). Another improvement is the use of an adaptive parameter which is multiplied with ${}^{R}\alpha_i$ ensuring in the appropriate contributions of system associated elemental sensitivities towards the combined sensitivities (or ${}^{C}\alpha_i$).

4 SIMULATIONS, RESULTS AND DISCUSSION

In order to assess the appropriateness of the proposed coupling strategy, two test cases are considered, both addressing questions fundamental to this work:

Test Case - I. What influence does the updated heuristic sensitivity definition (discussed in section 0) has on the coupled solution when compared to the authors' earlier implementation of heuristics?

Test Case - II. Whether to perform (structure + system) coupled optimization or not? (By uncoupled problem, the authors mean performing TO to obtain a structure and subsequently include system design)

Test Cases considers a standard 2D cantilever problem with the left edge fixed and a vertically downward force being applied to the bottom right corner. The parameters used to simulate the test cases are reported in Table 1.

Parameter	Description	Value
λ	Parameter used for the single objective weighted sum formulation	1
$E_{Structure}$	Modulus of elasticity used for structure	1
E_{Void}	Modulus of elasticity used for the void region	1e-6
E_{System}	Modulus of elasticity used for system	1e-3
ν	Poisson's ratio used for all materials	0.3
R_{min}	Filter used to avoid checker-boarding	2
er	Evolution rate used for BESO	2%
Vol_{frac}	Target volume fraction used for optimization	40%
$Iter^{limit}$	Number of optimization iterations after which the process is terminated	60

Table 1: Parameters used for the coupled optimization

Results for the simulation of Test Case – I are presented below. Figure 6 shows the characteristic straight-line route between two placement locations signifying the system dominated solution even though similar weighting for structure and system contributions (i.e. lamda = 1) have been used for the older heuristic definition of [10]. On the contrary, a more representative solution for the coupled problem is obtained using the proposed heuristics (see Figure 7). The main difference between solutions of Figure 6 and Figure 7 arises due to: a) the ${}^{R}\alpha_i$ values being bounded and b) the inclusion of the adaptive parameter which ensures in the appropriate contributions of system associated elemental sensitivities towards the combined sensitivities.

Results for the simulation of Test Case – II which aims to address the most important question of this work (i.e. whether to perform (structure + system) coupled optimization or not?) is shown qualitatively as Figure 8 and presented quantitatively in Table 2. It is evident after comparing the results from Table 2 that the system design benefits the most under a coupled optimization formulation.

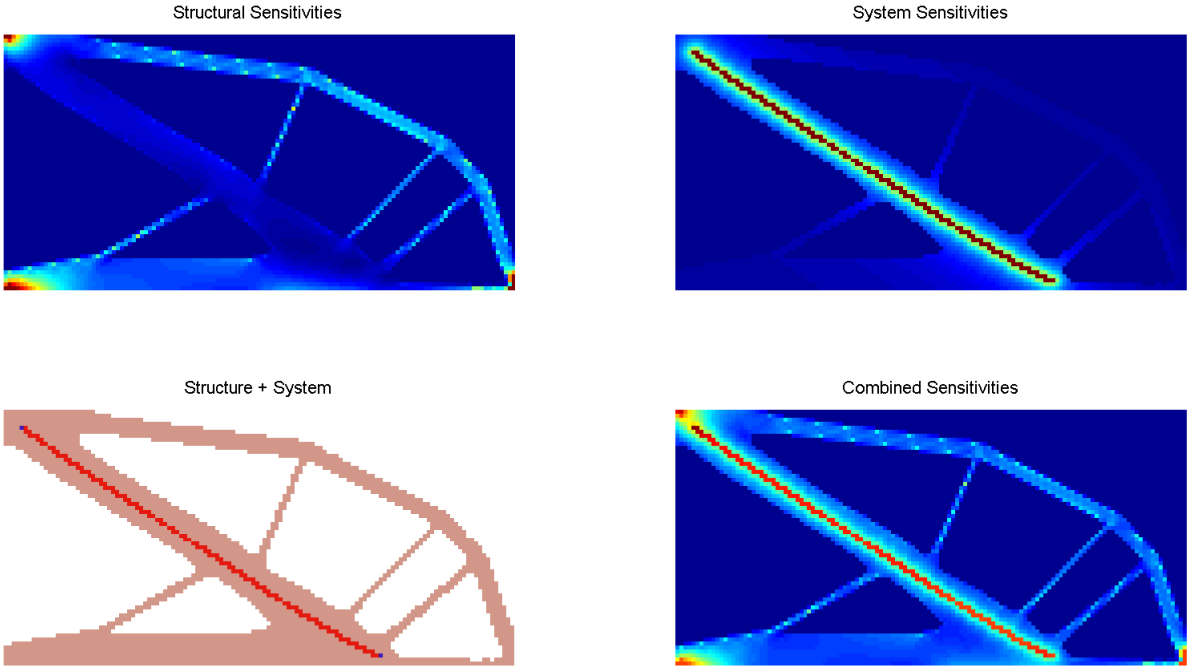


Figure 6: Results of the coupled optimization problem performed using the older heuristic definition of [10]

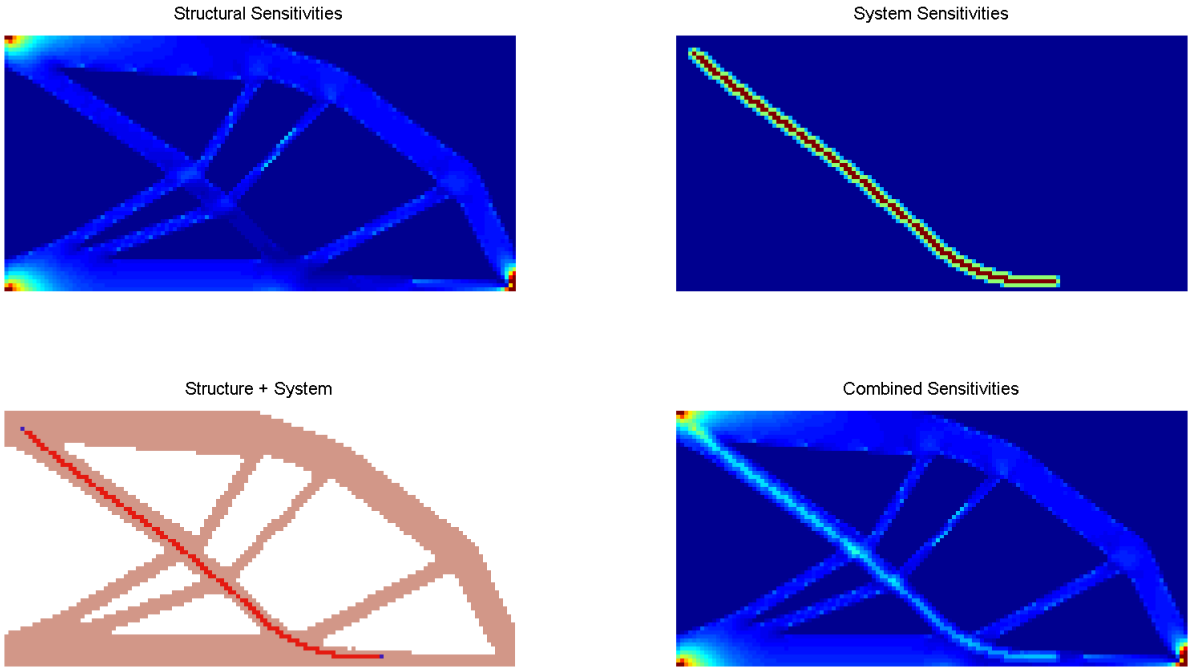


Figure 7: Results of the coupled optimization problem performed using the proposed heuristic definition

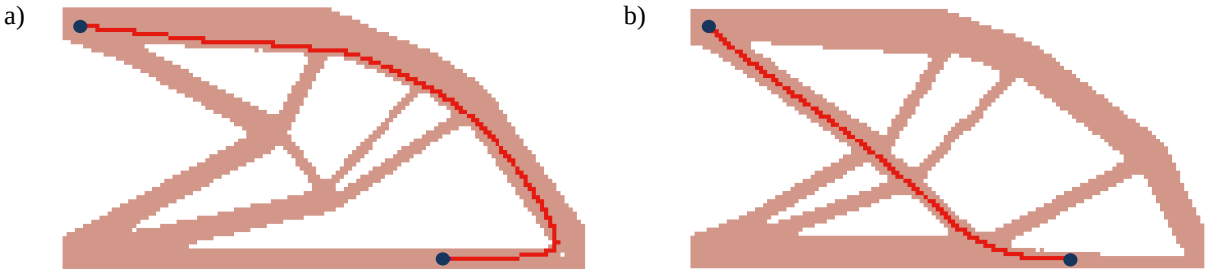


Figure 8: Comparison in results for uncoupled and coupled optimization - a) routing performed on a (converged) topology optimized structure, b) routing performed as a coupled (structure + system) optimization problem.

	Uncoupled solution	Coupled solution	difference
Path length (pixels)	196	138	30%
Total Strain Energy	39.9	40.2	<1%
Max. displacement (pixels)	76.7	77.4	<1%

Table 2: Comparison in performance between uncoupled and coupled optimization

5 CONCLUDING REMARKS

This paper has presented a coupled optimization formulation for the design of additively manufactured multi-material parts with embedded functional systems (e.g. a structural part with electronic/electrical components and associated conductive paths). This marks a significant step towards being able to exploit the design freedom offered by these manufacturing processes.

The main contribution of this paper is the improved heuristic definition that allows in a more appropriate coupling strategy where the structural optimization of a part is carried out in conjunction with the system design. This is achieved by accommodating the effects of system integration on the structural response of the part at every iteration within a modified bidirectional evolutionary structural optimization.

The simulation results for the evaluated 2D cantilever test cases show the suitability of the proposed coupling method where the system sensitivities, specifically routing sensitivities, are combined with the structural sensitivities for a multifunctional design problem. This work is testament to the need for developing the coupling strategy further as doing so would enable in the design of a more capable, better integrated and optimal multifunctional parts.

REFERENCES

[1] A. J. Lopes, E. MacDonald, and R. B. Wicker, “Integrating stereolithography and direct print technologies for 3D structural electronics fabrication,” *Rapid Prototyp. J.*, vol. 18, no. 2, pp. 129–143, 2012.

[2] M. P. Bendsoe and O. Sigmund, *Topology optimization: Theory, Methods and Applications*. 2003.

[3] X. Huang and Y. M. Xie, *Evolutionary Topology Optimization of Continuum Structures*, First edit. Wiley Publication, 2010.

[4] M. Amos, C.-W. Matthew, and S. Wimhurst, *3D Printed Prosthetic Arm*. Report, Univeristy of Noittingham, Nottingham, UK, 2013.

[5] A. Panesar, D. Brackett, I. Ashcroft, R. Wildman, and R. Hague, "Design Framework for Multifunctional Additive Manufacturing: Placement and Routing of 3D Printed Circuit Volumes," *J. Mech. Des.*, vol. 137, no. 11, 2015.

[6] T. Lee and R. Kashyap, "Building skeleton models via 3-D medial surface / axis thinning algorithms," *Graph. Model. Image Process.*, vol. 56, no. 6, pp. 462–478, 1994.

[7] M. Kerschnitzki, P. Kollmannsberger, M. Burghammer, G. N. Duda, R. Weinkamer, W. Wagermaier, and P. Fratzl, "Architecture of the osteocyte network correlates with bone material quality," *J. Bone Miner. Res.*, vol. 28, no. 8, pp. 1837–1845, 2013.

[8] N. Abboud, M. Grötschel, and T. Koch, "Mathematical methods for physical layout of printed circuit boards: an overview," *OR Spectr.*, vol. 30, no. 3, pp. 453–468, May 2007.

[9] A. Panesar, D. Brackett, I. Ashcroft, R. Wildman, and R. Hague, "Design Optimization Strategy for Multifunctional 3D Printing," in *25th Interational Solid Freeform Fabrication Symposium*, 2014, pp. 592–605.

[10] A. Panesar, D. Brackett, I. Ashcroft, R. Wildman, and R. Hague, "Design optimization for multifunctional 3D printed structures with embedded functional systems," in *11th World Congress on Structural and Multidisciplinary Optimisation*, 2015.

[11] X. Huang and Y. M. Xie, "Bi-directional evolutionary topology optimization of continuum structures with one or multiple materials," *Comput. Mech.*, vol. 43, no. 3, pp. 393–401, Jul. 2008.

[12] X. Huang and Y. M. Xie, "A further review of ESO type methods for topology optimization," *Struct. Multidiscip. Optim.*, vol. 41, no. 5, pp. 671–683, Mar. 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SHAPE-TOPOLOGY OPTIMIZATION FOR DESIGNING SHELL STRUCTURES

Hirotaka Nakayama[1], **Masatoshi Shimoda**[2]

[1] Graduate School of Toyota Technological Institute
Nagoya, Japan
tontenq.ucd080@gmail.com

[2] Toyota Technological Institute
Nagoya, Japan
shimoda@toyota-ti.ac.jp

Keywords: Shape, Topology, Optimization, Simultaneous, Shell Structure, SIMP Method, H^1 Gradient Method.

Abstract. *In this study, we propose a new approach for creating the optimal shape and topology of shell structures. By implementing topology optimization in the variable design domain which is optimized by shape optimization at every iteration, the optimal topology and shape is simultaneously determined. The free-form optimization method and the SIMP method for shells are used for shape and topology optimization, respectively. Compliance is minimized under both volume constraints for shape and topology optimization, and out-of-plane shape variation and fictitious density are used as the design variable. The optimum design problem is formulated as a distributed-parameter optimization problem, and the sensitivity functions for shape variation and density are theoretically derived. Both the optimal the optimal shape variation and density distribution are determined by using the H^1 gradient method, where the sensitivity functions are applied under the Robin condition to vary the shape and density. With the proposed method, the compliance is minimized, while maintaining smooth surface and the density distribution. The results show that the proposed simultaneous optimization method provides stiffer and lighter structures with less numerical instabilities.*

1 INTRODUCTION

The researches in the field of shape and topology optimization method has been active in recent years, and successfully applied to optimize structures in many practical automotive or aerospace engineering problems. In the shape optimization problems, the main subject is to find the optimal geometry with the given loading and boundary conditions. In the area of shape optimization for shell structures, the optimization design problem is divided into in-plane and out-of-plane. The former design is used to determine the detail shape or realize weight reduction by changing boundaries. Thus, this boundary shape optimization is usually applied the last stage in design procedure. On the other hand, the latter design has great impact to improve stiffness or some performances by changing surfaces. As the shape of the structure is optimized by varying only the existing boundaries or surfaces of the geometry, it is not able to remove or introduce new boundaries in the design domain. Hence, a general need for optimization methods which is able to allow for a change of the material layout in the design domain.

Topology optimization has been primarily applied to structural problems to obtain lighter structures with optimal material layout. SIMP (Solid Isotropic Material with Penalization) method [1] has been wildly used for topology optimization. This approach is to seek density distribution in the design domain, where the elastic tensor of the intermediate density of material is penalized with exponential function of density.

With a conventional topology optimization method for shell structures, it cannot be expected to improve the stiffness because the material layout is determined only in the fixed design domain. On the other hand, shape optimization for shell surface can largely improve the stiffness, but cannot create new holes. Therefore, in this study, we proposed an approach which can create stiffer and lighter shell structures by implemented topology optimization in varying design domain of a shell structure which is optimized by shape optimization.

Some early research for combined shape and topology optimization can be found in papers by Bremicker et al. [2] or Schwarz et al. [3]. Note that the procedure was to find optimal topology first and obtained structure was dealt as the initial design, then boundary shape optimization was additionally applied within obtained design domain in order to increase the performances. Thus the obtained boundaries by topology optimization are modified by shape optimization. Ikeya et al. [4] extended the two-phase optimization method to produce stiffer and lighter shell by carrying out surface shape optimization first and then size optimization. This approach intended to obtain a weight reduced shell with high stiffness by moving shell surface, but the optimization methods are considered separately and it can't take into account the coupling effect of shape and sizing parameters. In other words, these researches are two steps optimization method considering integrated shape and topology or sizing optimization. Ansola et al. [5, 6] and Hassani et al. [7] introduced a simultaneous shape and topology optimization method for shell structures. In their methods, material layout is determined while shape is varying. However, the obtained structures depend on the initial setting of design variables due to the use of parametrization method for shape optimization.

In this study, H^1 gradient method [8-11], a gradient method in a function space, is employed in order to determine the optimal surface for shape optimization and to overcome numerical instabilities for topology optimization while minimizing an objective functional. In our previous work, we developed a free-form optimization method for determining the dynamically natural and optimal shell form. In this work, we apply this method to shape-topology optimization method, and the influence of the method is investigated.

The stiffness maximization problem is solved as a compliance minimization problem under volume constraints for both shape and topology optimization. Shape variation and fictitious

density are optimized by applying the sensitivity functions. The simultaneous optimization problem is solved by means of integrated iterative algorithms, which is simultaneously implemented shape and topology optimization.

2 FORMULATION OF OPTIMIZATION PROBLEM

2.1 Governing equation for a shell structure

As shown in Fig. 1 (a), a shell has an initial design domain Ω , mid-surface A with the boundary of ∂A , side surface S and thickness t. It is assumed for simplicity that a shell structure occupying a bounded domain is a set of infinitesimal flat surfaces. The notations $\{u_{0\alpha}\}_{\alpha=1,2}$, w and $\{\theta_\alpha\}_{\alpha=1,2}$ express the in-plane displacements, out-of-plane displacement and rotational angles of the mid-surface of the shell, respectively in Fig 1 (b). In this paper, the Mindlin-Reissner plate theory is applied concerning plate bending.

The weak form of the state equation with respect to $(\boldsymbol{u}_0,w,\boldsymbol{\theta})\in U$ can be expressed as:

$$a\left((\boldsymbol{u}_0,w,\boldsymbol{\theta}),(\bar{\boldsymbol{u}}_0,\bar{w},\bar{\boldsymbol{\theta}})\right)=l(\bar{\boldsymbol{u}}_0,\bar{w},\bar{\boldsymbol{\theta}}),\ (\boldsymbol{u}_0,w,\boldsymbol{\theta})\in U \tag{1}$$

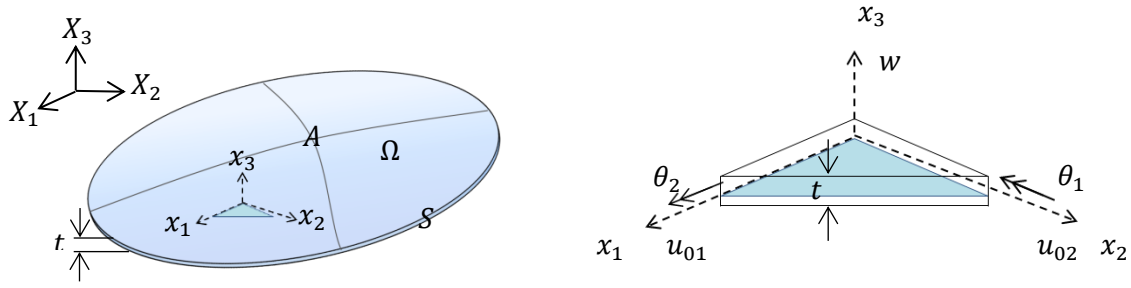
where the energy bilinear form $a(\cdot,\cdot)$ and the linear form $l(\cdot)$ for the state variables $(\boldsymbol{u}_0,w,\boldsymbol{\theta})$ are respectively defined as:

$$a\left((\boldsymbol{u}_0,w,\boldsymbol{\theta}),(\bar{\boldsymbol{u}}_0,\bar{w},\bar{\boldsymbol{\theta}})\right)=\int_\Omega\left\{C_{\alpha\beta\gamma\delta}(u_{0\alpha,\beta}-x_3\theta_{\alpha,\beta})(\bar{u}_{0\gamma,\delta}-x_3\bar{\theta}_{\gamma,\delta})+C^s_{\alpha\beta}(w_{,\alpha}-\theta_\alpha)(\bar{w}_{,\alpha}-\bar{\theta}_\alpha)\right\}d\Omega$$

$$\left(\ =\int_A\left\{c^B_{\alpha\beta\gamma\delta}\kappa_{\gamma\delta}\bar{\kappa}_{\alpha\beta}+c^M_{\alpha\beta\gamma\delta}\varepsilon_{0\gamma,\delta}\bar{\varepsilon}_{0\alpha,\beta}+kc^s_{\alpha\beta}\gamma_\alpha\bar{\gamma}_\beta\right\}dA\ \right) \tag{2}$$

$$l(\bar{\boldsymbol{u}}_0,\bar{w},\bar{\boldsymbol{\theta}})=\int_A(f_\alpha\bar{u}_{0\alpha}-m_\alpha\bar{\theta}_\alpha+q\bar{w})dA+\int_A t(b_\alpha\bar{u}_{0\alpha}+b_3\bar{w})dA+\int_{\partial A}(N_\alpha\bar{u}_{0\alpha}-M_\alpha\bar{\theta}_\alpha+Q\bar{w})dS \tag{3}$$

It should be noted that U in Eq. (1) is Sobolev space that satisfies the given Dirichlet conditions on each subboundary. In this paper, the subscripts of the Greek letters are expressed as $\alpha=1,2$, and the tensor subscript notation uses Einstein's summation convention and a partial differential notation. Loads acting relative to the local coordinate system $(x_1,x_2,0)$ are defined as: an out-of-plane load q, an in-plane load $\boldsymbol{f}=\{f_\alpha\}_{\alpha=1,2}$ and an out-of-plane moment $\boldsymbol{m}=\{m_\alpha\}_{\alpha=1,2}$, an in-plane load $\boldsymbol{N}=\{N_\alpha\}_{\alpha=1,2}$, a shearing force Q, a bending moment $\boldsymbol{M}=\{M_\alpha\}_{\alpha=1,2}$ and a body force $h\boldsymbol{b}=\{hb_i\}_{i=1,2,3}$ are considered as the external forces. In addition, $\{C_{\alpha\beta\gamma\delta}\}_{\alpha,\beta,\gamma,\delta=1,2}$ and $\{C^S_{\alpha\beta}\}_{\alpha,\beta=1,2}$ express an elastic tensor including bending and membrane stresses, and an elastic tensor with respect to the shearing stress, respectively. $\{c^B_{\alpha\beta\gamma\delta}\}_{\alpha,\beta,\gamma,\delta=1,2}$, $\{c^S_{\alpha\beta}\}_{\alpha,\beta=1,2}$ and $\{c^M_{\alpha\beta\gamma\delta}\}_{\alpha,\beta,\gamma,\delta=1,2}$ express orthotropic elastic tensors with respect to bending, shear and membrane component, respectively. The constants k expresses a shear correction factor (assuming k=5/6). The notations $\{\kappa_{\alpha\beta}\}_{\alpha,\beta=1,2}$ and $\{\gamma_\alpha\}_{\alpha=1,2}$ express the curvatures and the transverse shear strains.



(a) Geometry of shell (b) Local coordinates and DOF of flat surface

Figure 1: Shell as a set of infinitesimal flat surfaces.

2.2 Compliance minimization problem

In this paper, compliance minimization problem under volume constraints is considered. Letting the volume and the state equations be the constraint conditions, and the compliance be the objective functional to be minimized, a distributed-parameter optimization problem for finding the optimal design velocity field $\boldsymbol{V}$ and density field ρ can be formulated as:

$$\textit{Given} \qquad A,\ \hat{M}_S,\ \hat{M}_T \tag{4}$$

$$\textit{Find} \qquad \boldsymbol{V},\ \rho \tag{5}$$

$$\textit{Minimized} \qquad l(\boldsymbol{u}_0, w, \theta) \tag{6}$$

$$\textit{Subject to} \qquad M_S = \int \rho t dA \ \le \hat{M}_S\ ,\ M_T = \int \rho d\Omega \ \le \hat{M}_T \ \text{and Eq. (1)} \tag{7}$$

where M and $\hat{M}$ denote the volume and its constraint value, respectively.

2.3 Sensitivity analysis

Here, the Lagrange multiplier method is used to transform this constrained shape and topology optimization problem to the non-constrained one. Letting $(\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})$ and Λ denote the Lagrange multipliers for the state equation and the volume constraints, respectively, the Lagrange functional L associated with this problem can be expressed as:

$$L\left(\Omega, (\boldsymbol{u}_0, w, \boldsymbol{\theta}), (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}}), \Lambda\right) = l(\boldsymbol{u}_0, w, \boldsymbol{\theta}) + l(\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}}) - a\left((\boldsymbol{u}_0, w, \boldsymbol{\theta}), (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})\right)$$
$$+ \Lambda(M - \hat{M}_S) + \Lambda(M - \hat{M}_T) \tag{8}$$

Using the design velocity field $\boldsymbol{V}$ and density ρ to represent the amount of domain variation, the material derivative $\dot{L}$ of the Lagrange functional L can be expressed as:

$$\dot{L} = l(\boldsymbol{u}_0', w', \boldsymbol{\theta}') + l(\bar{\boldsymbol{u}}_0', \bar{w}', \bar{\boldsymbol{\theta}}') - a\left((\boldsymbol{u}_0', w', \boldsymbol{\theta}'), (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})\right) - a\left((\boldsymbol{u}_0, w, \boldsymbol{\theta}), (\bar{\boldsymbol{u}}_0', \bar{w}', \bar{\boldsymbol{\theta}}')\right)$$
$$+ \Lambda'(M_S - \hat{M}_S) + \Lambda'(M_T - \hat{M}_T) + \left\langle G\boldsymbol{n}, \boldsymbol{V}\right\rangle_S + \left\langle G_\rho, \rho\right\rangle_T \quad ,\ \boldsymbol{V} \in C_\Theta \tag{9}$$

$$\left\langle G\boldsymbol{n}, \boldsymbol{V}\right\rangle_S = \int_A G\boldsymbol{n} \cdot \boldsymbol{V} dA = \int_A G_A V_n dA + \int_A G_f V_n dA \tag{10}$$

$$\left\langle G_\rho, \rho\right\rangle_T = \int_A G_\rho \delta\rho dA \tag{11}$$

where C_Θ is the suitably smooth function space that satisfies the constraints of the domain variation and notation $\boldsymbol{n}$ denotes the unit normal vector to the shell surface.

KKT (Karush-Kuhn-Tucker) optimality conditions for this problem are expressed as follows:

$$a\left((\boldsymbol{u}_0, w, \boldsymbol{\theta}), (\bar{\boldsymbol{u}}_0', \bar{w}', \bar{\boldsymbol{\theta}}')\right) = l(\bar{\boldsymbol{u}}_0', \bar{w}', \bar{\boldsymbol{\theta}}'), \quad (\boldsymbol{u}_0, w, \boldsymbol{\theta}) \in U \ , \ \forall (\bar{\boldsymbol{u}}_0', \bar{w}', \bar{\boldsymbol{\theta}}') \in U \tag{12}$$

$$a\left((\boldsymbol{u}_0', w', \boldsymbol{\theta}'), (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})\right) = l(\boldsymbol{u}_0', w', \boldsymbol{\theta}'), \quad (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}}) \in U \ , \ \forall (\boldsymbol{u}_0', w', \boldsymbol{\theta}') \in U \tag{13}$$

$$\Lambda(M - \hat{M}_S) = 0, \quad \Lambda \geq 0, \quad M - \hat{M}_S \leq 0 \tag{14}$$

By assuming the KKT optimality conditions hold true, the shape and density sensitivity functions G_A and G_ρ for this problem are derived as:

$$G_A = -C_{\alpha\beta\gamma\delta}(u_{0\alpha,\beta} - \frac{t}{2}\theta_{\alpha,\beta})(u_{0\gamma,\delta} - \frac{t}{2}\theta_{\gamma,\delta}) + C_{\alpha\beta\gamma\delta}(u_{0\alpha,\beta} + \frac{t}{2}\theta_{\alpha,\beta})(u_{0\gamma,\delta} + \frac{t}{2}\theta_{\gamma,\delta}) + \Lambda t H \tag{15}$$

$$G_\rho = \frac{\partial c^B_{\alpha\beta\gamma\delta}}{\partial \rho}\theta_{\gamma,\delta}\theta_{\alpha,\beta} + \frac{\partial c^M_{\alpha\beta\gamma\delta}}{\partial \rho}u_{0\gamma,\delta}u_{0\alpha,\beta} + \kappa\frac{\partial c^S_{\alpha\beta}}{\partial \rho}\gamma_\alpha\gamma_\beta + \Lambda \tag{16}$$

where H is twice the mean curvature of mid-area A. In the derivation of G_A and G_f , we assumed that the loading boundaries and surfaces were not varied in the optimization process.

In the SIMP method, the elastic tensor of an intermediate density can be expressed as:

$$C_{\alpha\beta\gamma\delta} = C^0_{\alpha\beta\gamma\delta}\rho^p \tag{17}$$

where $C^0_{\alpha\beta\gamma\delta}$ is the initial value of the elastic tensor of material and p is the penalization factor.

3 SIMULTANEOUS OPTIMIZATION

In the proposed simultaneous shape and topology optimization method, the optimal design velocity field $\boldsymbol{V}$ and the optimal density field ρ are determined by using H^1 gradient method. With the proposed method, the update shape and topology of a shell structure is simultaneously implemented without seeing parameter and with smooth surface, while minimizing the objective functional.

3.1 Optimization algorithm

Fig. 2 shows a flowchart of the simultaneous optimization method. At the first step, the equilibrium equation in Eq. (1) is solved using the FEM. In the second step, shape and density sensitivity functions are calculated by Eq. (15) and (16). In the third step, give virtual loading and internal heat generation based on shape and density functions, and then pseudo displacement and temperature field are obtained, respectively. In the fourth step, the equilibrium equation and Poisson's equation are resolved using the virtual loading and internal heat generation, respectively. This third and fourth step is procedure of H^1 gradient method to determine the smooth shell surface and smooth density field. By applying H^1 gradient method, design velocity field $\boldsymbol{V}$ and density field ρ are updated. After that, this optimization procedure is repeated until the calculated objective function and volumes are converged.

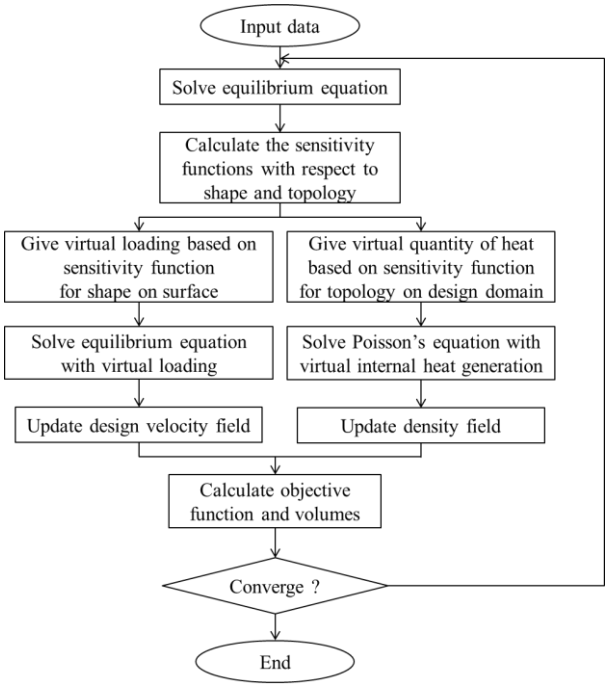


Figure 2: Flowchart of simultaneous optimization method.

3.2 H^1 gradient method for shell structures

The original H^1 gradient method under the Robin condition was introduced by Azegami [8], and Shimoda [10] modified it for free-form shell optimization. To obtain optimal design velocity field $\boldsymbol{V}$, the following weak formed governing equation is introduced.

$$a\left((\boldsymbol{V}_{0\alpha}, V_3, \ \boldsymbol{\theta}), (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})\right) + \alpha\left\langle (\boldsymbol{V}\cdot\boldsymbol{n})\boldsymbol{n}, (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})\right\rangle = -\left\langle G\boldsymbol{n}, (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}})\right\rangle,$$

$$(\boldsymbol{V}_{0\alpha}, V_3, \ \boldsymbol{\theta}) \in C_\Theta, \forall (\bar{\boldsymbol{u}}_0, \bar{w}, \bar{\boldsymbol{\theta}}) \in C_\Theta \qquad (18)$$

where α is distributed spring constant.

In this paper, we apply this method to topology optimization. The concept illustration is shown in Fig. 3. In the SIMP method, it causes the numerical instabilities [12] such as gray scale, checkerbording pattern and mesh dependency problem because material layout is expressed by density distribution. The proposed H^1 gradient method for topology optimization is also effective to overcome these instabilities.

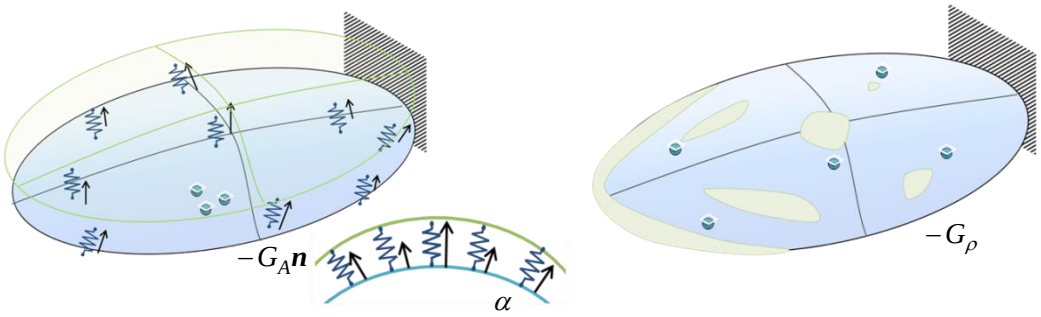
To obtain the optimal density field ρ , the following weak formed Poisson's equation, or the governing equation for ρ is introduced.

$$a(\rho, v) + \varphi_\rho \langle \rho - \rho_0, v\rangle = -\left\langle G_\rho, \Delta s\rho\right\rangle, \quad \rho \in C_\rho, \ \forall v \in C_\rho \qquad (19)$$

$$a(\rho, v) = \int_A \rho_{,i} k_{ij} v_{,j} dA \qquad (20)$$

where $\Delta s(>0)$ is sufficiently small constant. ρ_0 and C_ρ denote the reference density and the kinematic admissible function space that satisfies Dirichlet condition for density variation, respectively. The notations φ_ρ and k_{ij} are equivalent to the heat transfer coefficient and the thermal conductivity tensor in the steady heat transfer equation, respectively.

In this method, $-G_\rho$ is applied as internal heat generation to the design domain on a pseudo-elastic shell. The density field ρ is calculated as the pseudo-temperature distribution of Poisson's equation and is used to update the original density.



(a) For shape optimization (b) For topology optimization

Figure 3: Concept illustrations of H^1 gradient method.

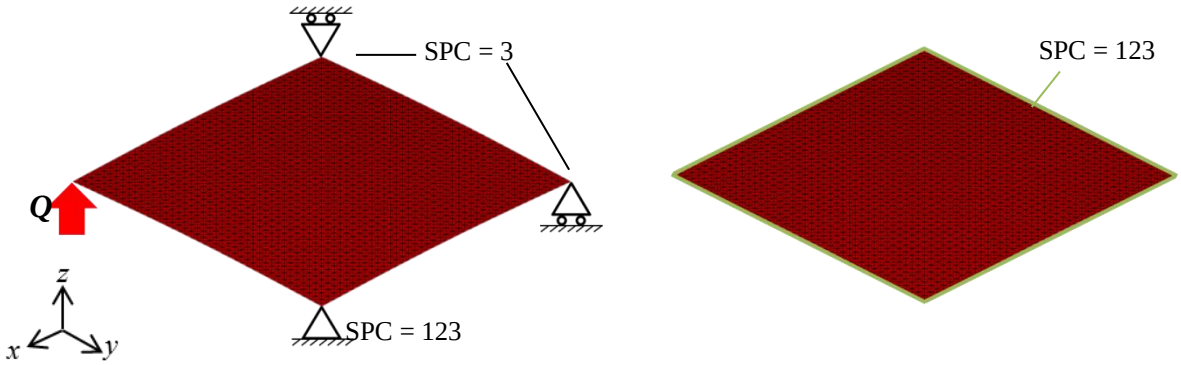
4 NUMERICAL EXAMPLES

In this section, two examples are considered to demonstrate the validity and effectiveness of the simultaneous shape and topology optimization method. In all cases, the material constants are used as the Young's module $E = 210$ GPa, the transverse elasticity module $G = 80.7$ GPa and Poisson ratio $\nu = 0.3$. Furthermore, minimum value of the normalized density is set $\rho_{\min} = 0.01$ to avoid the numerical error.

4.1 Torsion plate problem

Fig. 4 shows the design domain for a square plate. In the stiffness analysis, the plate is supported on three corners and upward force is applied on the other corner in Fig. 4 (a). In the velocity analysis, all edges are simply supported as shown in Fig. 4 (b). The volume constraints of shape and topology are set to 105% and 50% of the initial value, respectively. The material of the shell is modeled using power law SIMP model with a penalty factor of $p = 5$ and the optimization is carried out using 6400 elements with triangle mesh.

Fig. 5 shows the optimal result. As can be seen, a net-shaped topology is created on the X-shaped surface. Fig. 6 shows the iterations history of the compliance and the volumes along with the configurations. Note that the value of compliance is stably reduced by approximately 82% compared with the initial value.



(a) Stiffness analysis (b) Velocity analysis

Figure 4: Boundary conditions for a square plate under torsion.

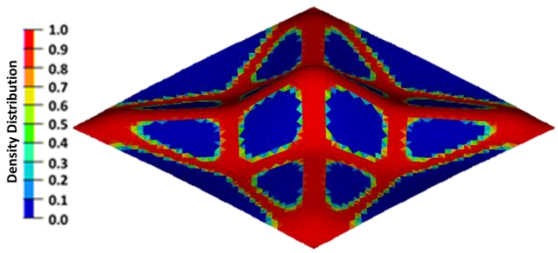


Figure 5: Optimal shape and material layout.

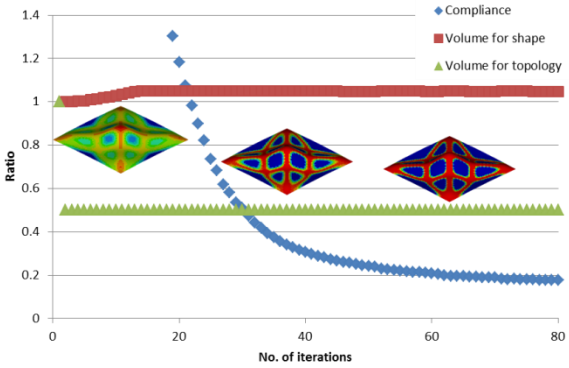


Figure 6: Iterations history of torsion plate.

4.2 Link part problem

The proposed method is applied to a link part model. The initial shape and the problem definition are illustrated in Fig. 7. In the stiffness analyses shown in Fig. 7 (a), the bending moments are applied at the two joints and the other joint is supported. In the velocity analysis shown in Fig. 7 (b), it is assumed that all joints are supported. The volume constraints of shape and topology are set to 100% and 70% of the initial value, respectively. The penalty parameter is set $p = 5$ and the triangle mesh is created by auto-mesh function of HyperMesh [12].

The obtained optimal shape and topology is illustrated in Fig. 8. As it is observed, the constraint area is expanded in the x-direction while the side, bottom and curve area are vertically narrowed, respectively. The material is distributed on the corners and the holes are created on the curve and narrowed area due to reduction the volume. The iterations history is illustrated in Fig. 9. The compliance is gradually decreased by approximately 66% under the constant volume. Hence, the proposed method is successful for finding the optimal shape and material layout.

(a) Stiffness analysis (b) Velocity analysis

Figure 7: Boundary conditions for link part model under bending moment.

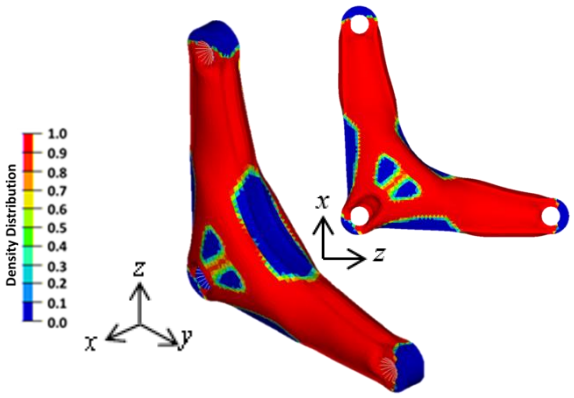


Figure 8: Optimal shape and material layouts

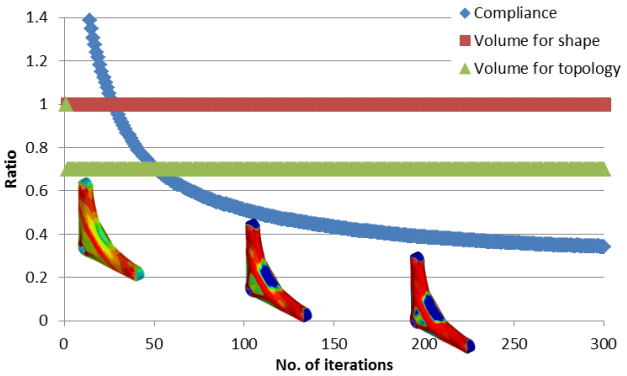


Figure 9: Iterations history of link part model

5 CONCLUSION

This paper described simultaneous shape and topology optimization method for shell structures to minimize the compliance under both shape and topology volume constraints. From a view of stiffness, the conventional topology optimization in a fixed design domain causes the stiffness reduction from the initial solid condition due to the reduction the volume constraint. To overcome this problem, varying design domain is introduced, where the optimal density distribution is determined on varied design domain determined by the shape optimization. The proposed simultaneous optimization method provided lighter and stiffer shell structures where both shape and topology optimization was carrying out concurrently. In the shape optimization problem, the free-form optimization method was applied to vary the shape surface. For the topology optimization problem, the SIMP method was employed to obtain the density distribution. The shape and topology sensitivity functions were derived from governing equation for shell structures and applied to each method. Using H^1 gradient method with sensitivity analysis to optimize shape and material layout, smooth surface was obtained and numerical instabilities for topology were thoroughly solved while reducing the objective functional without shape and topology parameterization.

REFERENCES

[1] M. Bendsøe, Optimal shape design as a material distribution problem, Struct. Optim, 1, 193-202, 1989.

[2] M. Bremicker, M. Chirehdast, N. Kikuchi and P. Y. Palambros, Integrated topology and shape optimization in structural design, Mech. Struct. & Mach., 19 (4), 551-587, 1991.

[3] S. Schwarz, K. Maute and E. Ramm, Topology and shape optimization for elastoplastic structural response, Comput. Methods Appl. Mech. Engrg. 190, 2135-2155 , 2001.

[4] K. Ikeya, M. Shimoda and J. X. Shi, Multi-objective free-form optimization for shape and thickness of shell structure, Composite Structures, 135, 262-275, 2016.

[5] R. Ansola, J. Canales, J.A. Tarrago and J. Rasmussen, An integrated approach for shape and topology optimization of shell structures, Computers and Structures, 80, 449-458, 2002.

[6] R. Ansola, J. Canales, J.A. Tarrago and J. Rasmussen, Combined shape and reinforcement layout optimization of shell structures, Struct Multidisc Optim, 27, 219-227, 2004.

[7] B. Hassani, M. S. Tavakkoli and H. Ghasemnejad, Simultaneous shape and topology optimization of shell structures, Struct Multidisc Optim, 48, 221-233, 2013.

[8] H. Azegami, A Solution to Domain Optimization Problems, Trans. of Jpn. Soc. of Mech. Engs., 60, 1479-1486, 1994.

[9] S. Riehl and P. Steinmann, A staggered approach to shape and topology optimization using the traction method and an evolutionary-type advancing front algorithm, Comput. Methods Appl. Mech. Engrg., 287, 1-30, 2015.

[10] M. Shimoda and Y. Liu, A Non-parametric free-form optimization method for shell Structures, Structural and Multidisciplinary Optimization, 50, 409-423, 2014.

[11] H. Azegami, S. Kaizu and K. Takeuchi, Regular solution to topology optimization problems of continua, JSIAM Letters, 3, 1-4, 2011.

[12] O. Sigmund, Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima, Structural Optimization, 16, 68-75, 1998.

[13] http://www.altairhyperworks.com/product/HyperMesh

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON A NASH GAME FOR TOPOLOGY OPTIMIZATION UNDER LOAD-UNCERTAINTY – FINDING THE WORST LOAD

Carl-Johan Thore[1]

[1]Division of Solid Mechanics, Department of Management and Engineering, Linköping University
Linköping University, 581 83 Linköping, Sweden
e-mail: carl-johan.thore@liu.se

Keywords: Worst load, Global optimization, Branch and Bound, Topology optimization, Nash game

Abstract. *Topology optimization (TO) enables creation of structures which are highly efficient with respect to precisely those conditions formulated mathematically in the optimization problem. To ensure that optimized structures are robust with respect to variations in geometric and material parameters as well as external loads it is therefore necessary to set up problem formulations in which such uncertain variations are accounted for. The author and co-workers have recently proposed to formulate TO under load-uncertainty as a two-person game in which the two players are the "designer", which controls the design variables, and "nature", which controls variables parametrizing the external loads. For this game a Nash-equilibrium is then sought which solves a standard TO problem to find the best design for a given load and an optimization problem – the "load-problem" – to find the "worst" load(s) for a given design (a problem of obvious interest also on its own). This work focuses on the load-problem – a non-convex problem with a relatively small number of variables and simple constraints – and investigates the practical possibility of solving it to global optimality using a simplicial branch-and-bound algorithm. The algorithm is tested on a topology optimized design to which a number of point loads are applied and allowed to vary in magnitude and direction in order to maximize a smooth approximation of the maximum von Mises stress. Numerical results show promise, but also that more work is needed to be able to solve large-scale (in terms of both the size of the structural model and the number of independent loads) instances of the load-problem to global optimality.*

1 INTRODUCTION

Topology optimization (TO) can be used to design load-carrying structures that are highly efficient with respect to *precisely* those conditions formulated mathematically in the optimization problem. Deviations, uncertain and inevitable in practise, from those conditions may cause optimized structures to behave erratically or even collapse. This leads to the idea of robust TO wherein uncertainties are accounted for explicitly in the optimization problem, either by assuming stochastic variations of problem parameters or using a *worst-case* approach [1]. Here we consider the latter approach and its use under load-uncertainty. This paper therefore concerns the problem of finding the "worst" load(s) for a given functional f; i.e. the load(s) which maximize (or minimize) f. The setting, studied by the author and co-workers in [2, 3, 4], is a two-player Nash game formulation for robust TO wherein a "designer" controls the design variables in $\boldsymbol{x}$ and "nature" the load variables in $\boldsymbol{r}$. Assuming the same pay-off functional f for both players a Nash equilibrium $(\boldsymbol{x}^*, \boldsymbol{r}^*)$ is defined by

$$\boldsymbol{x}^* \in \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\min}\, f(\boldsymbol{x}, \boldsymbol{r}^*) \tag{1a}$$

$$\boldsymbol{r}^* \in \underset{\boldsymbol{r}\in\mathcal{T}}{\arg\max}\, f(\boldsymbol{x}^*, \boldsymbol{r}) \tag{1b}$$

where $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{T} \subset \mathbb{R}^d$ is the strategy set of respective player. Ideally (and suggested by the notation), $\boldsymbol{x}^*$ and $\boldsymbol{r}^*$ should be globally optimal solutions in their respective problem. Unfortunately the (standard) TO problem (1a) typically has a very large number of variables and solving it to global optimality is currently not a viable option. In problem (1b) – the "load-problem" – however, the design is fixed and a useful parametrization of the load can often be obtained using a relatively small number of variables. Obtaining a globally optimal solution is also in one sense more important in (1b) than in (1a) as it ensures robustness of the optimized design; i.e., that there are no other loads realizable by the given parametrization that result in worse performance. For these reasons the focus here is the load-problem, for which globally optimal solutions are sought.

If f is the so-called *compliance* and the parametrization of the load is suitably chosen the load-problem is readily solved: the worst load can be obtained by solving an eigenvalue problem [5] or a generalized eigenvalue problem [6]. Here however, we let f be an ℓ_p-norm approximation of the maximum von Mises stress. In this case no "closed-form" solution is, to the author's knowledge, available and the load-problem has to be treated as a non-convex, non-linear optimization problem (NLP). We propose to solve this problem to global optimality by a branch and bound algorithm rather than some type of heuristic method (such as gradient-based multistart or genetic algorithms). The obvious advantage of this choice is that we get a guarantee of the quality of the solution in the sense that the computed optimal value deviates at most a user-specified number ε from the globally optimal value.

Related work in solid mechanics includes [7] and [8] which describe the use of exact global optimization to evaluate a fatique criterion and the minimum limit load factor in worst case plastic analysis, respectively. Branch and bound has also been used for design of trusses under fixed loading conditions [9, 10].

In the following we describe the structural model and the parametrization of the external loads. We then define the ℓ_p-norm approximation of the maximum von Mises stress to be used as objective function in the load-problem. Thereafter a simplical branch and bound method is described (the method itself is rather standard [11, 12] and is described here mainly for the reader's convenience). Finally some numerical examples are described to give an idea of how

the method works and performs.

2 THE MODEL AND ITS PARAMETRIZATION

We consider an elastic body undergoing small deformations. Using the finite element method we obtain the global equilibrium equation

$$\boldsymbol{K}(\boldsymbol{x})\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{r}) \tag{2}$$

for the nodal displacements in $\boldsymbol{u} \in \mathbb{R}^N$. Following the SIMP-approach to TO the global stiffness matrix

$$\boldsymbol{K}(\boldsymbol{x}) = \sum_{e=1}^{m} \rho_e(\boldsymbol{x})^q \boldsymbol{K}_e,$$

where the (relative) density-variables ρ_e depend on the design variables in $\boldsymbol{x}$ through a, in this case linear, filter [13], and $q > 1$. The matrices $\boldsymbol{K}_e$ are element stiffness matrices for material of unit (relative) density. The design $\boldsymbol{x}$ is assumed to be such that $\boldsymbol{K}(\boldsymbol{x})$ is positive definite.

The load vector in (2) is taken as

$$\boldsymbol{f}(\boldsymbol{r}) = \boldsymbol{f}_0 + \sum_{i=1}^{n} \boldsymbol{Q}_i^{\mathsf{T}} \boldsymbol{r}_i, \tag{3}$$

where $\boldsymbol{f}_0$ is a fixed vector, and $\boldsymbol{r}_i \in \mathbb{R}^s$, where s is the number of spatial dimensions, account for load variations during the optimization. The matrices $\boldsymbol{Q}_i$ can be used to specify which nodes are loaded and the maximum load variation in different directions. The vectors $\boldsymbol{r}_i$ are collected in $\boldsymbol{r} \in \mathcal{T} = \{\boldsymbol{r} \in \mathbb{R}^d \mid ||\boldsymbol{r}_i|| \leq 1,\ i = 1, \ldots, n\}$, where $d = sn$ and $||\cdot||$ is the Euclidean vector norm.

3 THE GLOBAL STRESS MEASURE

In this section we omit all dependencies on $\boldsymbol{x}$ as it is fixed to $\boldsymbol{x}^*$ in (1b).

Given the matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

the von Mises stress at stress evaluation point a can be written as

$$\sigma_a^{vM}(\boldsymbol{r}) = \left[\boldsymbol{\sigma}_a(\boldsymbol{r})^{\mathsf{T}} \boldsymbol{A} \boldsymbol{\sigma}_a(\boldsymbol{r})\right]^{1/2}, \tag{4}$$

where the stress vector is, via (2), given by

$$\boldsymbol{\sigma}_a(\boldsymbol{r}) = \boldsymbol{E}_a \boldsymbol{B}_a \boldsymbol{u}(\boldsymbol{r}) = \boldsymbol{E}_a \boldsymbol{B}_a \boldsymbol{K}^{-1} \boldsymbol{f}(\boldsymbol{r}). \tag{5}$$

Here $\boldsymbol{B}_a$ is the (expanded) strain-displacement matrix evaluated at a and

$$\boldsymbol{E}_a = \rho_e^r \boldsymbol{E}_{0a},$$

where $r > 0$ is a constant, the index e indicates the finite element in which point a is found, and the symmetric, positive semi-definite matrix $\boldsymbol{E}_{0a}$ is the constitutive matrix for material of unit relative density.

Following a currently popular approach to TO with stress constraints/objectives [14, 15] we use as a global stress measure

$$\left(\sum_{a=1}^{S}\sigma_a^{vM}(\boldsymbol{r})^p\right)^{1/p}, \tag{6}$$

where $p > 0$. This is the ℓ_p-norm of the vector of von Mises stresses and it converges from above to $\max_{a\in\{1,\dots,S\}}\sigma_a^{vM}(\boldsymbol{r})$ as p tends to infinity. Since for $p > 0$ the function $x \to x^{1/p}$ is monotone increasing on $[0,\infty)$ we omit the outer exponentiation in (6) and consider maximizing

$$\overline{\sigma}(\boldsymbol{r}) = \sum_{a=1}^{S}\sigma_a^{vM}(\boldsymbol{r})^p. \tag{7}$$

Since the design $\boldsymbol{x}$ is fixed, (7) can be evaluated very efficiently by a rewriting of (4). Let $\boldsymbol{Q}^\mathsf{T} = [\boldsymbol{Q}_1^\mathsf{T}\ \dots\ \boldsymbol{Q}_2^\mathsf{T}] \in \mathbb{R}^{N\times d}$. Then, recalling (5) and (3),

$$\sigma_a^{vM}(\boldsymbol{r}) = \left[(\boldsymbol{f}_0+\boldsymbol{Q}^\mathsf{T}\boldsymbol{r})^\mathsf{T}\boldsymbol{K}^{-1}\boldsymbol{B}_a^\mathsf{T}\boldsymbol{E}_a\boldsymbol{A}\boldsymbol{E}_a\boldsymbol{B}_a\boldsymbol{K}^{-1}(\boldsymbol{f}_0+\boldsymbol{Q}^\mathsf{T}\boldsymbol{r})\right]^{1/2} = \left[\boldsymbol{u}_0^\mathsf{T}\boldsymbol{W}_a\boldsymbol{u}_0+\boldsymbol{r}^\mathsf{T}\boldsymbol{U}^\mathsf{T}\boldsymbol{W}_a\boldsymbol{U}\boldsymbol{r}+2\boldsymbol{u}_0^\mathsf{T}\boldsymbol{W}_a\boldsymbol{U}\boldsymbol{r}\right]^{1/2},$$

where $\boldsymbol{W}_a = \rho_e^r\boldsymbol{B}_a^\mathsf{T}\boldsymbol{E}_{0a}\boldsymbol{A}\boldsymbol{E}_{0a}\boldsymbol{B}_a$ is a very sparse matrix, and $\boldsymbol{u}_0 \in \mathbb{R}^N$ and $\boldsymbol{U} \in \mathbb{R}^{N\times d}$ solves the $sn+1$ linear systems

$$\boldsymbol{K}[\boldsymbol{u}_0\ \boldsymbol{U}] = [\boldsymbol{f}_0\ \boldsymbol{Q}^\mathsf{T}].$$

These equations need only be solved once before the optimization process starts, and the von Mises stress is then given by

$$\sigma_a^{vM}(\boldsymbol{r}) = \left[v_a + 2\boldsymbol{v}_a^\mathsf{T}\boldsymbol{r}+\boldsymbol{r}^\mathsf{T}\boldsymbol{V}_a\boldsymbol{r}\right]^{1/2},$$

where $v_a = \boldsymbol{u}_0^\mathsf{T}\boldsymbol{W}_a\boldsymbol{u}_0$, $\boldsymbol{v}_a = \boldsymbol{U}^\mathsf{T}\boldsymbol{W}_a\boldsymbol{u}_0 \in \mathbb{R}^d$, and $\boldsymbol{V}_a = \boldsymbol{U}^\mathsf{T}\boldsymbol{W}_a\boldsymbol{U} \in \mathbb{R}^{d\times d}$.

4 THE LOAD-PROBLEM

In the special case when the fixed part of the load $\boldsymbol{f}_0 = \boldsymbol{0}$, (7) and (4) gives $\overline{\sigma}(\boldsymbol{r}) = \overline{\sigma}(-\boldsymbol{r})$. This implies (see Appendix A) that

$$\{\,\overline{\sigma}(\boldsymbol{r}) \mid \boldsymbol{r}\in\mathcal{T}\,\} = \{\,\overline{\sigma}(\boldsymbol{r}) \mid \boldsymbol{r}\in\mathcal{T},\ \boldsymbol{n}^\mathsf{T}\boldsymbol{r}\geq 0\,\}$$

for some vector $\boldsymbol{n}$. In view of this we now introduce

$$\mathcal{T}_+ = \{\,\boldsymbol{r}\in\mathbb{R}^d \mid ||\boldsymbol{r}_i||\leq 1,\ i=1,\dots,n,\ \boldsymbol{n}^\mathsf{T}\boldsymbol{r}\geq 0\,\} \tag{8}$$

(the case when $\boldsymbol{f}_0 \neq \boldsymbol{0}$ is retrieved by letting $\boldsymbol{n} = \boldsymbol{0}$) and consider problem (1b), with $\sigma = -\overline{\sigma}$ in place of f and feasible set $\mathcal{T}_+$ instead of $\mathcal{T}$:

$$\min_{\boldsymbol{r}\in\mathcal{T}_+}\sigma(\boldsymbol{r}). \tag{9}$$

This problem is one of minimizing a smooth, concave function [2] over a compact, convex set. Existence of globally optimal solutions is thus guaranteed, and the problem is of a type much studied in the global optimization literature [12, 11].

4.1 A BRANCH-AND-BOUND ALGORITHM FOR THE LOAD-PROBLEM

We describe a branch-and-bound algorithm to solve (9) where the feasible set $\mathcal{T}_+$ is partitioned into simplices $\mathcal{S}_i$. A d-simplex in $\mathbb{R}^d$ is defined as the convex hull of $d+1$ affinely independent vertices $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$, i.e.

$$\boldsymbol{r} \in \mathcal{S}_i \;\Leftrightarrow\; \boldsymbol{r} = \sum_{i=0}^{d} \lambda_i \boldsymbol{v}_i, \;\; \sum_{i=0}^{d} \lambda_i = 1, \;\lambda_i \geq 0, \; i = 0, \ldots, d. \tag{10}$$

To specify the algorithm we need to know: i) how to construct an initial simplex; ii) how to obtain upper and lower bounds on a simplex; iii) how to perform the partition of $\mathcal{T}_+$; and iv) how to check whether $\mathcal{T}_+ \cap \mathcal{S}_i = \emptyset$. These items are described next.

4.1.1 AN INITIAL SIMPLEX

Following [12, Section 3.5.3] we choose an initial simplex $\mathcal{S}_0 \supset \mathcal{T}_+$ as

$$\mathcal{S}_0 = \left\{ \boldsymbol{r} \in \mathbb{R}^d \mid r_i \geq \gamma_i, i = 1, \ldots, d, \; \sum_{i=1}^{d} r_i \leq \gamma \right\},$$

where $\gamma, \gamma_1, \ldots, \gamma_d$ are obtained by solving the convex problems

$$\gamma_i = \min_{\boldsymbol{r} \in \mathcal{T}_+} \; r_i, \quad i = 1, \ldots, d, \qquad \text{and} \qquad \gamma = \max_{\boldsymbol{r} \in \mathcal{T}_+} \sum_{i=1}^{d} r_i.$$

The vertices of $\mathcal{S}_0$ are then given by $\boldsymbol{v}_0 = (\gamma_1, \ldots, \gamma_d)^\mathsf{T}$ and

$$\boldsymbol{v}_i = (\gamma_1, \ldots, \gamma_{i-1}, \alpha_i, \gamma_{i+1}, \ldots, \gamma_d)^\mathsf{T}, \quad i = 1, \ldots, d,$$

where $\alpha_i = \gamma - \sum_{j \neq i} \gamma_j$.

4.1.2 UPPER BOUNDING

Assuming $\mathcal{T}_+ \cap \mathcal{S}_i \neq \emptyset$, an upper bound U_i is obtained on $\mathcal{T}_+ \cap \mathcal{S}_i$ by solving the problem

$$U_i = \min_{\boldsymbol{r} \in \mathcal{T}_+ \cap \mathcal{S}_i} \sigma(\boldsymbol{r}). \tag{11}$$

It is possible to treat this problem in two different ways: (i) compute the polytope representation of $\mathcal{S}_i$ [16, p. 33] and optimize directly in $\boldsymbol{r}$; or (ii) use the vertex representation (10) with the λ_i-s as variables. Here we choose the former strategy because of convenience in the implementation. For problems with more variables than the ones treated here, the fact that the polytope representation requires solution of some additional linear systems compared to the vertex representation may be one reason to prefer the latter.

To represent $\mathcal{S}_i$ as a polyhedron we follow [16, p. 33] and introduce a matrix

$$\boldsymbol{B}_i = [\boldsymbol{v}_1 - \boldsymbol{v}_0 \;\; \ldots \;\; \boldsymbol{v}_d - \boldsymbol{v}_0]^{-1} \in \mathbb{R}^{d \times d},$$

where the inverse is well-defined since $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$ are affinely independent. Now [16, p. 33]

$$\mathcal{S}_i = \left\{ \boldsymbol{r} \in \mathbb{R}^d \mid \boldsymbol{B}_i \boldsymbol{r} \geq \boldsymbol{B}_i \boldsymbol{v}_0, \; \mathbf{1}^\mathsf{T} \boldsymbol{B}_i \boldsymbol{r} \leq \mathbf{1}^\mathsf{T} \boldsymbol{B}_i \boldsymbol{v}_0 + 1 \right\},$$

where $\mathbf{1}$ is a vector of ones.

4.1.3 LOWER BOUNDING

Since $\sigma(\boldsymbol{r})$ is concave, its convex envelope over $\mathcal{S}_i$ is [12, Theorem 1.22] the affine function $\ell(\boldsymbol{r}) = \boldsymbol{a}^\mathsf{T}\boldsymbol{r} + b$, where $\boldsymbol{a}$ and b comprise the unique solution to the linear system

$$\boldsymbol{a}^\mathsf{T}\boldsymbol{v}_i + b = \sigma(\boldsymbol{v}_i), \quad i = 0, \dots, d.$$

Assuming $\mathcal{T}_+ \cap \mathcal{S}_i \neq \emptyset$, a lower bound L_i on $\mathcal{T}_+ \cap \mathcal{S}_i$ is obtained by solving

$$L_i = \min_{\boldsymbol{r} \in \mathcal{T}_+ \cap \mathcal{S}_i} \ell(\boldsymbol{r}). \tag{12}$$

Since the intersection of two convex sets is convex, this problem consists of minimizing a linear function over a convex set.

In this case, solutions to the lower bounding problem are feasible in the original problem (9), so, in order to obtain an upper bound U_i, simply evaluating σ at a solution to (12) is an alternative to solving (11)

4.1.4 SUBDIVISION OF SIMPLICES

Subdivision of a simplex $\mathcal{S}_i$ is done by bisection along one of its longest edges. Letting $\boldsymbol{v}_k$ and $\boldsymbol{v}_{k+1}$ be the end-points of this edge we take the bisection point as $\boldsymbol{w} = 0.5(\boldsymbol{v}_k + \boldsymbol{v}_{k+1})$ and thus obtain two new simplices $\mathcal{S}_{i1}$ and $\mathcal{S}_{i2}$ with vertices $(\boldsymbol{v}_0, \dots, \boldsymbol{v}_k, \boldsymbol{w}, \boldsymbol{v}_{k+2}, \dots, \boldsymbol{v}_d)$ and $(\boldsymbol{v}_0, \dots, \boldsymbol{v}_{k-1}, \boldsymbol{w}, \boldsymbol{v}_{k+1}, \dots, \boldsymbol{v}_d)$.

4.1.5 INTERSECTION BETWEEN $\mathcal{T}_+$ AND $\mathcal{S}_i$

It could happen that the intersection $\mathcal{T}_+ \cap \mathcal{S}_i$ is empty, in which case $\mathcal{S}_i$ should be discarded from further consideration. Here we propose to check for an empty intersection in three steps[1]:

1. If any of the vertices of $\mathcal{S}_i$ lie in $\mathcal{T}_+$, then $\mathcal{T}_+ \cap \mathcal{S}_i \neq \emptyset$ and $\mathcal{S}_i$ should not be discarded.

2. Let $\boldsymbol{I}_j$ be a diagonal matrix with diagonal elements $2j-1$ and $2j$ set to one and the other to zero. If
$$\max_{j \in \{1,\dots,n\}} \left[-\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + 2 \min_{i \in \{1,\dots,d\}} \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_i \right] > 1,$$
then $\mathcal{T}_+ \cap \mathcal{S}_i = \emptyset$ (see Appendix B) and $\mathcal{S}_i$ should be discarded.

3. If neither of Step 1 or 2 hold we proceed to solving the convex feasibility problem
$$\begin{aligned} z^* = \min_{z \in \mathbb{R},\, \boldsymbol{r} \in \mathcal{S}_i} \; & z \\ \text{s. t.} \; & \begin{cases} \boldsymbol{n}^\mathsf{T}\boldsymbol{r} \geq 0 \\ ||\boldsymbol{r}_i||^2 - 1 \leq z, \quad i = 1, \dots, n. \end{cases} \end{aligned} \tag{13}$$
If $z^* \leq 0$ we conclude that $\mathcal{T}_+ \cap \mathcal{S}_i \neq \emptyset$, and $\mathcal{S}_i$ should not be discarded.

[1] Since the criteria in Step 2 is *sufficient* for an empty intersection, it is also possible to skip Step 3 altogether without losing candidate solutions to (9).

4.2 THE COMPLETE ALGORITHM

Let $\mathcal{A}$ denote a set of *problems*. A problem i in $\mathcal{A}$ is defined simply by the vertices of a simplex $\mathcal{S}_i$ and is associated with a lower bound L_i. The proposed algorithm for solving (9) to global optimality (to within a certain tolerance) can now be described as follows:

Algorithm 1

Choose $\varepsilon > 0$. Construct an initial simplex $\mathcal{S}_0$ as described in section 4.1.1 and add it to $\mathcal{A}$. Solve the lower bounding problem (12) and take the objective function value as the best lower bound L. Set the best upper bound U to ∞.

1. Select the problem, i, with the least lower bound and delete it from $\mathcal{A}$.
2. (Optional) Solve the upper bounding problem (11). If the optimal value U_i is smaller than U then set $U := U_i$.
3. Divide $\mathcal{S}_i$ into $\mathcal{S}_{i1}$ and $\mathcal{S}_{i2}$.
4. For $j = 1, 2$: If $\mathcal{S}_{ij} \cap \mathcal{T}_+ \neq \emptyset$, solve the lower bounding problem on $\mathcal{S}_{ij}$. If $L_{ij} < U$, then add problem ij to $\mathcal{A}$.
5. Set L to the least lower bound found in $\mathcal{A}$.
6. If $U - L < \varepsilon$, then STOP. Else goto Step 1.

Let σ^* denote the global minimum value of problem (9). Then according to Theorem 5.26 in [11] Algorithm 1 converges in a finite number of iterations to a point $\boldsymbol{r}^*$ such that $\sigma(\boldsymbol{r}^*) \leq \sigma^* + \varepsilon$[2].

5 NUMERICAL EXAMPLES

We consider solutions to some instances of (9) based on the design (i.e. $\boldsymbol{x}^*$) shown in the left plot of Fig. 1. This design was obtained in [2] by varying, under an upper bound on the total mass, $\boldsymbol{x}$ to minimize the ℓ_{24}-norm of the von Mises-stress vector and varying the load to maximize the same quantity. The L-shaped design domain is indicated by thin dashed lines. The finite element model consists of 6400 four-noded, bilinear elements, and stress is evaluated at the geometric center of each element. To see the effect of varying the design we also consider a problem where the design variables are set to one in the entire design domain. This design is denoted by "s" in Tab. 1 below; the one seen in Fig. 1 by "c".

The load is given by (3) with $\boldsymbol{f}_0 = \boldsymbol{0}$, i.e., $\boldsymbol{f}(\boldsymbol{r}) = \sum_{i=1}^{n} \boldsymbol{Q}_i^\mathsf{T} \boldsymbol{r}_i$. Each $\boldsymbol{Q}_i$ consists here of zeros except for an identity matrix placed at the positions corresponding to the degrees of freedom of a loaded node; i.e. we consider point loads at various parts of the structure varying freely expect that their magnitude is limited to unity and that $\boldsymbol{n}^\mathsf{T}\boldsymbol{r} \geq 0$. Here we take $\boldsymbol{n} = (1\ 0\ \ldots\ 0)^\mathsf{T}$; i.e., we require that $r_1 \geq 0$. Stresses in a small patch of 3 by 4 elements around the point of application of each force are set to zero; c.f. [14, 15].

The right plot in Fig. 1 shows normalized maximum von Mises stress and its approximation using the ℓ_{24}-norm as functions of the angle θ defining the direction of the unit load applied at the tip of the beam in the left plot (here $n = 1$). Three local maxima are clearly discerned in the solid graph, illustrating the non-convexity of problem (9).

Algorithm 1 has been implemented in Matlab R2015b. The NLPs are solved by an interior-

[2] *Exactness in the limit*, as required by the cited theorem, follows from the Lipschitz continuity of σ and that ℓ interpolates σ at the vertices of each simplex. *Exhaustivness of the subdivision process* follows from Proposition 3.14 in [12].

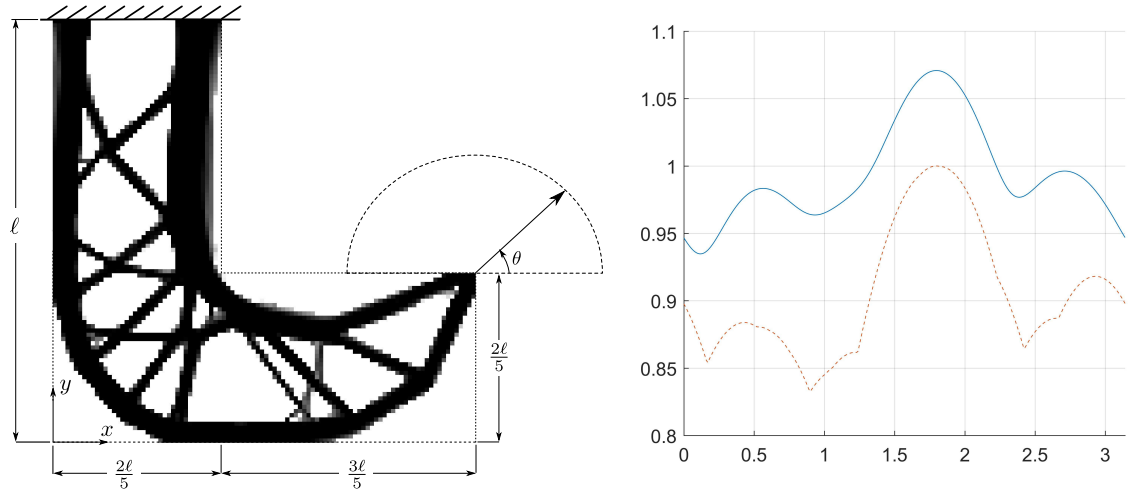


Figure 1: Left: Design with boundary conditions and loading. $\ell = 100$ [mm]. A point load is applied at the tip of the beam and the arrowhead is allowed to vary within the half-circle delimited by the dashed line. Right: Normalized maximum von Mises stress (dashed line) and its ℓ_{24}-norm-approximation (solid line) as functions of the angle θ in the left plot.

point algorithm available via the function `fmincon` in the Matlab Optimization Toolbox. The global stress measure (7) and its gradient are implement as C-functions interfaced to Matlab via the MEX API. Exact first-order derivatives are used for all problems. A BFGS approximation of the Hessian of the Lagrangian is used to solve (9), whereas exact Hessians are used for the other problems. In the model we use $q = 3$, $r = 0.5$; filter radius 3 [mm]; an isotropic material with Young's modulus 1000 [N/mm^2] and Poisson's ratio 0.3; and domain thickness 1 [mm]. The stopping tolerance in Algorithm 1 is set to $\varepsilon = 10^{-3}$. Stopping tolerances for `fmincon` are set to `TolX` $= 10^{-12}$, `TolCon` $= 10^{-10}$ and `TolFun` $= 10^{-10}$. The upper bounding problem in Step 2 of Algorithm 1 is solved every 50:th iteration. All computations are carried out by an Intel Core I7-4712MQ.

Table 1 shows some numerical results. In all cases the NLPs in Algorithm 1 were, after appropriate scaling, solved by `fmincon` using between 10 and 30 iterations for problems (12) and (13), and occasionally up to around 100 for problem (11). For the cases $n = 1$, $d = 2$ in the table, the total CPU time is on the order of 10 seconds and the number of iterations of Algorithm 1 around 200. For the three cases with $n = 2$, $d = 4$, the number of iterations jump to the order of 10000 and CPU times to between 700 and 1200 [s]. The absolute gap $U - L$ as a function of the number of iterations for the case in row three is shown in Fig. 2. Changing the design from the one seen in Fig. 1 to the simpler design "s" in the forth row of Tab. 1 yields a slight increase in computational time compared to the third row. For the cases $n = 3$, $d = 6$, the iteration count increased further. The total time increases with increasing p, with $p = 24$ resulting in around 38 hours of CPU time.

6 CONCLUSIONS

A branch and bound method for finding loads that maximize a given functional, in this case an ℓ_p-norm approximation of the maximum von Mises stress, has been described. The proposed method is quite general and can be applied to a wide range of functionals and for many different

n	d	p	des.	load position	iter.	f. eval.	low. bnd.	$\lvert\mathcal{A}\rvert$	CPU [s]
1	2	24	c	(100,40)	191	1129	343	17	14
1	2	24	c	(40,0)	161	1227	298	13	21
2	4	24	c	(100,40), (0,40)	13761	$1.3{\cdot}10^5$	$2.4{\cdot}10^4$	877	875
2	4	24	s	(100,40), (0,40)	20001	$1.8{\cdot}10^5$	$3.5{\cdot}10^4$	820	1197
2	4	24	c	(40,0), (0,40)	10631	$1{\cdot}10^5$	$1.8{\cdot}10^4$	668	710
3	6	4	c	(100,40), (0,40), (40,0)	$8.7{\cdot}10^4$	$1.2{\cdot}10^6$	$1.6{\cdot}10^5$	5913	6912
3	6	8	c	(100,40), (0,40), (40,0)	$3.3{\cdot}10^5$	$4.2{\cdot}10^6$	$5.7{\cdot}10^5$	15835	24418
3	6	24	c	(100,40), (0,40), (40,0)	$2{\cdot}10^6$	$25{\cdot}10^6$	$3.4{\cdot}10^6$	97540	139435

Table 1: Numerical results. Columns show number of individual loads; number of variables; ℓ_p-norm exponent; type of design; point of application of loads in the coordinate system shown in Fig. 1; number of iterations of Algorithm 1; number of evaluations of (7); number of times the lower bounding problem (12) has been solved; maximum number of active problems; and total time

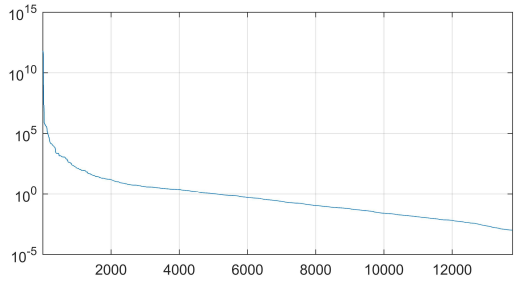


Figure 2: Absolute gap $U - L$ versus number of iterations in Algorithm 1 for the case $n = 2$, $d = 4$ described in the third row of Tab. 1.

loading parametrizations. A test problem based on a topology optimized design showed that the approach is promising, but, although further optimization of the code might bring down computational times somewhat, the large number of iterations observed indicate that more work is needed to be able to solve large-scale (in terms of both the size of the finite element model and the number of independent loads) instances of the load-problem to global optimality. Finally we remark that although the load-problem is here described as part of a Nash game for robust TO it is obviously very relevant on its own.

ACKNOWLEDGEMENTS

This research was supported in part by the Swedish Foundation for Strategic Research, Grant No. AM13-0029. Thanks to Associate Professor Johan Löfberg for useful comments on an early draft.

References

[1] A Ben-Tal, L El Ghaoui, and AS Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[2] E Holmberg, C-J Thore, and A Klarbring. A game theory approach for robust topology optimization with uncertain loading. *Under review*, 2015.

[3] E Holmberg. *Topology optimization considering stress, fatique and load uncertainities*. PhD thesis, Linköping University, 2015.

[4] E Holmberg, C-J Thore, and A Klarbring. Game theory approach for topology optimization with uncertain loads. In *Svenska Mekanikdagar 10-12 Juni, Linköping*, 2015.

[5] E Holmberg, C-J Thore, and A Klarbring. Worst-case topology optimization of self-weight loaded structures using semi-definite programming. *Structural and Multidisciplinary Optimization*, 52(5):915–928, 2015.

[6] C-J Thore, E Holmberg, and A Klarbring. Large-scale robust topology optimization under load-uncertainty. In *11th World Congress on Structural and Multidisciplinary Optimisation*, 2015.

[7] H Svärd. A branch and bound algorithm for evaluation of the Findley fatigue criterion. *International Journal of Fatigue*, 73:27–38, 2015.

[8] Y Kanno and I Takewaki. Worst case plastic limit analysis of trusses under uncertain loads via mixed 0-1 programming. *Journal of Mechanics of Materials and Structures*, 2(2):245–273, 2007.

[9] M Stolpe. Global optimization of minimum weight truss topology problems with stress, displacement, and local buckling constraints using branch-and-bound. *International Journal for Numerical Methods in Engineering*, 61:1270–1309, 2004.

[10] M Stolpe and A Kawamoto. Design of planar articulated mechanisms using branch and bound. *Mathematical Programming*, 103:357–397, 2005.

[11] M Locatelli and F Schoen. *Global Optimization – Theory, Algorithms and Applications*. SIAM, 2013.

[12] R Horst, PM Pardalos, and N Van Thoai. *Introduction to global optimization*. Kluwer Academic Publishers, 1995.

[13] TE Bruns and DA Tortorelli. Topology optimization of non-linear elastic structures and compliant mechanisms. *Computer Methods in Applied Mechanics and Engineering*, 190(26):3443–3459, 2001.

[14] E Holmberg, B Torstenfelt, and A Klarbring. Stress constrained topology optimization. *Structural and Multidisciplinary Optimization*, 48(1):33–47, 2013.

[15] DM De Leon, J Alexandersen, JS Fonseca, and O Sigmund. Stress-constrained topology optimization for compliant mechanism design. *Structural and Multidisciplinary Optimization*, 52(5):929–943, 2015.

[16] S Boyd and L Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[17] U Raber. A simplicial branch-and-bound method for solving nonconvex all-quadratic programs. *Journal of Global Optimization*, 13(4):417–432, 1998.

[18] N Andréasson, A Evgrafov, and M Patriksson. *An Introduction to Continuous Optimization*. Studentlitteratur, 2005.

A IMAGES OF $\mathcal{T}$ AND $\mathcal{T}_+$

Proposition. Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfy $g(\boldsymbol{r}) = g(-\boldsymbol{r})$ for all $\boldsymbol{r} \in \mathcal{T} = \{\boldsymbol{r} \in \mathbb{R}^d \mid ||\boldsymbol{r}_i|| \leq 1,\ i = 1, \dots, n\}$. Then

$$\{\, g(\boldsymbol{r}) \,|\, \boldsymbol{r} \in \mathcal{T} \,\} = \{\, g(\boldsymbol{r}) \,|\, \boldsymbol{r} \in \mathcal{T}_+ \,\},$$

where $\mathcal{T}_+ = \{\, \boldsymbol{r} \in \mathbb{R}^d \mid ||\boldsymbol{r}_i|| \leq 1,\ i = 1, \dots, n,\ \boldsymbol{n}^\mathsf{T}\boldsymbol{r} \geq 0 \,\}$ for some vector $\boldsymbol{n} \in \mathbb{R}^d$.

Proof. We have

$$\{\, g(\boldsymbol{r}) \,|\, \boldsymbol{r} \in \mathcal{T} \,\} = \{\, g(\boldsymbol{r}) \,|\, \boldsymbol{r} \in \mathcal{T},\ \boldsymbol{n}^\mathsf{T}\boldsymbol{r} \geq 0 \,\} \cup \{\, g(\boldsymbol{r}) \,|\, \boldsymbol{r} \in \mathcal{T},\ \boldsymbol{n}^\mathsf{T}\boldsymbol{r} \leq 0 \,\} = U_+ \cup U_- \quad (14)$$

Since $||\boldsymbol{r}_i|| \leq 1$ and $\boldsymbol{n}^\mathsf{T}\boldsymbol{r} \leq 0$ if and only if if and only if $|| - \boldsymbol{r}_i|| \leq 1$ and $\boldsymbol{n}^\mathsf{T}(-\boldsymbol{r}) \geq 0$ we can write

$$U_- = \{\, g(\boldsymbol{r}) \,|\, -\boldsymbol{r} \in \mathcal{T},\, \boldsymbol{n}^\mathsf{T}(-\boldsymbol{r}) \geq 0 \,\} = \{\, g(-\boldsymbol{r}) \,|\, -\boldsymbol{r} \in \mathcal{T},\, \boldsymbol{n}^\mathsf{T}(-\boldsymbol{r}) \geq 0 \,\}.$$

Now defining $\boldsymbol{s} \equiv -\boldsymbol{r}$ we see that $U_- = U_+$, so that, going back to (14),

$$\{\, g(\boldsymbol{r}) \,|\, \boldsymbol{r} \in \mathcal{T} \,\} = U_+ \cup U_- = U_+ \cup U_+ = U_+.$$

□

B A SUFFICIENT CONDITION FOR $\mathcal{T}_+ \cap \mathcal{S}_i = \emptyset$

Proposition. Let $\boldsymbol{I}_j$ be a diagonal matrix with ones in diagonal elements $2j-1$ and $2j$ and zeros otherwise. If

$$\max_{j \in \{1,\dots,n\}} \left[-\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + 2 \min_{i \in \{1,\dots,d\}} \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_i \right] > 1, \quad (15)$$

then $\mathcal{T}_+ \cap \mathcal{S}_i = \emptyset$.

(A version of this criteria was originally proposed by Raber [17].)

Proof. Let $\boldsymbol{r}$ denote a vector from the origin to an arbitrary point in $\mathcal{S}_i$. Then

$$\boldsymbol{r}^\mathsf{T}\boldsymbol{I}_j\boldsymbol{r} = (\boldsymbol{v}_0 + (\boldsymbol{r} - \boldsymbol{v}_0))^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{v}_0 + (\boldsymbol{r} - \boldsymbol{v}_0)) = \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + 2\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{r} - \boldsymbol{v}_0) + (\boldsymbol{r} - \boldsymbol{v}_0)^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{r} - \boldsymbol{v}_0) \geq \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + 2\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{r} - \boldsymbol{v}_0), \quad (16)$$

where the last inequality follows from $\boldsymbol{I}_j$ being positive semi-definite. We now write $\boldsymbol{r} = \boldsymbol{v}_0 + \sum_{i=1}^d \lambda_i(\boldsymbol{v}_i - \boldsymbol{v}_0)$, where $\lambda_i \geq 0$, $i = 1, \dots, d$ and $\sum_{i=1}^d \lambda_i \leq 1$. Substitution in (16) gives

$$\boldsymbol{r}^\mathsf{T}\boldsymbol{I}_j\boldsymbol{r} \geq \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + 2\boldsymbol{v}_0^\mathsf{T}\left(\sum_{i=1}^d \lambda_i \boldsymbol{I}_j(\boldsymbol{v}_i - \boldsymbol{v}_0)\right) \geq \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + \min_{\lambda_i \geq 0,\, i=1,\dots,d,\, \sum_{i=1}^d \lambda_i \leq 1} 2\boldsymbol{v}_0^\mathsf{T}\left(\sum_{i=1}^d \lambda_i \boldsymbol{I}_j(\boldsymbol{v}_i - \boldsymbol{v}_0)\right) \geq \boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + \min_{i \in \{1,\dots,d\}} 2\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{v}_i - \boldsymbol{v}_0). \quad (17)$$

The last inequality follows from the fact that for a linear program with a bounded feasible set, at least one extreme point of this set will be a solution [18, Theorem 8.10]; i.e, at least one of the vertices of $\mathcal{S}_i$ will be a solution. The last inequality in (17) shows that $\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j\boldsymbol{v}_0 + \min_{i\in\{1,\dots,d\}} 2\boldsymbol{v}_0^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{v}_i-\boldsymbol{v}_0) > 1$ implies $\boldsymbol{r}^\mathsf{T}\boldsymbol{I}_j\boldsymbol{r} > 1$ and thus, since $\boldsymbol{r}\in\mathcal{S}_i$ is arbitrary, that $\mathcal{T}_+\cap\mathcal{S}_i=\emptyset$. □

Remark. The smaller the simplex, the smaller is the term $(\boldsymbol{r}-\boldsymbol{v}_0)^\mathsf{T}\boldsymbol{I}_j(\boldsymbol{r}-\boldsymbol{v}_0)$ in (16) and the better is the accuracy of the criterion (15).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ANALYSIS AND DESIGN OF REINFORCED CONCRETE STRUCTURES AS A TOPOLOGY OPTIMIZATION PROBLEM

M. Bruggi

Department of Civil and Environmental Engineering DICA
Politecnico di Milano
Piazza Leonardo da Vinci, 32 20133 Milano Italy
e-mail: matteo.bruggi@polimi.it

Keywords: Topology Optimization, No–Tension Materials, Reinforced–Concrete Structures, Size Optimization.

Abstract. *Technical codes for buildings deal with cracked reinforced concrete structures assuming concrete as a compression–only material, whereas rebar provides the structural component with the required tensile strength [1]. Numerical methods can handle reinforced concrete structures calling for demanding non–linear analysis. Indeed, well–known convergence issues arise when copying with concrete as a compression–only material. Recently, an alternative energy–based approach has been proposed to solve the equilibrium of a linear elastic no–tension medium exploiting its hyper–elasticity [2]. A topology optimization problem distributes an equivalent orthotropic material to minimize the strain energy of the no-tension body, thus avoiding more demanding non–linear analysis. This contribution provides an extension to the analysis and optimal design of reinforced concrete structures. Following [3], truss members are modeled within a two–dimensional no–tension continuum in order to model structural elements made of reinforced concrete. The solution of the equilibrium is straightforward within the approach proposed in [2], thus allowing performing analysis at the serviceability limit state with cracked sections. Also, introducing the areas of the reinforcement bars as an additional set of unknowns, a problem of size optimization is outlined to cope with the optimal rebar of r.c. structures. Preliminary numerical simulations are shown to assess the proposed procedure.*

1 INTRODUCTION

An extensive research has been done in the last decades addressing the non–linear analysis of reinforced concrete (r.c.) structures, see e.g. the seminal paper [4]. At the serviceability limit state, the practical detailing of a reinforced concrete structure is based on simple assumptions, i.e. the adoption of linear elastic modeling for the composite structural member and the assumption that concrete strength in tension is negligible [5]. Implementing this theoretical model within a numerical code is not trivial, due to well–known instabilities arising when addressing a linear elastic no–tension body through conventional incremental approaches. Recently, [6] has proposed to solve the equilibrium of a no–tension solid resorting to a minimization problem that adopts the displacement field as unknown and the strain energy of the hyper–elastic no–tension body as objective function.

An alternative energy–based approach has been proposed in [2]. A topology optimization problem [7] distributes an equivalent orthotropic material to minimize the strain energy of the compression–only body, thus resorting to an established and computationally efficient formulation of compliance minimization that avoids more demanding non–linear analysis.

A possible extension to the simplified analysis of reinforced concrete structures is investigated in this contribution. Following [3], truss members are modeled within a two–dimensional no–tension continuum in order to address two–dimensional structural elements made of reinforced concrete. The solution of the equilibrium is straightforward within the approach proposed in [2], thus allowing performing analysis at the serviceability limit state with cracked sections. Preliminary numerical simulations are shown to investigate the capabilities of the proposed procedure.

Moreover, introducing the areas of the reinforcement bars as an additional set of unknowns for the minimization problem, size optimization is outlined as a possible extension of the proposed procedure to cope with the optimal rebar of r.c. structures. A first numerical test shows that the computational burden tied to this extension is almost equivalent to that required for the solution of the equilibrium equations.

The outline of the paper is as follows. Section 2.1 recalls fundamentals of the topology optimization problem used to address linear elastic no–tension structures according to [2, 8], whereas Section 2.2 introduces the minimum compliance problem to cope with r.c. members. Section 3 presents preliminary numerical simulations performed on a benchmark example and Section 4 provides remarks and outlines the ongoing research.

2 PROBLEM FORMULATION

2.1 Equilibrium of no–tension structures as a topology optimization problem

The analysis of a 2D no–tension continuum is re–formulated as a topology optimization problem for minimum compliance. The equilibrium of any compression–only structure is solved seeking for the distribution of an 'equivalent' orthotropic material that minimizes the potential energy of the hyper–elastic solid. A one–shot energy–based procedure computes the non–incremental solution under given loads, provided that the applied forces are compatible with the no–tension constraint.

A 2D Cartesian reference is considered to address a no–tension isotropic solid in the region Ω. Prescribed displacements $\underline{u}_0$ are enforced along its constrained boundary, Γ_u, whereas traction $\underline{t}_0$ is assigned along its free boundary, Γ_t. An equivalent orthotropic material is introduced to mimic the behavior of a no–tension body. $(\widetilde{z}_1, \widetilde{z}_2)$ are the symmetry axes of the equivalent phase that are assumed to be aligned with the directions (z_I, z_{II}) of the principal stresses

(σ_I, σ_{II}) at any point in Ω. The parameter θ defines the orientation of the axes $(z_I,\ z_{II})$ with respect to the given Cartesian reference.

In weak form, the energy–based minimization problem can be stated as follows:

$$\begin{cases} \min\limits_{\rho_{min}\leq\rho_1,\rho_2\leq 1} & \displaystyle\int_\Omega \mathbf{D}(\rho_1,\rho_2,\theta)\underline{\varepsilon}(\underline{u})\ \underline{\varepsilon}(\underline{u})\ d\Omega \\ \text{s.t.} & \displaystyle\int_\Omega \mathbf{D}(\rho_1,\rho_2,\theta)\underline{\varepsilon}(\underline{u})\ \underline{\varepsilon}(\underline{v})\ d\Omega = \int_{\Gamma_t} \underline{t}_0\cdot\underline{v}\ d\Gamma \text{ and } \underline{u}\mid_{\Gamma_u}=\underline{u}_0 \quad \forall \underline{v}, \\ & \theta \mid \widetilde{z}_1 = z_I \text{ and } \widetilde{z}_2 = z_{II}, \\ & \rho_1,\rho_2 \mid \sigma_I \leq 0 \text{ and } \sigma_{II}\leq 0. \end{cases} \tag{1}$$

The above equation adopts the compliance (i.e. twice the strain–energy) as objective function, $\underline{u}$ is the displacement field, $\underline{\varepsilon}(\underline{u})$ the strain tensor and $\mathbf{D}$ the fourth–order elasticity tensor of the 'equivalent' orthotropic material. Using a suitable transformation matrix $\mathbf{T}(\theta)$, one has that $\mathbf{D} = \mathbf{T}^{-1}\widetilde{\mathbf{D}}\mathbf{T}^{-t}$. Exploiting the Voigt's notation, $\widetilde{\mathbf{D}}$, which is the constitutive tensor written in the material frame $(\widetilde{z}_1, \widetilde{z}_2)$, reads:

$$\widetilde{\mathbf{D}} = \frac{1}{1-\widetilde{\nu}_{12}\widetilde{\nu}_{21}} \begin{bmatrix} \widetilde{E}_1 & \widetilde{\nu}_{12}\widetilde{E}_2 & 0 \\ \widetilde{\nu}_{21}\widetilde{E}_1 & \widetilde{E}_2 & 0 \\ 0 & 0 & (1-\widetilde{\nu}_{12}\widetilde{\nu}_{21})\widetilde{G}_{12} \end{bmatrix}, \tag{2}$$

where $\widetilde{E}_1$ and $\widetilde{E}_2$ are the Young's moduli of the 'equivalent' orthotropic material, $\widetilde{\nu}_{12}, \widetilde{\nu}_{21}$ its Poisson's ratios (being $\widetilde{\nu}_{12}/\widetilde{E}_1 = \widetilde{\nu}_{21}/\widetilde{E}_2$) and $\widetilde{G}_{12}$ its shear modulus.

A generalization of the so–called SIMP model [9] relates the elastic constants of the equivalent material and those (E, ν) of the isotropic no–tension material according to the following interpolations:

$$\widetilde{E}_i = \rho_i^p E, \quad \widetilde{\nu}_{ij} = \sqrt{\frac{\rho_i^p}{\rho_j^p}}\nu, \quad \widetilde{G}_{ij} = \sqrt{\rho_i^p\rho_j^p}\frac{E}{2(1+\nu)}, \tag{3}$$

for $i, j = 1, 2$. Each minimization unknown ρ_i ranges between $\rho_{min} > 0$ and 1 to penalize or preserve stiffness along the relevant axis depending on the sign of the principal stress. To avoid any tensile stress in the solid, Eqns. (1.4) enforce vanishing stiffness of the equivalent orthotropic material along the direction of any arising positive principal stress. The adopted lower bound ρ_{min} avoids singularity of the stiffness matrix when solving the discrete formulation, whereas $p = 3$ is conventionally assumed [7].

2.2 Equilibrium of r.c. structures as a topology optimization problem

The generalization of the above problem to the finite element analysis of a cracked reinforced concrete structure seen as a strengthened no–tension body is straightforward. Following [3], truss elements are modeled within a two–dimensional continuum in order to cope with the arising composite structure. The formulation is presented in its discrete form.

A finite element discretization made of N truss elements is used for the reinforcement bars, along with M four–node plane elements for the underlying compression–only structure. An element–wise constant discretization is adopted to cope with ρ_1, ρ_2, θ, see Section 2.1. The stiffness matrix of the j–th truss–like element is denoted by $\mathbf{K}_j^r$, whereas x_{1i} and x_{2i} are the discrete minimization unknowns that govern the stiffness of the 'equivalent material' along its symmetry axes, being t_i the value of the orientation parameter in the i–th element. Denoting by

$\mathbf{K}_i(x_{1i}, x_{2i}, t_i)$ the stiffness matrix of the i–th plane element, the implemented discrete form for the minimization of the strain energy of a 2D r.c. structure reads:

$$\begin{cases} \min\limits_{x_{min} \leq x_{1i}, x_{2i} \leq 1} & \mathcal{C} = \sum_{i=1}^{M} \mathbf{U}_i^T \, \mathbf{K}_i(x_{1i}, x_{2i}, t_i) \, \mathbf{U}_i + \sum_{j=1}^{N} \mathbf{U}_j^T \, \mathbf{K}_j^r \, \mathbf{U}_j \\ \text{s.t.} & \sum_{i=1}^{M} \mathbf{K}_i(x_{1i}, x_{2i}, t_i) \, \mathbf{U} + \sum_{j=1}^{N} \mathbf{K}_j^r \, \mathbf{U}_j \; = \; \mathbf{F}, \\ & t_i \;\mid\; \widetilde{z}_1 = z_I \text{ and } \widetilde{z}_2 = z_{II}, \\ & x_{1i}, x_{2i} \;\mid\; \sigma_{i,I} \leq 0 \text{ and } \sigma_{i,II} \leq 0, \end{cases} \tag{4}$$

where $\mathbf{U}_i$ is the vector of the d.o.f.s of the 2D finite elements, $\mathbf{U}_j$ is the vector of the d.o.f.s of the truss elements and F is the array of the nodal loads. Eqn.(4) is fully along the lines of Eqn.(1). It is recalled that the structural compliance $\mathcal{C}$ is the work of the external loads computed at equilibrium. Dealing with a composite structure, the overall strain energy depends on the amount stored in the underlying compression–only material and that stored in the steel reinforcement.

Sequential convex programming [10] and analytical computation of the sensitivities [7] can be used to solve the minimization problem stated in Eqn.(4).

Instead of implementing demanding sets of stress constraints, the penalization approach already used in [2, 8] is herein adopted to cope with Eqn.(4.4). A set of penalized densities $\widehat{x}_{1i}$, $\widehat{x}_{2i}$ can be introduced for a straightforward computation of a modified strain energy $\widehat{\mathcal{C}}$, in which the terms related to any possible positive principal stress arising in the underlying compression–only material are reduced by a parameter $k = 0.5$. Providing the optimizer with the reduced objective function $\widehat{\mathcal{C}}$ and its sensitivities $\partial\widehat{\mathcal{C}}/\partial x_{1i}$, $\partial\widehat{\mathcal{C}}/\partial x_{2i}$, variables $\mathbf{x}_1, \mathbf{x}_2$ are updated preventing any distribution of stiff material along the weak direction(s) of the no–tension body.

Introducing a new set of minimization unknowns to cope with bars having different sections, minor modifications are required in Eqn.(4) to cope with a problem of size optimization. A preview of this approach is shown in Section 3.2, addressing the optimal detailing of r.c. structures.

3 NUMERICAL SIMULATIONS

A set of preliminary numerical simulations are presented in this section, adopting the formulation described above to cope with the analysis of the reinforced concrete cantilever represented in Figure 1. The left edge is clamped and a nodal force $P = 1\,\text{kN}$ is applied at the upper right corner.

Concrete is modeled as a linear elastic no–tension material with Young modulus $E_c = 20,000\,\text{MPa}$ and Poisson's ratio $\nu_c = 0.15$, whereas the prescribed steel reinforcement is discretized through linear elastic truss elements with Young modulus $E_s = 210,000\,\text{MPa}$. The

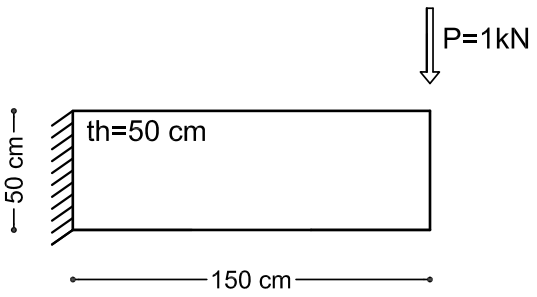


Figure 1: Geometry and boundary conditions for the numerical application.

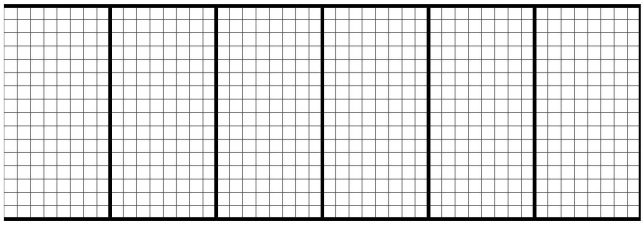


Figure 2: Finite element mesh and steel reinforcement layout.

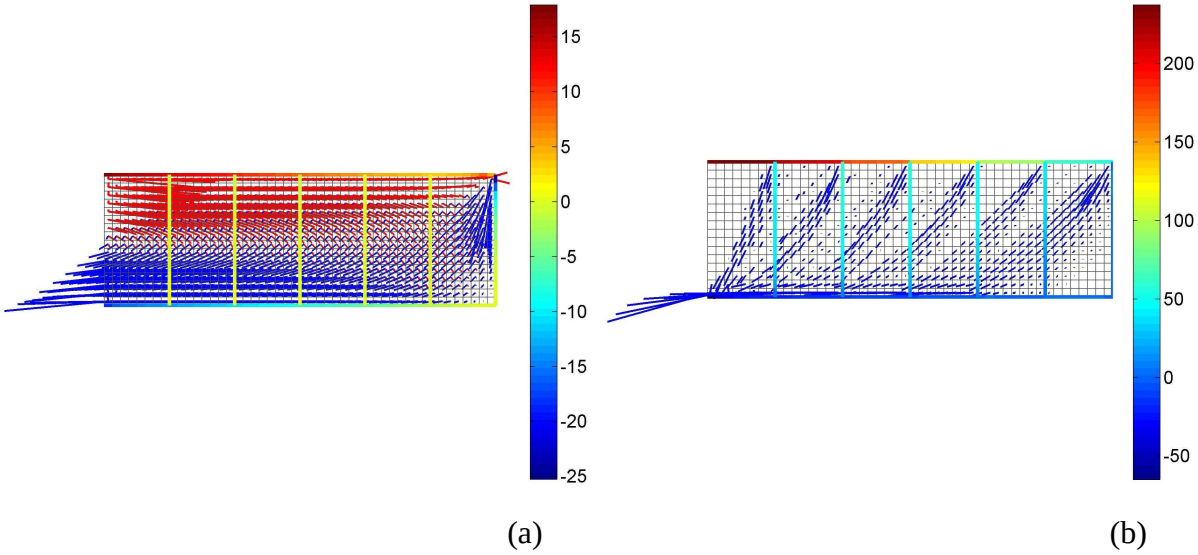


Figure 3: Analysis of the r.c. cantilever: principal stresses in the concrete specimen along with uniaxial stresses in the reinforcement (MPa) computed for a conventional linear elastic beam (a) and a no–tension beam (b).

area of each bar is twice that of a circular section with diameter $\phi_s = 12\,\text{mm}$.

Section 3.1 assesses the proposed procedure of analysis, whereas Section 3.2 outlines a possible extension to the size optimization of the steel reinforcement.

3.1 The analysis problem

Figure 2 shows the two–dimensional finite element discretization used to handle the concrete cantilever along with the truss elements that model stirrups and longitudinal reinforcement, both with the same diameter ϕ_s. For simplicity's sake, no reinforcing covering is considered in this preliminary investigation. Stirrup spacing is uniform along the beam and equal to half the height of the square section.

Figure 3(a) shows results computed through a linear elastic analysis that models concrete as a material with equal behavior in tension and compression. Principal stresses in the concrete specimen and uniaxial stresses in the reinforcement are represented in the same picture. Due to the enforced compatibility, the steel reinforcement and the concrete beam carry different amounts of the external load, depending on their stiffness. As expected, the upper part of the beam is tensile–stressed (red vectors), whereas the region under the neutral axis is compressed (blue vectors). Minor axial stresses are found in the longitudinal reinforcement; stirrups are nearly unloaded.

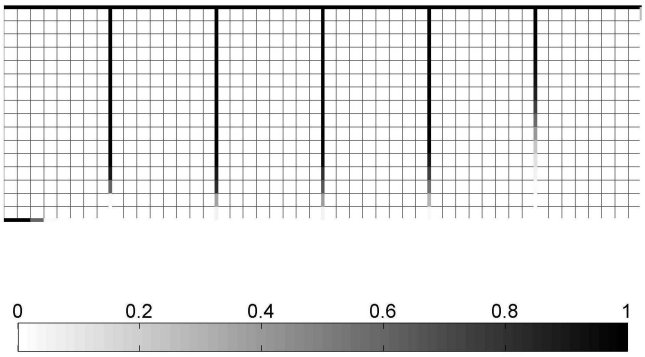


Figure 4: Optimization of the r.c. cantilever: achieved layout.

Figure 3(b) shows results computed through the minimization problem of Eqn. (4), which allows modeling concrete as a linear elastic no–tension material. Results shown in this picture are quite different with respect to those of the previous one. No tensile stress arises in the concrete body (no red vector is found). The tensile stresses in the upper longitudinal reinforcement are around $230\,\mathrm{MPa}$, whereas compressive stresses in the lower horizontal bars are around $50\,\mathrm{MPa}$.

The direction of the principal stresses computed in the two–dimensional concrete domain clearly shows the expected activation of a strut–and–tie model for shear resistance that involves the stirrups, now tensile–stressed by nearly $50\,\mathrm{MPa}$. Note that the inclination of the struts in the arising truss–like structure is not the same along the cantilever. This can be directly compared with established results of strut–and–tie modeling, see e.g. [1, 11].

3.2 The optimization problem

The formulation in Eqn. (4) can be straightforwardly modified to cope with the detailing of the optimal amount of steel reinforcement in the cantilever.

Replacing the stiffness of the j–th reinforcement bar $\mathbf{K}_j^r$ with $x_j\mathbf{K}_j^r$, being $0 \leq x_j \leq 1$ a sizing unknown that allows for a variation of the diameter of the bar section in the range $0 \leq \phi_s \leq 12\,\mathrm{mm}$, the problem in Eqn. (4) minimizes the strain energy of the composite structure solving, within the same formulation, the equilibrium of the no–tension cantilever and the size optimization of the prescribed reinforcement bars.

Figure 4 shows the distribution of the unknowns x_j found by the algorithm to detect optimal sections, enforcing that the allowed global amount of reinforcement is half the case with $x_j = 1,\ \forall j$. Figure 5 represents the principal compressive stresses found in the concrete specimen along with the uniaxial stresses computed in the reinforcement. Removing the horizontal reinforcement lying in the compressive–stressed region and suitable portions of the stirrups, a lighter layout is achieved than in Figure 2 without introducing any remarkable variation in the computed stress field. The achieved solution is trivial, but allows outlining capabilities of the proposed formulation that can be conveniently exploited in case of more complex geometry, load and reinforcement patterns.

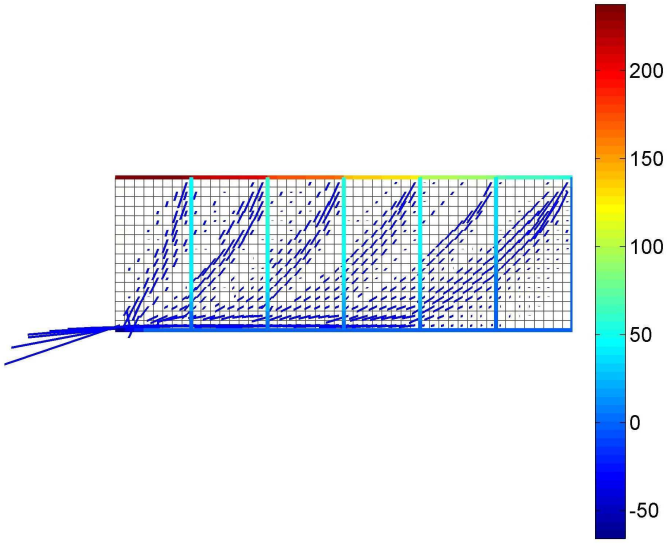


Figure 5: Optimization of the r.c. cantilever: principal stresses in the concrete specimen along with uniaxial stresses in the reinforcement (MPa).

Figure 6 reports the history plot of the objective function for the analysis problem and the optimization problem that have been considered above. Both curves are similar, convergence is smooth and the computational cost is limited, as for conventional problems of topology optimization for minimum compliance.

4 CONCLUSIONS

A topology optimization problem has been formulated as an alternative approach to cope with the analysis of cracked reinforced concrete structures, assuming the hyper–elastic no–tension model for concrete. An energy–based formulation which was originally developed for plain elements has been herein extended to handle compression–only composite structures embedding bars of steel reinforcement.

Results found by the preliminary analysis commented above are in very good agreement with established theories addressing the behavior of cracked reinforced–concrete structures, see e.g. [11]. The proposed approach allows for a direct computation of compressive stresses acting in the concrete domain along with axial stresses in the steel reinforcement. This matches well–known methods used in engineering practice that neglect the tensile strength of concrete, see e.g. [5].

Moreover, a problem of size optimization has been outlined to cope with the optimal rebar of r.c. sections through minor modifications of the proposed numerical procedure.

The ongoing research is mainly focused on the assessment of the analysis problem, the development of the optimization problem and the extension of the proposed approach to cope with the optimal fiber–reinforcement of existing structures, see in particular [12, 13] for plain and reinforced concrete structures and [14, 15] for masonry structures.

Referring to computational issues, adaptive techniques are currently under investigation to improve the accuracy in the evaluation of both the displacement field and the stress field in the no–tension layer while decreasing the computational effort, see in particular [16, 17].

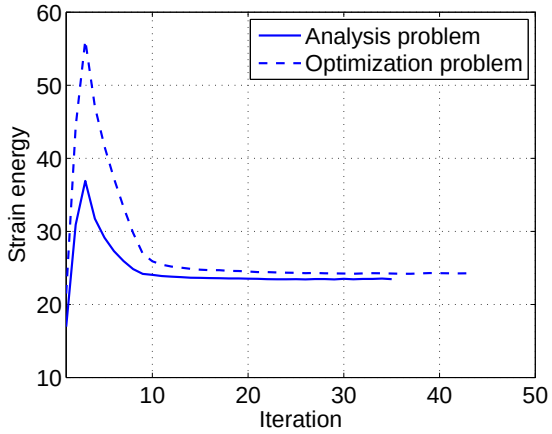


Figure 6: Convergence curves for the analysis problem and the optimization problem (Nm).

REFERENCES

[1] P. Marti, Basic tools of reinforced concrete beam design. *Journal of the American Concrete Institute*, **82**, 46–56, 1985.

[2] M. Bruggi, Finite element analysis of no–tension structures as a topology optimization problem. *Structural and Multidisciplinary Optimization*, **50**(6), 957–973, 2014.

[3] A.L. Gaynor, J.K. Guest, C.D. Moen, Reinforced concrete force visualization and design using bilinear truss-continuum topology optimization. *Journal of Structural Engineering*, **139**, 607–618, 2013.

[4] H.G. Kwak, F.C. Filippou, Nonlinear FE analysis of R/C structures under monotonic loads. *Computers and Structures*, **65**, 1–16, 1997.

[5] EN 1992-1-1, Eurocode 2: Design of concrete structures Part 1-1: General rules and rules for buildings, 2004.

[6] M. Angelillo, L. Cardamone, A. Fortunato, A numerical model for masonry–like structures. *Journal of Mechanics of Materials and Structures* **5**, 583–615, 2010.

[7] M.P. Bendsøe, O. Sigmund, *Topology optimization theory, methods and applications*, New York, Springer, 2003.

[8] M. Bruggi, A. Taliercio, Analysis of no–tension structures under monotonic loading through an energy–based method. *Computers and Structures* **159**, 14–25, 2015.

[9] M.P. Bendsøe, Optimal shape design as a material distribution problem. *Structural Optimization*, **1**, 193–202, 1989.

[10] K. Svanberg, Method of moving asymptotes - a new method for structural optimization. *International Journal for Numerical Methods in Engineering*, **24**, 193-202, 1987.

[11] J. Schlaich, K. Schaefer, M. Jennewein, Toward a consistent design of structural concrete. *PCI Journal*, **32**(3), 74–150, 1987.

[12] M. Bruggi, A. Taliercio, Optimal strengthening of concrete plates with unidirectional fiber-reinforcing layers. *International Journal of Solids and Structures*, **67**–**68**, 311–325, 2015.

[13] J. Cunha, L.P. Chaves, The use of topology optimization in disposing carbon fiber reinforcement for concrete structures. *Structural and Multidisciplinary Optimization*, **49**(6), 1009–1023, 2014.

[14] M. Bruggi, G. Milani, A. Taliercio, Design of the optimal fiber-reinforcement for masonry structures via topology optimization. *International Journal of Solids and Structures*, **50**(13), 2087–2106, 2013.

[15] M. Bruggi, G. Milani, A. Taliercio, Simple topology optimization strategy for the FRP reinforcement of masonry walls in two-way bending. *Computers and Structures*, **138**, 86–101, 2014.

[16] Y. Wang, Z. Kang, Q. He, Adaptive topology optimization with independent error control for separated displacement and density fields. *Computers and Structures*, **135**, 50–61, 2014.

[17] M. Bruggi, M. Verani, A fully adaptive topology optimization algorithm with goal–oriented error control. *Computers and Structures*, **89**(15–16), 1481–1493, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ADJOINT OPTIMAL CONTROL PROBLEMS FOR FLUID-STRUCTURE INTERACTION SYSTEMS

D. Cerroni[1], R. Da Vià[1], S. Manservisi[1], F. Menghini[1*], and L. Zaniboni

[1]University of Bologna
Via dei Colli 16, Laboratory of Montecuccolino - DIN

[*] e-mail: filippo.menghini3@unibo.it

Keywords: Adjoint Optimal Control, Fluid-Structure Interaction.

Abstract. *In the last years adjoint optimal control has been increasingly used for design and simulations in several research fields. Applications to Computational Fluid Dynamics problems dedicated to the study of transient-diffusion equations, shape optimization problems, fluid-solid conjugate heat transfer and turbulent flows can be found in literature. The study of Fluid-Structure Interaction problems gained popularity recently because of many interesting applications in engineering and biomedical fields. In this paper we study adjoint optimal control problems for Fluid-Structure Interaction systems in order to improve the advantages of using FSI simulations when designing engineering devices where fluid-dynamical interactions between a fluid and a solid play a significant role. We assess distributed optimal control problems with the purpose to control the fluid behavior by moving the solid region to obtain a desired fluid velocity in specific parts of the domain. The adjoint equations of the FSI monolithic system are derived and the optimality system solved for some simple cases with an in-house finite element code with mesh-moving capabilities for the study of large displacements in the solid. The approach presented in this work is general and can be used to assess different objectives and types of control in future works.*

1 INTRODUCTION

Optimization has always been used to improve the performance of engineering devices. Nowadays several approaches to optimization are available, such as single and multi-objective, adjoint or sensitivities based method, evolutionary algorithms and many others. In literature there are several works dealing with this subject, for a quick review the interested reader can see [1, 2, 3, 4, 5, 6, 7, 8] and references therein.

Fluid-Structure Interaction (FSI) problems can be defined as a system where the fluid flow alters the tensional state of a solid moving shape and the solid deformation has an important effect on the fluid flow. Examples of this type of systems are quite common in engineering, like floating structures, wind turbines, man-made drones and in the study of biological systems such as artery or heart valves. In literature these topics are investigated deeply and the interested reader can see [9, 10, 11, 12, 13, 14]. Many attempts to use optimization techniques can be found, for example in [15], where the authors propose a solution method for the problem of optimizing a non-linear aeroelasticity system in a steady-state flow by using a sensitivity method. The work in [16] deals with general shape optimization methods based on design sensitivity analysis. In [17] a partitioned approach for a NURBS-parametrized shape optimization problem is studied. In the recent work [18] the authors propose a reduced order model, based on a sequential quadratic programming approach, to accelerate the shape optimization process.

In this work we focus on adjoint based optimal control approach which is very popular in industrial applications where Computational Fluid Dynamics simulations should be performed [19, 20, 21, 22, 23]. Moreover this method has a solid mathematical background and the existence of local optimal solutions can be investigated in many interesting cases, see for examples [21, 22, 24]. We are interested in a monolithic approach to the solution of the FSI system leading to a stable and well defined solution in a finite element setting [25, 26]. Inside this framework we study a distributed optimal control problem applied to the monolithic FSI system. The objective of the control is a velocity profile matching in a specific region of the domain. This is accomplished by enforcing a force in the solid domain which deforms the shape and thus the fluid flow profile. In the next section we derive the optimality system which consists of the state, the adjoint system and the control equation. In order to solve the optimality system we propose a simple steepest descent algorithm. In Section 3 we report some numerical results obtained by the implementation of the optimal control algorithm in a finite element parallel code designed for multiphysics simulations.

2 OPTIMALITY SYSTEM

In this section we present the mathematical model of the FSI problem together with the derivation of the optimality system. We use standard notation $H^s(\Omega)$ for the Sobolev spaces with norm $\|\cdot\|_s$ ($H^0(\Omega) = L^2(\Omega)$ and $\|\cdot\|_0 = \|\cdot\|$). Let $H^s_0(\Omega)$ be the space of all functions in $H^s(\Omega)$ that vanish on the boundary Γ_d of the bounded open set Ω and $H^{-s}(\Omega)$ be the dual space of $H^s_0(\Omega)$. The trace space for the functions in $H^1(\Omega)$ is denoted by $H^{1/2}(\Gamma)$.

Let us consider the domain $\Omega \subset R^N$ which consists of a fluid Ω^f and a solid part Ω^s. The solid domain is defined through the solid displacement field $\mathbf{l}$ as

$$\Omega^s(\mathbf{l}) = \{\mathbf{x} \in R^3 \text{ such that } \mathbf{x} = \mathbf{x}_0 + \mathbf{l}\}\,. \tag{1}$$

The position vector $\mathbf{x}_0$ defines the initial solid domain Ω^s_0 before the deformation $\mathbf{l}$. We write $\Omega^s(\mathbf{l})$ for the final domain and $\Omega^s(\mathbf{0}) = \Omega^s_0$ for the initial undeformed domain. In the rest of the paper we drop the notation $(\mathbf{l})$ and $(\mathbf{0})$ over the domain whenever it is not necessary.

The mathematical model of the steady state FSI problem in strong form is

$$\boldsymbol{\nabla} \cdot \mathbf{v} = 0 \qquad \text{on} \quad \Omega^f \,, \tag{2}$$

$$\rho(\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{v} - \boldsymbol{\nabla} \cdot \mathbf{T} - \mathbf{f}_f = 0 \qquad \text{on} \quad \Omega^f \,, \tag{3}$$

$$\boldsymbol{\nabla} \cdot \mathbf{S}(\mathbf{l}) + \mathbf{f}_s = 0 \qquad \text{on} \quad \Omega^s \,, \tag{4}$$

where $\mathbf{v}$ is the fluid velocity, ρ the fluid density, $\mathbf{T}$ the viscous stress tensor of the fluid, $\mathbf{S}$ the strain tensor of the solid, $\mathbf{f}$ a force acting on the solid and Ω^f and Ω^s the fluid and solid domain, respectively. The tensors $\mathbf{T}$ and $\mathbf{S}$ have to be defined by a constitutive model. A simple choice is a Newtonian fluid for $\mathbf{T}$ and a linear elastic solid for $\mathbf{S}$,

$$\mathbf{T}(p, \mathbf{v}) = -p\mathbf{I} + \mu_l \boldsymbol{\nabla}\mathbf{v} \,, \tag{5}$$

$$\mathbf{S}(\mathbf{l}) = \lambda_s(\boldsymbol{\nabla} \cdot \mathbf{l})\mathbf{I} + \mu_s \boldsymbol{\nabla}\mathbf{l} \,, \tag{6}$$

where p is the fluid pressure, μ_l is the dynamic viscosity of the fluid and λ_s and μ_s are the first and second Lamè parameters of the solid. When the parameter $\lambda \to \infty$ the solid becomes incompressible and another equation is needed for the incompressibility constraint. In the following we study incompressible solid materials only.

The system (2-4) has to be completed with appropriate boundary and interface conditions, which are

$$\begin{aligned}
\mathbf{v} &= \mathbf{v_0} && \text{on} \quad \Gamma_{fd} \,, \\
\mathbf{l} &= \mathbf{l_0} && \text{on} \quad \Gamma_{sd} \,, \\
\mathbf{T} \cdot \mathbf{n} &= \mathbf{0} && \text{on} \quad \Gamma_{fn} \,, \\
\mathbf{S} \cdot \mathbf{n} &= \mathbf{0} && \text{on} \quad \Gamma_{sn} \,, \\
\mathbf{T} \cdot \mathbf{n} &= \mathbf{S} \cdot \mathbf{n} && \text{on} \quad \Gamma_{i} \,, \\
\mathbf{v} &= \mathbf{0} && \text{on} \quad \Gamma_{i} \,,
\end{aligned} \tag{7}$$

where Γ_{fd} and Γ_{sd} are the surfaces where a Dirichlet boundary condition is imposed for the fluid velocity and solid displacement, Γ_{fn} and Γ_{sn} are the surfaces where Neumann homogenous conditions are imposed for both the fields and Γ_i is the interface between the solid and the fluid domain, $\Gamma_i = \Omega^f \cap \Omega^s$. In FSI problems the boundary conditions are of primary importance because conditions on the fluid-solid interface can lead to convergence issues so they must be correctly imposed. It is very common to solve FSI problems using dedicated solvers for the fluid and solid domains. This method is called segregated approach and it is widely used in commercial CFD codes. A known drawback of this method is the need to impose interface conditions in an iterative way because the two solvers are coupled through the field values on the interface which give the appropriate boundary conditions to the other solver. This approach can lead to convergence issues, especially when large deformations occur in the solution. On the other hand, by using a monolithic approach with a finite element method, the interface conditions are imposed directly in the same solver and there is no need to iterate to obtain the correct interface values. In this work we use a monolithic approach, so the same solver is used for both the fluid and the solid domain. However the derivation of the optimality system is performed *a priori* and though is valid for both segregated and monolithic approaches.

To derive the optimality system with the Lagrange multiplier method we first establish the objective of the control. In this work we study a fluid velocity matching problem in which the control is a force inside the solid domain. This can be done since we have a complete FSI

formulation in which the fluid and solid domains are considered in the mathematical model definition. The objective or cost functional can be defined

$$\mathcal{J}(\mathbf{v},\mathbf{f}) = \frac{1}{2}\int_{\Omega^d} w\,(\mathbf{v}-\mathbf{v}_d)^2\,d\Omega + \frac{1}{2}\int_{\Omega^c} \beta\,\mathbf{f}^2\,d\Omega\,. \tag{8}$$

The function w is a weight function of the coordinate $\mathbf{x}$ which can be used to set a control in a specific region of the fluid domain. The positive number β is a regularization parameter which is needed in the optimal control formulation to obtain a control function in the space of non singular integrable functions $L^2(\Omega^c)$. If a too high value of β is chosen the control becomes too smooth and the objective cannot be achieved well, while convergence problems can arise in the numerical solution of the problem if β is chosen too low.

In order to derive the optimality system we write the full constrained Lagrangian of the problem which consists of the objective functional and state equations multiplied by the appropriate Lagrangian multipliers

$$\begin{aligned}\mathcal{L}(p,\mathbf{v},\mathbf{l},\widehat{\mathbf{l}},\mathbf{f},p_a,\mathbf{v}_a,\widehat{\mathbf{l}}_a,\widehat{\mathbf{s}}_a,\boldsymbol{\beta}_a) = \mathcal{J}(\mathbf{v},\mathbf{f}) + \int_{\Omega^f}(\nabla\cdot\mathbf{v})\,p_a\,d\Omega + \\ \int_{\Omega^f}[(\rho(\mathbf{v}\cdot\nabla)\mathbf{v}) + \boldsymbol{\nabla} p + \boldsymbol{\nabla}\cdot(\mu_f\boldsymbol{\nabla}\mathbf{v})]\cdot\boldsymbol{v}_a\,d\Omega + \\ \int_{\Omega^s}[-\boldsymbol{\nabla}\cdot(\mu_s\boldsymbol{\nabla}\mathbf{l}+\lambda_s\mathbf{I}(\nabla\cdot\mathbf{l})) - \mathbf{f}]\cdot\boldsymbol{v}_a\,d\Omega + \int_{\Omega^s}\boldsymbol{\nabla}^2\widehat{\mathbf{l}}\cdot\widehat{\mathbf{l}}_a\,d\Omega + \\ \int_{\Gamma_i}\widehat{\mathbf{s}}_a\cdot\left[(\widehat{\mathbf{l}}-\mathbf{l}) + \frac{\mathbf{v}}{h}\right]d\Gamma + \int_{\Omega^s}\boldsymbol{\beta}_a\cdot\left[\mathbf{v}-h(\mathbf{l}-\widehat{\mathbf{l}})\right]d\Omega\,.\end{aligned} \tag{9}$$

In (9) the auxiliary mesh displacement field $\widehat{\mathbf{l}}$ is defined over the solid domain $\Omega^s(\mathbf{l})$, which is the solution of a Laplacian operator. Moreover the velocity field is defined also on the solid domain through the last term in (9). By taking the Frèchet derivatives of (9) with respect to the adjoint variables we recover the state system (2-4) in weak form with the correct boundary and interface conditions. When the derivatives are taken with respect to the state variables and some simplifications are performed on the equations, the adjoint system in weak form reads

$$\int_{\Omega^f}\delta p\boldsymbol{\nabla}\cdot\mathbf{v}_a\,d\Omega = 0 \qquad \forall\delta p\in L^2(\Omega^f)\,, \tag{10}$$

$$\begin{aligned}\int_{\Omega^d}(\mathbf{v}-\mathbf{v}_d)\cdot\delta\mathbf{v}\,d\Omega + \int_{\Omega^s}[\mu_s\boldsymbol{\nabla}\boldsymbol{v}_a:\boldsymbol{\nabla}\delta\mathbf{v} + \lambda_s(\boldsymbol{\nabla}\cdot\boldsymbol{v}_a)(\boldsymbol{\nabla}\cdot\delta\mathbf{v})]\,d\Omega + \\ \int_{\Omega^f}(\boldsymbol{\nabla}\cdot\delta\mathbf{v})\,p_a\,d\Omega + \int_{\Omega^f}[(\rho(\delta\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v})\cdot\boldsymbol{v}_a + (\rho(\mathbf{v}\cdot\boldsymbol{\nabla})\delta\mathbf{v})\cdot\boldsymbol{v}_a + \\ \mu_f\boldsymbol{\nabla}\boldsymbol{v}_a:\boldsymbol{\nabla}\delta\mathbf{v}]\,d\Omega = 0 \qquad \forall\delta\mathbf{v}\in\mathbf{H}^1_{\Gamma_d^f\cup\Gamma_d^s}(\Omega)\,,\end{aligned} \tag{11}$$

The shape derivatives with respect to the fluid and solid domain are taken into account. Since Ω^d and Ω^c are fixed and the system is solved by using a monolithic approach then the shape derivative contribution can be only seen in the $\mathbf{l}_a$ adjoint equation. We do not need the solution of the $\mathbf{l}_a$ adjoint equation to compute the distributed control $\mathbf{f}$ and therefore it is not reported. We remark that the use of a segregated approach, where the solid and the fluid are solved on separated domains, needs to take into account of shape derivatives which appear split into two contributions on different domains. This is one of the main advantage of the monolithic approach.

If one is interested in the strong form of the *adjoint system*, for instance to obtain a proper finite volume discretization, it is necessary to perform some integrations by parts on the terms where the variations δ are differentiated. After performing the integration by parts, we recover the adjoint state $(\mathbf{v}_a^f, \mathbf{v}_a^s, p_a) \in \mathbf{H}^1_{\partial\Omega^f - \Gamma_i}(\Omega^f) \cap \mathbf{H}^2(\Omega^f) \times \mathbf{H}^1_{\partial\Omega^s - \Gamma_i}(\Omega^s) \cap \mathbf{H}^2(\Omega^s) \times L^2_0(\Omega^f) \cap \mathbf{H}^1(\Omega^f)$, by solving

$$\boldsymbol{\nabla} \cdot \mathbf{v}_a^f = 0\,, \tag{12}$$

$$-\rho(\boldsymbol{\nabla}\mathbf{v})^T\mathbf{v}_a^f + \rho[(\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v}_a^f] + \boldsymbol{\nabla} p_a - \boldsymbol{\nabla}\cdot(\mu_l \boldsymbol{\nabla}\mathbf{v}_a^f) = w(\mathbf{v}-\mathbf{v}_d)\,, \tag{13}$$

$$\boldsymbol{\nabla}\cdot\mathbf{S}(\mathbf{v}_a^s) = 0\,. \tag{14}$$

with boundary conditions defined as

$$\mathbf{v}_a^s = \mathbf{v}_a^f \qquad \text{on } \Gamma_i \tag{15}$$

$$\mathbf{S}(\mathbf{v}_a^s)\cdot\mathbf{n} = \left(\mu_l(\boldsymbol{\nabla}\mathbf{v}_a^f + (\boldsymbol{\nabla}\mathbf{v}_a^f)^T) - p_a\right)\cdot\mathbf{n} \qquad \text{on } \Gamma_i \tag{16}$$

$$\mu_l(\boldsymbol{\nabla}\mathbf{v}_a)\cdot\mathbf{n} = -(\mathbf{v}\cdot\mathbf{n})\,\mathbf{v}_a\,, \qquad p_a = 0 \qquad \text{on } \Gamma_{fn}\,. \tag{17}$$

To obtain the definition of the control $\mathbf{f}$ we take the derivative of 9 with respect to $\mathbf{f}$ and we have

$$\mathbf{f} = \mathbf{v}_a/\beta\,. \tag{18}$$

The distributed control is then equal to the adjoint variable $\mathbf{v}_a$ scaled by the regularization parameter β.

We use the following steepest descent algorithm to solve the optimal control problem. In Algorithm 1 we set the tolerance on the optimal control convergence $\tau = 10^{-6}$. The step length r allows us to under-relax the solution of the optimal control problem and to avoid

Algorithm 1 Steepest descent algorithm to find the optimal solution

function FIND OPTIMAL
  set a state $(\mathbf{v}^0, p^0, \mathbf{l}^0)$ satisfying (2-4) ▷ *Setup of the state - Reference case*
  compute the functional $\mathcal{J}^0$ in (8)
  set $\mathbf{f}^0 = 0, r^0 = r_0$
  for $i = 1 \to i_{max}$ **do**
    Solve the system (10)-(11) to obtain the adjoint state $(\mathbf{v}_a^i, p_a^i)$
    for $j = 1 \to j_{max}$ **do**
      compute the control $\mathbf{f}^i = \mathbf{f}^{i-1} + r^{i,j}\,\mathbf{v}_a^i/\beta$
      solve (2-4) for the state $(\mathbf{v}^{i,j}, p^{i,j}, \mathbf{l}^{i,j})$ with the new control $\mathbf{f}^i$.
      compute the new functional $\mathcal{J}^{i,j+1}$ in (8)
      if $\|\mathcal{J}^{i,j+1} - \mathcal{J}^{i,j}\|/\mathcal{J}^{i,j} < \tau$ **then**
        convergence reached ▷ *end of the algorithm*
      else if $\mathcal{J}^{i,j+1} > \mathcal{J}^{i,j}$ **then**
        set $r^{i,j+1} = 2/3\, r^{i,j}$, $j = j+1$ and go to 8 ▷ *loop on j again*
      else if $\mathcal{J}^{i,j+1} < \mathcal{J}^{i,j}$ **then**
        set $r^{i,j+1} = 3/2\, r^{i,j}$, $i = i+1$ and go to 6 ▷ *loop on i again*
      end if
    end for
  end for
end function

strong oscillations. In the loop j, the search of a decreasing functional, we solve iteratively the state system by adjusting this parameter in order to obtain a proper decrease of the functional. When this is accomplished we save the control and we iterate again on the main loop i where the adjoint system is solved. This algorithm requires several solutions of the state and adjoint systems in order to find the optimal control, however it does not need a great amount of memory which is limited to a standard CFD simulation.

We implemented this algorithm in a finite element code which is parallelized with openMPI libraries and uses a multigrid solver with mesh-moving capability [26, 27]. We have used standard Taylor-Hood finite elements for the velocity and pressure solution in order to fulfill the inf-sup condition and a SUPG stabilization technique. The displacements are approximated with standard quadratic elements. In the following we report the results obtained with the proposed algorithm in some simple test cases. First we overview the reference FSI problem with no control, then report the optimal solutions found for different objective velocity profiles and values of β ranging from 10^{-2} to 10^{-6}. The test case setting and the results are presented in a non-dimensional form.

3 NUMERICAL RESULTS

In this section we report the numerical results obtained by applying Algorithm 1 to a two-dimensional test case. The geometry of the test case is shown in Figure 1, where the blue region represents the fluid and the red region of the solid domain. The inlet of the flow is defined by the green line AB, the outlets by the two black lines CD and EF on the right. The overall domain is

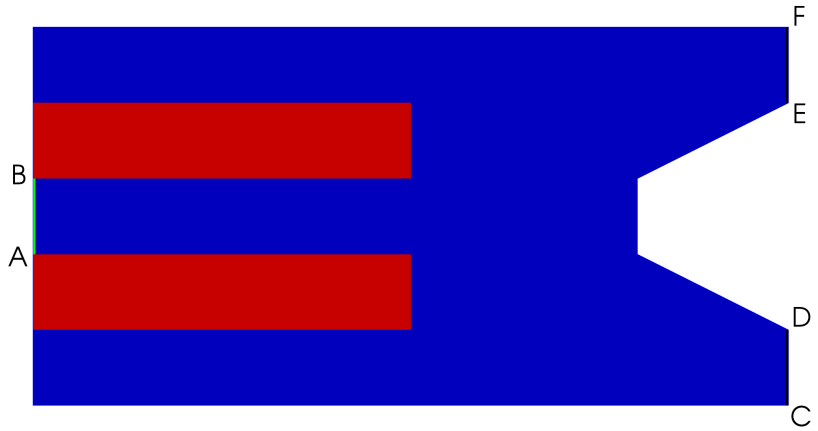


Figure 1: Geometry of the test case. The blue region is the fluid and the red one is the solid domain. On the left the green line AB is the inlet, on the right the black lines CD and EF are the outlets of the flow. The rest of the boundary is a non deformable solid wall.

Property	value	Property	value
E	2×10^2	μ	1×10^{-2}
ρ_s	1	ρ_l	1

Table 1: Physical data for the solid and the fluid.

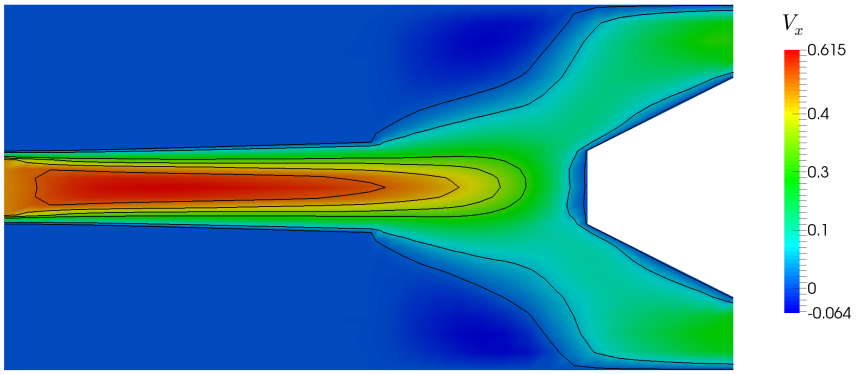


Figure 2: Axial velocity of the reference test case. The velocity v_x is reported with colors and with six lines of iso-magnitude.

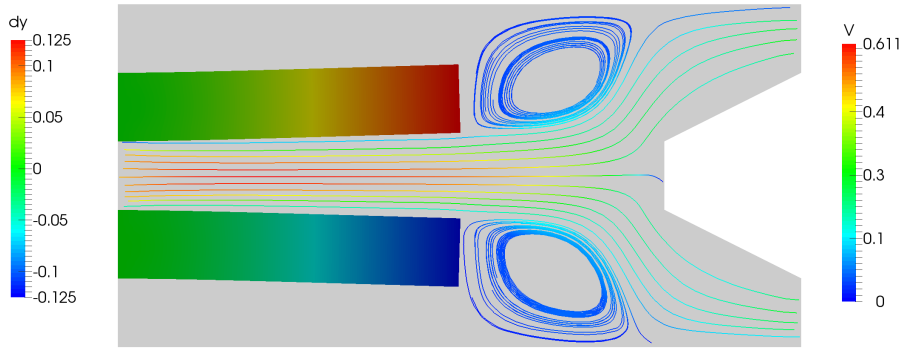


Figure 3: Reference test case, flow field and displacement in the vertical direction. The streamlines are colored with the velocity magnitude and the displacement in the vertical direction dy is reported on the solid bar with colors.

10 wide on the x-axis and 5 high on the y-axis, while lines AB, CD and EF are 1 long each.

The fluid enters from the left with a flat profile, it develops in the region between the two solid bars and hits the solid non deformable wall on the right where the flow is split in two equal secondary flows that leave the domain from the bottom and the outlets on the right. The inlet velocity is set to $\mathbf{v} = (0.5, 0)$ and the other boundaries are considered non deformable solid walls, so the velocity is $\mathbf{v} = (0, 0)$. The solid bars have a boundary that is assumed to be non deformable, so here the solid displacement is set to $\mathbf{l} = (0, 0)$. The physical properties are reported in Table 1. Given these properties, the inlet velocity and the transverse length we can compute the Reynolds number of the flow as $Re = 500$.

We first report the reference flow field without control and then set up the optimal control problem. In Figure 2 the axial velocity v_x is reported with colors and with six lines of iso-magnitude. The flow is accelerated between the two solid bars and then it slows down and split

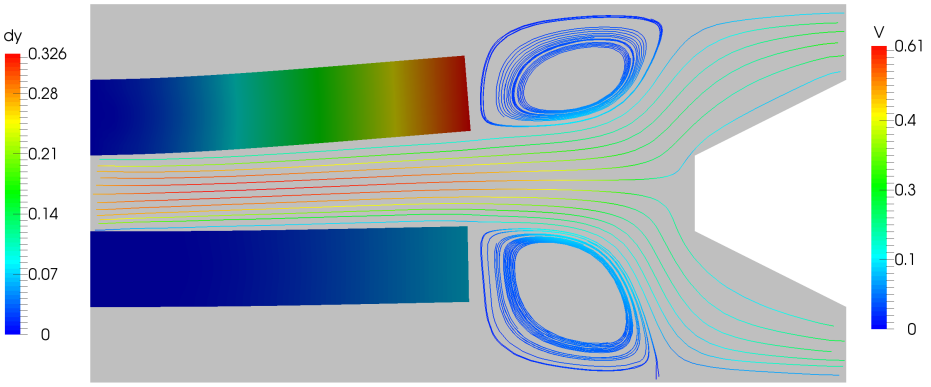


Figure 4: Controlled test case with $\mathbf{v}_{d1}$ and $\beta = 10^{-6}$, flow field and displacement in the vertical direction. The streamlines are colored with the velocity magnitude and the displacement in the vertical direction dy is reported on the solid bar with colors.

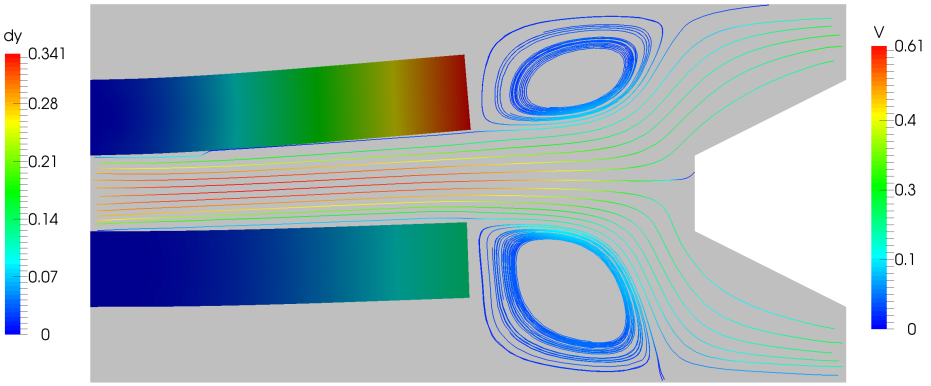


Figure 5: Controlled test case with $\mathbf{v}_{d2}$ and $\beta = 10^{-6}$, flow field and displacement in the vertical direction. The streamlines are colored with the velocity magnitude and the displacement in the vertical direction dy is reported on the solid bar with colors.

before leaving the domain with an average velocity around 0.2. In Figure 3 the streamlines of the velocity field are reported and colored with the magnitude of the velocity field. With this visualization of the flow two slow recirculation vortexes appear in the region near the tips of the solid bars. The solid bars displacement can be seen in this Figure and the value of $\mathbf{l}_y = dy$ is reported with colors on the solid bars. The whole problem is symmetrical with respect to the line $y = 2.5$.

To obtain a proper benchmark for our Algorithm we assign the objective of the functional

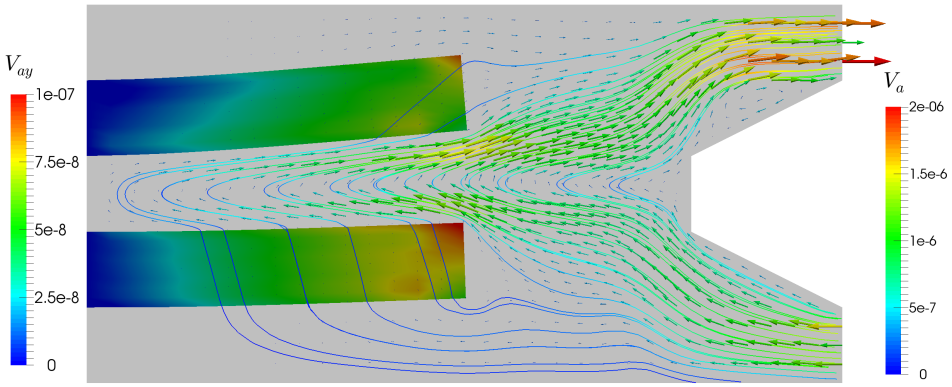


Figure 6: Controlled test case with $\mathbf{v}_{d2}$ and $\beta = 10^{-6}$, adjoint velocity field. The streamlines and arrows are colored with the adjoint velocity magnitude and in the solid bars the vertical adjoint variable is reported with a different color scale.

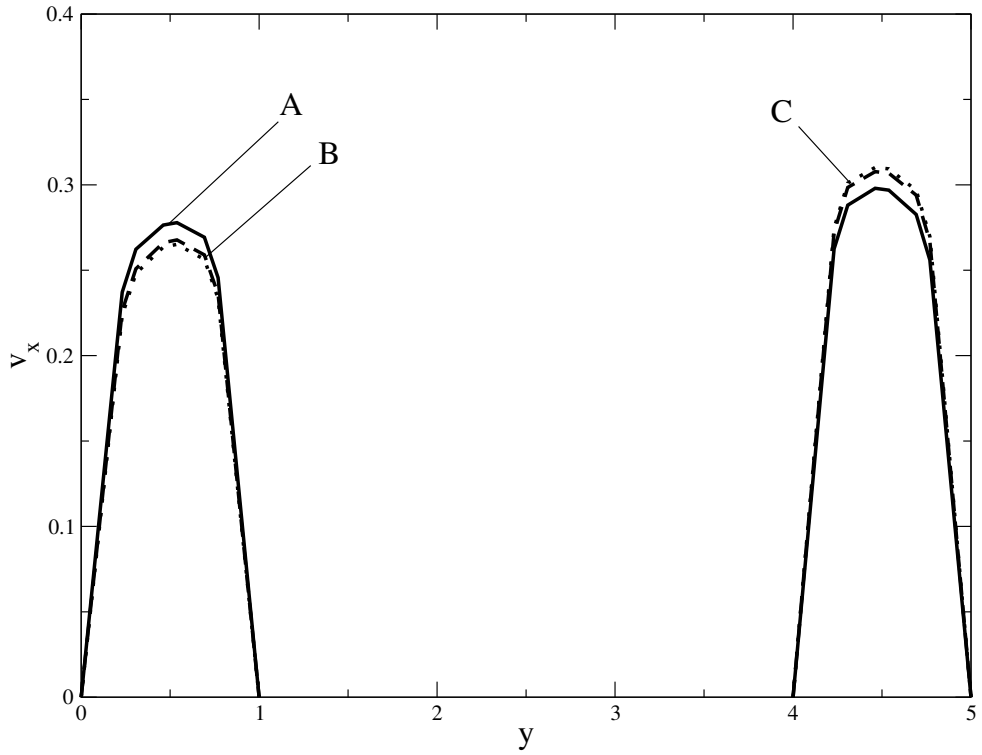


Figure 7: Axial velocity v_x plotted on a vertical line at $x = 10$, comparison between reference case (A, continuous line), test case with $\mathbf{v}_{d1}$ and $\beta = 10^{-6}$ (B, dashed line) and test case with $\mathbf{v}_{d2}$ and $\beta = 10^{-6}$ (C, dotted line).

with a weight function

$$w(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad x > 9, y > 4 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

and set the desired velocity $\mathbf{v}_{d1} = (0.25, 0)$ or $\mathbf{v}_{d2} = (0.4, 0)$ in this region. By doing so we aim at breaking the symmetry of the problem by increasing the axial velocity near the upper outlet.

β	∞	10^{-2}	10^{-4}	10^{-6}
Case with $\mathbf{v}_{d1}$	0.00521914	0.00519039	0.00511677	0.00511527
Case with $\mathbf{v}_{d2}$	0.0312969	0.0297863	0.0295657	0.0295634

Table 2: Objective functionals computed with different β values in the test cases with $\mathbf{v}_{d1}$ and $\mathbf{v}_{d2}$. The reference case with no control is labeled with $\beta = \infty$.

The optimal control problem has been studied for the two desired velocities $\mathbf{v}_{d1}$ and $\mathbf{v}_{d2}$ and for three values of the parameter $\beta = 10^{-2}$, 10^{-4} and 10^{-6}. In Figure 4 the results of the algorithm obtained by using $\mathbf{v}_{d1}$ and $\beta = 10^{-6}$ are reported with streamlines that visualizes the flow field and with the displacement in the vertical direction $\mathbf{l}_y = dy$. The controlling force pushes the bars upper and changes their shape allowing the fluid to be mainly directed towards the upper outlet thus increasing the axial velocity in this region. Also the vortexes change, with the one on the bottom becoming wider than the upper one. In order to improve the control power we can use $\mathbf{v}_{d2}$. In Figure 5 we report the results of the optimal control algorithm by using $\mathbf{v}_{d2}$ and $\beta = 10^{-6}$. The main flow is very similar to the one depicted in Figure 4, the differences lying in a higher displacement in the vertical direction $\mathbf{l}_y = dy$ with the bottom bar displaced too and in a more pronounced flow pattern towards the upper outlet. The adjoint velocity is the controlling force in the solid domain but its source is located in the fluid domain. In Figure 6 this variable is reported with streamlines and arrows colored with its magnitude to visualize the adjoint flow field in the test case with $\mathbf{v}_{d2}$ and $\beta = 10^{-6}$. In the solid bars the vertical adjoint velocity is reported with colors and rescaled. The adjoint flow field enters the domain from the bottom outlet, it develops inside the two bars, crosses them in the internal region and then exits through the upper outlet. One has to keep in mind that the value of the adjoint velocity is related with the value of β since the physical controlling force that is used in the state FSI problem is the adjoint divided by β.

We report in Figure 7 a plot of the axial velocity v_x on a vertical line parallel to the y-axis at $x = 10$, on the outlet. The reference result is reported with a continuous line and it is compared with the controlled result obtained by using $\mathbf{v}_{d1}$ (dashed line) and $\mathbf{v}_{d2}$ (dotted line), both with $\beta = 10^{-6}$. A small asymmetry can be seen in the reference case between the upper and the bottom velocities. However this is negligible compared with the clear velocity increasing that can be seen in the upper outlet in the controlled cases and this is followed by a flow decreasing on the bottom outlet, due to the mass conservation constraint. To gain a more quantitative comparison of the results we report in Table 2 the objective functional (8) as computed in the reference and in the controlled cases, with the different β. One can appreciate the decrease in the functional in both the test cases with decreasing β. We can therefore conclude that the optimal control algorithm is performing well in the search of a minimum of the objective functional. However the movement of the solid cannot increase or decrease in large amount the outlet velocity due to the mass constraint at the inlet.

4 CONCLUSION

In this work a distributed optimal control for the fluid-structure interaction problem has been studied. The objective of the optimal control problem is a velocity matching profile in a specific region of the fluid domain and the control acts through a force in the solid that changes the shape of the solid domain. The optimality system has been derived from the complete Lagrangian of the problem and an algorithm for the numerical solution of this system has been presented. The results obtained show the feasibility of this approach and the possible use of this method in

many industrially relevant applications. However some improvements are needed in the FSI solver because too large deformations with small change in topology are not allowed unless remeshing or more complex techniques are used to handle this problem. With this limitation the optimal control algorithm might not find the local minimum if this is obtained with a large shape change. Nevertheless this algorithm is useful to obtain a reduction in the objective functional that can be a first step in industrial design and the shape cannot be changed dramatically due to other constraints. In future works we plan to assess other objectives and types of control in order to better show the capability of this optimization approach for the design and improvement of engineering devices where the interaction between fluid and solid plays a significant role.

REFERENCES

[1] A. Chinchuluun, P.M. Pardalos, R. Enkhbat, I. Tseveendorj, *Optimization and Optimal Control: Theory and Applications*, Springer Optimization and Its Applications, Springer, 2010.

[2] B.R. Noack, M. Morzynski, G. Tadmor, *Reduced-Order Modelling for Flow Control*, CISM International Centre for Mechanical Sciences, Springer, 2011.

[3] G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, S. Ulbrich, *Constrained Optimization and Optimal Control for Partial Differential Equations*, International Series of Numerical Mathematics, Birkhäuser Mathematics, Springer, 2012.

[4] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operation Research and Financial Engineering, Springer, 2006.

[5] M.D. Gunzburger, *Perspectives in Flow Control and Optimization*, Advances in Design and Control, SIAM, 2003.

[6] C.W. Rowley and D.R. Williams, *Dynamics and Control of High-Reynolds-Number Flow over Open Cavities*, Annual Review of Fluid Mechanics, vol. 38, pp. 251-276, 2006.

[7] J. Kim and T.R. Bewley, *A Linear Systems Approach to Flow Control*, Annual Review of Fluid Mechanics, vol. 39, pp. 383-417, 2007.

[8] M Gunzburger and S Manservisi, *Analysis and approximation for linear feedback control for tracking the velocity in Navier-Stokes flows*, Computer Methods in Applied Mechanics and Engineering, Volume 189 (3), pp. 803-823 (2000)

[9] S. Turek, J. Hron, *Proposal for Numerical Benchmarking of Fluid-Structure Interaction between an Elastic Object and Laminar Incompressible Flow*, Lecture Notes in Computational Science and Engineering, Vol. 53, pp. 371-385, 2006.

[10] L. Formaggia, A. Quarteroni and A. Veneziani, *Cardiovascular Mathematics*, Springer, 2009.

[11] P. Tallec and J. Mouro, *Fluid structure interaction with large structural displacements*, Computer Methods in Applied Mechanics and Engineering, Vol. 190 (24-25), pp. 3039-3067, 2001.

[12] M. Bukaè, S. Èaniè and B. Muha, *A partitioned scheme for fluid-composite structure interaction problems*, Journal of Computational Physics, Vol. 281, pp. 493-517, 2015.

[13] A. Gilmanova, T.B. Lea, F. Sotiropoulos, *A numerical approach for simulating fluid structure interaction of flexible thin shells undergoing arbitrarily large deformations in complex domains*, Journal of Computational Physics, Vol. 300, pp. 814-843, 2015.

[14] A. Calderer, S. Kang, F. Sotiropoulos, *Level set immersed boundary method for coupled simulation of air/water interaction with complex floating structures*, Journal of Computational Physics, Vol. 277, pp. 201-227, 2014.

[15] K. Maute, M. Nikbay and C. Farhat, *Sensitivity analysis and design optimization of three-dimensional nonlinear aeroelastic systems by the adjoint method*, International Journal for Numerical Methods in Engineering, Vol. 56 (6), pp. 911-933, 2003.

[16] E. Lund, H. Moller and L.A. Jakobsen, *Shape design optimization of stationary fluid-structure interaction problems with large displacements and turbulence*, Structural and Multidisciplinary Optimization, Vol. 25 (5), pp. 383-392, 2003.

[17] M. Hojjat, E. Stavropoulou, T. Gallinger, U. Israel, R. Wüchner, K.U. Bletzinger, *Fluid-structure interaction in the context of shape optimization and computational wind engineering*, Fluid Structure Interaction II, Vol. 73, pp. 351-381, 2010.

[18] N. Aghajari and M. Schäfer, *Efficient shape optimization for fluid-structure interaction problems*, Journal of Fluids and Structures, Vol. 57, pp. 298-313, 2015.

[19] Y. Yan and D.E. Keyes, *Smooth and robust solutions for Dirichlet boundary control of fluid-solid conjugate heat transfer problems*, Journal of Computational Physics, vol. 281, pp. 759-786, 2015.

[20] E.M. Papoutsis-Kiachagiasa, S.A. Kyriacoua, and K.C. Giannakogloua, *The continuous adjoint method for the design of hydraulic turbomachines*, Computer Methods in Applied Mechanics and Engineering, Vol. 278, pp. 621-639, 2014.

[21] M. Gunzburger and S. Manservisi, *Analysis and approximation of the velocity tracking problem for Navier-Stokes flows with distributed control*, SIAM Journal of Numerical Analysis, vol. 37 (5), pp. 1481-1512, 2000.

[22] M. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier-Stokes flow with boundary control*, SIAM Journal on Control and Optimization, vol. 39 (2), pp. 594-634, 2000.

[23] S. Manservisi and F. Menghini, *Numerical simulations of optimal control problems for the Reynolds averaged Navier-Stokes system closed with a two-equation turbulence model*, Computers & Fluids, Vol. 125, pp. 130-143, 2016.

[24] S. Manservisi and F. Menghini, *Optimal control problems for the Navier-Stokes system coupled with the k-ω turbulence model*, Computers & Mathematics with Applications, 2015 doi:10.1016/j.camwa.2015.10.003

[25] J. Hron and S. Turek, *A monolithic FEM/multigrid solver for an ALE formulation of fluid-structure interaction with applications in biomechanics* Fluid-Structure Interaction, LNCSE, Vol. 53, pp. 146-170, 2006.

[26] D. Cerroni, S. Manservisi and F. Menghini, *An improved monolithic multigrid Fluid-Structure Interaction solver with a new moving mesh technique*, International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 9, pp. 227-234, 2015.

[27] J.H. Bramble, *Multigrid methods*, Pitman Research Notes in Math, 294, Longman, London, 1993.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

APPLICATION OF WHIRL FLUTTER OPTIMIZATION-BASED SOLUTION TO FULL-SPAN MODEL OF TWIN TURBOPROP AIRCRAFT

Jiri Cecrdle[1]

[1] Aeronautical Research and Test Institute
Beranovych 130, Prague, Czech Republic
e-mail: cecrdle@vzlu.cz

Keywords: Flutter, Whirl Flutter, Optimization, L-410NG aircraft.

Abstract. *This paper deals with whirl flutter analysis of aircraft structures. It gives the theoretical background of the whirl flutter phenomenon and aircraft certification-related issues. After that, the ordinary and optimization-based analytical approaches are described. The optimization based analytical procedure is used to determine the whirl flutter stability boundaries for the certification speed. The focus is on the application of the optimization-based approach on a full-span model of an aircraft structure. The necessary adjustments of a stick computational model to make it applicable as a full-span and also the modification of the optimization solution are described. The methodology is demonstrated on the model of a twin-engine turboprop commuter aircraft. The evaluated results include the available choices of both propeller rotation directions, i.e. CW-CW and CW-CCW. Finally, the outlook and the future work on the described topic is outlined.*

1 INTRODUCTION

Turboprop aircraft are certified considering the whirl flutter stability. Whirl flutter is a specific type of flutter instability caused by the effect of rotating parts as a propeller or a gas turbine engine rotor. Rotating mass causes additional forces and moments and increases the number of degrees-of-freedom. Rotating propeller also causes aerodynamic interference effect with a nacelle and a wing structure. Whirl flutter instability is driven by motion-induced unsteady aerodynamic propeller forces and moments acting in the propeller plane. It is quite serious phenomenon, that may cause unstable vibration of a propeller mounting, even a failure of an engine installation or a whole wing. Airworthiness regulations require flutter analysis taking into account the influence of the degrees of freedom of the propeller plane of rotation and significant elastic, inertia and aerodynamic forces. Also the changes in the stiffness and damping of the propeller–engine–nacelle structure system must be considered (FAR/CS §23.629(e)(1)(2)) [1, 2].

The reliable stiffness data regarding the engine attachment are not usually at disposal until the ground vibration test of the prototype is done and the final updating of the analytical model is possible. Nevertheless, considering timesaving in the final development phase, it is worth performing whirl flutter calculations in the earlier phase. For this purpose, an optimization based analytical procedure to determine the critical values of an engine attachment stiffness parameters was prepared. It allows determination of whirl flutter and whirl divergence stability boundaries for the speed, which is set by regulations as the certification speed. The solution employs the gradient-based algorithm and includes modal-based and flutter-based design responses. Design variables are represented by the engine attachment stiffness parameters.

The initial solution was prepared for the half-span models of aircraft with either symmetric or antisymmetric boundary condition, which are usually used for flutter analyses. But, the applicability of half-span models is limited and application of a full-span model is necessary in some cases. The typical example is the whirl flutter. Considering the usual twin wing-mounted engine aircraft concept, one of the parameters influencing the whirl flutter stability is the direction of rotation of both propellers. Therefore, the applicability of the optimization-based solution was enlarged also to the full-span models. The optimization solution for a full-span model is, compare to a half-span model, more complicated. Also, the specific adjustment of a structural model is necessary. The solution is demonstrated on the example of the reference model of the twin-engine turboprop commuter aircraft. Whirl flutter stability boundaries are constructed for different conditions of propeller rotations.

2 THEORETICAL BACKGROUND

The principle of whirl flutter phenomenon is outlined on the simple mechanical system with two degrees-of-freedom. Propeller and hub are considered as rigid. An engine flexible mounting is represented by two rotational springs (stiffness K_Ψ, K_Θ) as illustrated in figure 1. Such a system has two independent mode shapes (yaw and pitch) with angular frequencies ω_Ψ and ω_Θ. Considering a propeller rotation with the angular velocity Ω, the primary motion changes and the gyroscopic effect makes both independent mode shapes merge into the whirl motion. A propeller axis of rotation shows an elliptical movement. A trajectory of this elliptical movement depends on both angular frequencies ω_Ψ and ω_Θ. The orientation of the gyroscopic movement is backward relative to the propeller rotation for the mode with the lower frequency (backward whirl mode) and forward relative to the propeller rotation for the mode with the higher frequency (forward whirl mode). Because the yaw and pitch motions have a 90° phase shift, the mode shapes in the presence of gyroscopic effects are complex.

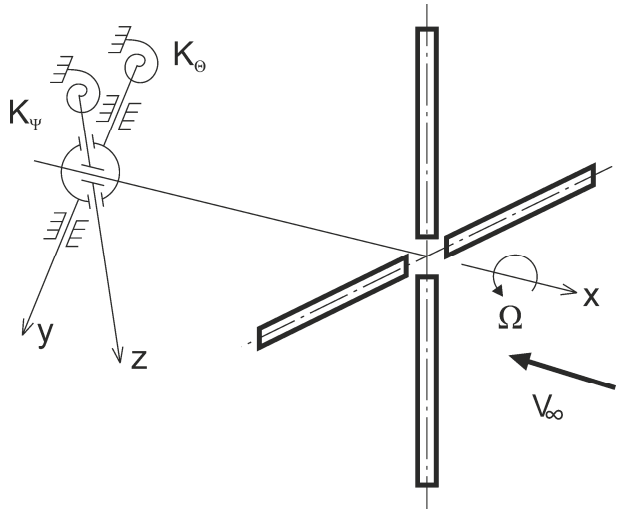


Figure 1. Gyroscopic system with propeller.

The gyroscopic motion results in changes of the propeller blades' angles of attack and consequently, leading to unsteady aerodynamic forces. These forces may under specific conditions induce whirl flutter instability. The flutter state is defined as the neutral stability with no damping of the system and the corresponding airflow ($V_\infty = V_{FL}$) is called critical flutter speed. If the air velocity is lower than flutter speed ($V_\infty < V_{FL}$), the system is stable and the gyroscopic motion is damped (see figure 2a). If the airspeed exceeds the flutter speed ($V_\infty > V_{FL}$), the system becomes unstable, and gyroscopic motion divergent (see figure 2b).

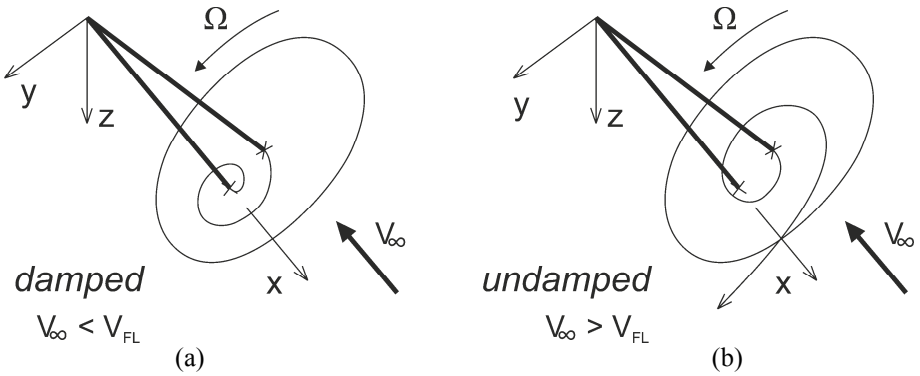


Figure 2. Stable (a) and unstable (b) state of gyroscopic vibrations for backward flutter mode.

The analytical solution is aimed to determine the aerodynamic force caused by the gyroscopic motion on each of propeller blades. Presented equations of motion were derived for the system shown in figure 1 using Lagrange's approach. The kinematical scheme is shown in figure 3. We select three angles (φ, Θ, Ψ) as the independent generalised coordinates. The propeller angular velocity is considered constant ($\varphi = \Omega\, t$). The rotating part is assumed cyclically symmetric with respect to both mass and aerodynamics (i.e., propeller of three blades in minimum). Non-uniform mass moments of inertia of the engine with respect to pitch and yaw axes ($J_Z \neq J_Y$) are considered.

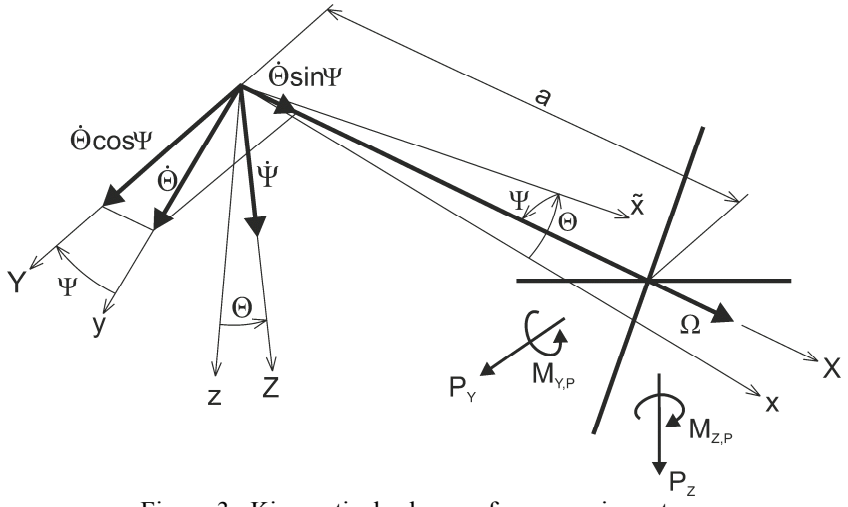


Figure 3. Kinematical scheme of gyroscopic system.

Considering small angles, the equations of motion become:

$$\begin{aligned} J_Y\ddot{\Theta} + \frac{K_\Theta \gamma_\Theta}{\omega}\dot{\Theta} + J_X \Omega \dot{\Psi} + K_\Theta \Theta &= M_{Y,P} - a.P_Z \\ J_Z\ddot{\Psi} + \frac{K_\Psi \gamma_\Psi}{\omega}\dot{\Psi} - J_X \Omega \dot{\Theta} + K_\Psi \Psi &= M_{Z,P} + a.P_Y \end{aligned} \tag{1}$$

where a is the distance between a propeller plane and a vibration mode node point. Neglecting the aerodynamic inertia terms ($\dot{\Theta}^* \approx \dot{\Theta}$; $\dot{\Psi}^* \approx \dot{\Psi}$), the propeller aerodynamic forces and moments at the propeller plane (P_Y; P_Z; $M_{Y,P}$; $M_{Z,P}$) are calculated as:

$$\begin{aligned} P_Y &= \pi\,\rho\,V_\infty^2 R^2 \left(c_{y\Theta}\Theta^* + c_{y\Psi}\Psi^* + c_{yq}\frac{\dot{\Theta}^* R}{V_\infty} + c_{yr}\frac{\dot{\Psi}^* R}{V_\infty} \right) \\ P_Z &= \pi\,\rho\,V_\infty^2 R^2 \left(c_{z\Theta}\Theta^* + c_{z\Psi}\Psi^* + c_{zq}\frac{\dot{\Theta}^* R}{V_\infty} + c_{zr}\frac{\dot{\Psi}^* R}{V_\infty} \right) \\ M_{Y,P} &= 2\pi\,\rho\,V_\infty^2 R^3 \left(c_{m\Theta}\Theta^* + c_{m\Psi}\Psi^* + c_{mq}\frac{\dot{\Theta}^* R}{V_\infty} + c_{mr}\frac{\dot{\Psi}^* R}{V_\infty} \right) \\ M_{Z,P} &= 2\pi\,\rho\,V_\infty^2 R^3 \left(c_{n\Theta}\Theta^* + c_{n\Psi}\Psi^* + c_{nq}\frac{\dot{\Theta}^* R}{V_\infty} + c_{nr}\frac{\dot{\Psi}^* R}{V_\infty} \right) \end{aligned} \tag{2}$$

where ρ is an air density and R is a propeller radius. Aerodynamic derivatives (c-terms) are given from the propeller blade integrals (by [3, 4] or by [5]). Using the quasi-steady theory [6], the effective angles (Θ^*, Ψ^*) become:

$$\begin{aligned} \Theta^* &= \left(\Theta + \frac{\dot{z}}{V_\infty} \right) - \frac{w_1}{V_\infty} \\ \Psi^* &= \left(\Psi - \frac{\dot{y}}{V_\infty} \right) + \frac{w_2}{V_\infty} \end{aligned} \tag{3}$$

where (w_1/V_∞) and (w_2/V_∞) are optional downwash and sidewash terms. Downwash and sidewash angles behind the propeller describe the interference between propeller and nacelle.

Induced downwash and sidewash angles are added to the effective static angles, which are represented by the remaining terms of eq. (3). Induced downwash and sidewash angles, which are dependent on the reduced frequency, can be obtained from the lift solution by partitioning the interference coefficients. The downwash effect influences the aerodynamic stiffness matrix; the influence to the aerodynamic damping matrix is neglected. An option to include the downwash and sidewash effects may be important for the aircraft configuration with wing-mounted engines.

Finally, seeking for the critical (flutter) state assuming the harmonic motion has a character of an eigenvalue problem. The final whirl flutter matrix equation can be expressed as:

$$\left(-\omega^2[M]+j\omega\left([D]+[G]+q_\infty\frac{4\pi R^4}{V_\infty}[D^A]\right)+\left([K]+q_\infty 2\pi R^3[K^A]\right)\right)\begin{bmatrix}\overline{\Theta}\\ \overline{\Psi}\end{bmatrix}=\{0\} \tag{4}$$

where [M] is mass matrix, [D] is structural damping matrix, [K] is structural stiffness matrix. $[D^A]$ and $[K^A]$ are aerodynamic damping and aerodynamic stiffness matrix, respectively. Finally, [G] is gyroscopic matrix. The critical state emerges when the angular velocity ω is real. The critical state can be reached by increasing either V_∞ or Ω. Increasing the propeller advance ratio (V_∞ / (ΩR)) has destabilizing effect. Another important parameter is distance of propeller plane and node points of engine vibration modes. Structural damping is a significant stabilization factor. The paper [7] describes an experiment during which friction was deeply suppressed and whirl flutter occurred. However, the notably slight structural damping made the model stable. By contrast, the influence of the propeller thrust is negligible. This small influence comes from the fact that the aerodynamic derivatives of the thrusted propeller and windmilling propeller variance can be high in the low speed region, but at high velocities (where whirl flutter is expected), the variance is less than 5% [8]. The most critical state is $\omega_\Theta = \omega_\Psi$ when the interaction of both independent motions is maximal and the trajectory of the gyroscopic motion is circular. Considering the rigid propeller blades, the whirl flutter inherently appears at the backward gyroscopic mode. A special case of eq. (4) for $\omega = 0$ is gyroscopic static divergence, which is characteristic by uni-directional divergent motion.

The described model that considers a rigid propeller is obviously applicable to conventional propellers, for which the propeller blade frequencies are much higher compared to the nacelle pitch and yaw frequencies. Considering the large multi-bladed propellers of heavy turboprop aircraft, the consideration of a rigid propeller appears too conservative and the blade flexibility must also be modeled. Obviously, the whirl flutter investigation of tilt-rotor aircraft must include even more complex analytical models [9].

The comprehensive information regarding the whirl flutter phenomenon can be found in [10].

3 STANDARD ANALYTICAL APPROACH

The standard whirl flutter solution is based on the strip aerodynamic theory [11] for the propeller at the windmilling mode. The propeller is assumed rigid. For the residual structure the unsteady doublett-lattice method including wing–body interference aerodynamic theory is used [12]. For the flutter stability solution the p-k method [13] is applied. The basic flutter equation in modal coordinates is:

$$\left[[M_{hh}]\lambda^2+\left([B_{hh}]-\frac{1}{4}\rho\,\bar{c}\,V_\infty\frac{[Q_{hh}^{\mathrm{Im}}]}{k}\right)\lambda+\left([K_{hh}]-\frac{1}{2}\rho\,V_\infty^2\,[Q_{hh}^{\mathrm{Re}}]\right)\right]\{u_h\}=0 \tag{5}$$

$[M_{hh}]$, $[B_{hh}]$ and $[K_{hh}]$ are modal mass, damping and stiffness matrices, respectively, a function of the Mach number (M) and the reduced frequency (k). Aerodynamic loads are incorporated into damping and stiffness matrices. $[Q_{hh}^{Re}]$ and $[Q_{hh}^{Im}]$ are the real and the imaginary part of a complex aerodynamic matrix, also a function of parameters M and k. The parameter ρ is the air density, $\overline{c}$ is a reference length, and $\{u_h\}$ is a modal amplitude vector. The eigenvalue λ is given as:

$$\lambda = \omega\left(\gamma \pm j\right) \tag{6}$$

and γ is a transient decay rate coefficient. Note that the structural damping coefficient (g) is expressed as:

$$g = 2.\gamma \tag{7}$$

The standard whirl flutter solution is performed for multiple velocities. The resulting quantities are V-g-f curves, i.e. the dependence of the damping and frequency of analyzed modes on the flight velocity. The state with the zero damping represents the critical flutter state and the corresponding flight velocity is the critical flutter speed.

4 OPTIMIZATION-BASED ANALYTICAL APPROACH

Optimization-based approach employs the gradient-based algorithms [14] for the whirl flutter solution. It makes possible the calculation of the flutter stability boundaries for the specified certification speed. In this case, the flutter speed is set equal to the certification speed, and the results are critical values of the structural parameters. The stability margin can be then obtained from these critical structural parameters. The analyzed states are then compared only with respect to the structural parameters and the relationship to the stability margin. Such an approach can save large amounts of time because the number of whirl flutter analyses required by the regulations is dramatically reduced.

The whirl flutter optimization employs two types of the design responses (eigenvalue and flutter). The eigenvalue equation is:

$$\left(\left[K\right]-\lambda_n\left[M\right]\right)\left\{\phi_n\right\}=0 \tag{8}$$

where λ_n and ϕ_n are the nth eigenvalue and eigenvector, respectively. [K] is the structural stiffness and [M] is the structural mass matrix. Eq. (8) can be differentiated with respect to the ith design variable x_i:

$$\left(\left[K\right]-\lambda_n\left[M\right]\right)\frac{\partial\left\{\phi_n\right\}}{\partial x_i}+\left(\frac{\partial\left[K\right]}{\partial x_i}-\lambda_n\frac{\partial\left[M\right]}{\partial x_i}\right)\left\{\phi_n\right\}=\frac{\partial\lambda_n}{\partial x_i}\left[M\right]\left\{\phi_n\right\} \tag{9}$$

When eq. (9) is premultiplied by ϕ_n^T, the first term becomes zero and eq. (9) can then be solved for the eigenvalue derivatives:

$$\frac{\partial\lambda_n}{\partial x_i}=\frac{\left\{\phi_n\right\}^T\left(\frac{\partial\left[K\right]}{\partial x_i}-\lambda_n\frac{\partial\left[M\right]}{\partial x_i}\right)\left\{\phi_n\right\}}{\left\{\phi_n\right\}^T\left[M\right]\left\{\phi_n\right\}} \tag{10}$$

In practice, the solution of eq. (10) is based on the semi-analytical approach. The derivatives of the mass and stiffness matrices are approximated using the finite differences.

The equation is solved for each retained eigenvalue referenced in the design model and for each design variable.

Aeroelastic flutter stability matrix equation is given by eq. (5). The equation represents the p-k method of the flutter solution, which is the only method applicable for the purpose of the design optimization. Flutter sensitivity computes the rates of change of the transient decay rate coefficient γ with respect to changes of the design variables. Eq. (5) is differentiated with respect to the design variables for the quantity $(\partial\gamma/\partial x_i)$. The solution is semi-analytical in nature with derivatives approximated using either forward differences or central differences.

Flutter sensitivities are computed as the rate of change of the transient decay coefficient γ with respect to changes in design variables $(\partial\gamma/\partial x_i)$.

The optimization-based whirl flutter solution is performed for a single velocity. The resulting quantities are structural parameters, for which the flutter speed is equal to the specified certification speed.

5 STRUCTURAL MODEL OF AIRCRAFT STRUCTURE

5.1 Half-span Model

Aeroelastic analyses of aircraft structures are usually performed using simple dynamic structural models (stick models). Stiffness characteristics of the structural parts are modeled by means of the massless beam elements, and inertial characteristics are modeled by concentrated mass elements including appropriate moments of inertia. The model also includes spring elements, various conditions and auxiliary elements (controls suspension, visualization, etc.). In the most cases, the analysis can be performed using half-span model. In this case, the half-values of the stiffness and inertial characteristics are applied at the plane of symmetry as well as either symmetric or antisymmetric boundary condition. Such a model is shown in figure 4. Attachment of the engine to the wing is realized simply by means of two spring elements, which model engine pitch and yaw vibration modes. Stiffness constant of the spring determines the natural frequency of the mode. Spring element is stationed at the node point of the mode.

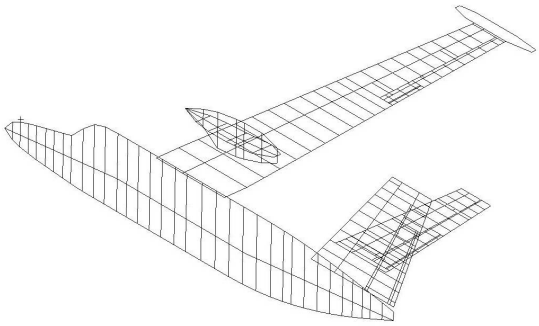


Figure 4. Half-span structural model of twin engine turboprop commuter aircraft.

5.2 Full-span Model

The applicability of half-span models is limited and application of a full-span model is necessary in some specific cases. In the case of the full-span model, both symmetric and antisymmetric engine vibration modes must be modeled. Such a model is shown in figure 5.

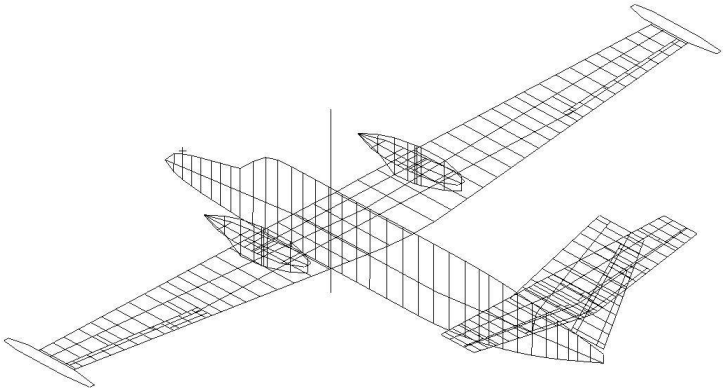
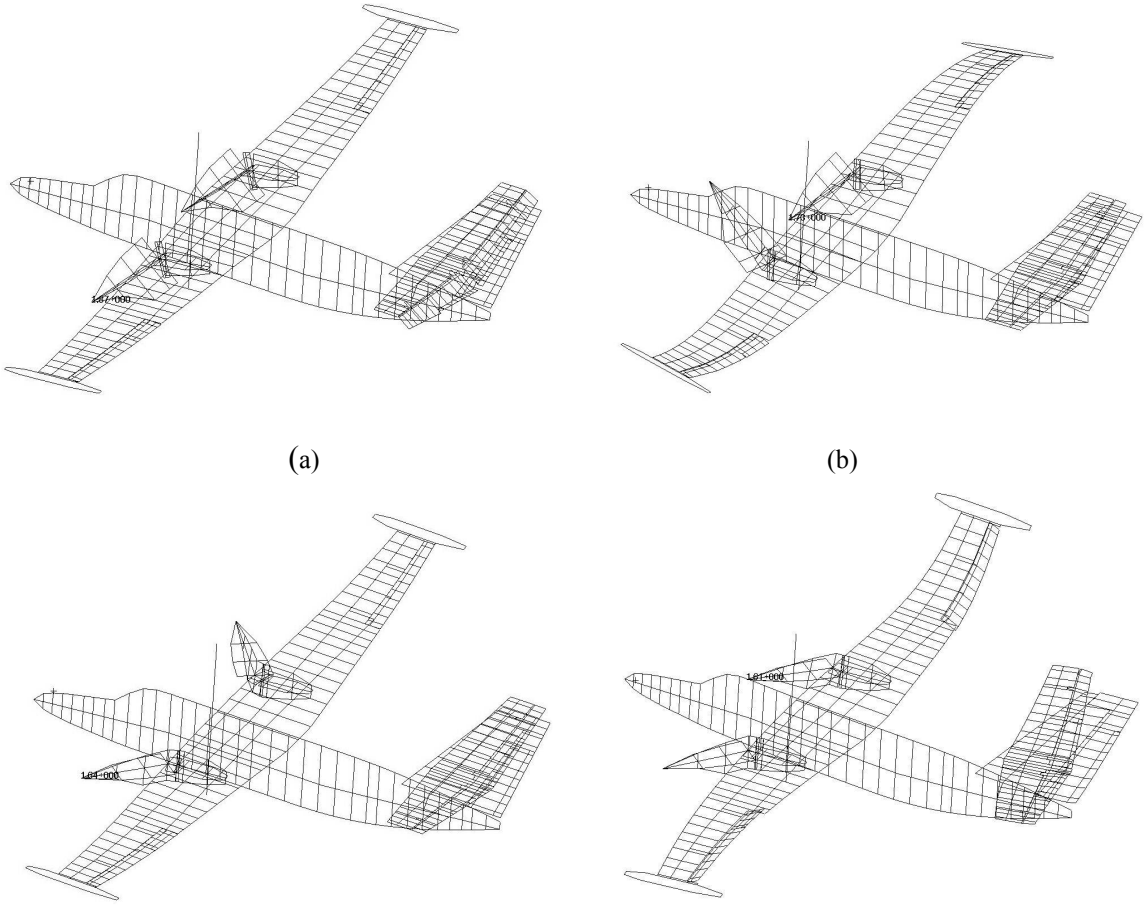


Figure 5. Full-span structural model of twin engine turboprop commuter aircraft.

The engine attachment includes four modes with diverse natural frequencies and diverse node points. The typical order of engine modes (by frequency) is 1) symmetric pitch, 2) symmetric yaw, 3) antisymmetric pitch and, 4) antisymmetric yaw. The node points of these modes are typically stationed in the direction from the rear to the front (in the flight direction).

The systems to model pitch and yaw engine vibration modes are separate. The appropriate rotational degree-of-freedom (i.e., around lateral or around vertical axis) is connected to the central system, which consists of the grounded spring element and two rod elements. The grounded spring is placed at the plane of symmetry at the station of the node of the symmetric mode while rod elements are placed at the station of the node point of the antisymmetric mode. Rod elements are oriented in the appropriate direction (i.e., laterally or vertically). Node point of the grounded spring is connected with the central node of rod elements by means of multi-point constraint. Apart from the appropriate degree-of-freedom, other ones are omitted from the analysis. Spring constant of the grounded spring element K_δ and the torsional stiffness of rod elements (GI_k) then determine natural frequencies of both symmetric and antisymmetric mode. Spring constant is decisive for the symmetric mode frequency, while rod torsional stiffness is decisive for the antisymmetric frequency. However, there is also cross-influence; and therefore, both parameters must be used to set both frequencies. The examples of mentioned engine vibration modes are shown in figure 6.

Note, that also both symmetric and antisymmetric control surface and tab flapping modes must be modeled on the full-span models. However, no detailed description of the modeling technique is provided here as these modes are not important from the whirl flutter phenomenon point of view.



(a) (b)

(c) (d)

Figure 6. Engine vibration modes: (a) symmetric pitch, (b) antisymmetric pitch, (c) symmetric yaw, (d) antisymmetric yaw.

6 DESCRIPTION OF OPTIMIZATION-BASED SOLUTION

The important parameters influencing the whirl flutter stability are engine pitch and yaw frequencies. The solution is therefore demonstrated on the example of variation of these parameters. We consider the inertia characteristics of the engine and propeller system as well as the characteristics of the residual structure to be reliably determined; thus, we will use the engine attachment stiffness properties as parameters for the optimization.

6.1 Solution for Half-span Model

We define two design variables: 1) effective stiffness of the engine attachment in pitch and 2) effective stiffness of the engine attachment in yaw. These design variables are directly related to the spring constants of two spring elements (pitch-$K_{\varphi V}$ and yaw-$K_{\varphi H}$).

The first preparatory step is intended to set the initial design variables for the main optimization. The target frequency ratio (TFR) is set, and both effective stiffnesses are adjusted to reach this target ratio. The simplest way is to set any of the mentioned stiffnesses as the design variable (while the other one is fixed) and use the optimization solution with the objective function (OBJ) defined as:

$$OBJ = min\left[ABS\left(\frac{f_2}{f_1} - TFR \right) \right] \tag{11}$$

where ABS denotes for the absolute value. Symbols f_1 and f_2 represent both pitch and yaw engine frequencies (f_2 is the higher frequency while f_1 is the lower frequency). Note that the yaw frequency is usually higher compare to the pitch frequency; nevertheless, the solution may be done regardless the frequency order. This preparatory analysis gives the initial values of $K_{\varphi V}$ and $K_{\varphi H}$ for the main optimization. The ratio of both frequencies (pitch-f_Θ and yaw-f_Ψ) is equal to the TFR.

The main optimization includes both design variables ($K_{\varphi V}$ and $K_{\varphi H}$). The design constraints include the requirement to keep the target frequency ratio. For the practical applications, specification of a constraint with the ±2% band is usually used:

$$ABS\left(\frac{\left(\frac{f_2}{f_1}-TFR\right)}{TFR}\right)<0.02 \tag{12}$$

Another constraint includes the requirement to keep the flutter stability (i.e., negative damping) at the selected certification speed V_{cert}. This requirement is expressed as:

$$g\left(V=V_{cert}\right)<0 \tag{13}$$

In the practical solution, the constraint is modified as:

$$-\infty<\left(\frac{g\left(V=V_{cert}\right)-0.03}{0.1}\right)<-0.3 \tag{14}$$

The interval shift from the null value is given due to the numerical character of the solution preventing the division by zero. This constraint should also prevent another type of flutter instability below the certification speed that may be caused by the design variable changes. The constraint should therefore be applied to all modes included in the solution.

The objective function is defined simply as minimization of pitch and yaw frequency sum as:

$$OBJ=min\left(f_1+f_2\right) \tag{15}$$

As the output, we will obtain the engine pitch and yaw stiffness for which the flutter speed is equal to the specified certification speed and the yaw-to-pitch frequency ratio is equal to specified target value. The optimization is then repeated for a several yaw-to-pitch frequency ratios, typically ranging from 1.05 to 2.0 to get enough points to construct a stability boundary curve. Note that the described solution is applicable for the whirl divergence as well.

The procedure described above is applicable to the "no downwash" option (see section 2). Provided the downwash effects are to be included, the procedure is extended by the extra steps described here. The basic assumption is, that the minor change in the engine pitch and yaw frequencies will have negligible effect on the downwash terms. The downwash terms are therefore calculated only for the structure optimized excluding the downwash as described above. After that, the downwash terms are calculated and the main optimization step is repeated in the same manner as described above.

Due to the possibility of switching of the engine pitch or yaw mode frequency with the other modes, it is worth to re-order the modes at the end of the optimization iteration, provided such mode switch appear. The re-ordering is based on the cross-orthogonality correlation analysis of both sets of modes (before and after the optimization iteration) using the modal assurance criterion, which is expressed as:

$$MAC(\psi_1,\psi_2)=\frac{\left|\left(\{\psi_1\}^t\{\psi_2\}\right)\right|^2}{\left(\{\psi_1\}^t\{\psi_1\}\right)\left(\{\psi_2\}^t\{\psi_2\}\right)} \tag{16}$$

where ψ_1 and ψ_2 are correlated mode shapes. Only engine pitch or yaw modes are switched, if necessary.

6.2 Solution for Full-span Model

The optimization solution for a full-span model is more complicated. We define four design variables: 1) effective stiffness of the engine attachment for symmetric pitch, 2) effective stiffness of the engine attachment for antisymmetric pitch, 3) effective stiffness of the engine attachment for symmetric yaw and 4) effective stiffness of the engine attachment for antisymmetric yaw. These design variables are related to the spring constants of two grounded spring elements ($K_{\delta 1}$ and $K_{\delta 2}$) and to torsional stiffness of two pairs of rod elements [$(GI_k)_1$ and $(GI_k)_2$]. However, the relation is not direct here due to the above mentioned cross-influences (see section 5.2).

We define three frequency ratios: 1) pitch frequency ratio (VFR = $f_{A\Theta}/f_{S\Theta}$), 2) yaw frequency ratio (HFR = $f_{A\Psi}/f_{S\Psi}$) and finally 3) critical frequency ratio (CFR). Critical modes are those ones, the combination of which causes a flutter instability. Choice of the critical modes is dependent on the relation of directions of rotation of both propellers, and on the mode order. Considering the identical directions (i.e., CW-CW or CCW-CCW), the critical modes are symmetric pitch and antisymmetric yaw (it corresponds to a half-span symmetric case) or antisymmetric pitch and symmetric yaw (it corresponds to a half-span antisymmetric case). Considering the inverse directions (i.e., CW-CCW or CCW-CW), the critical modes are symmetric pitch and symmetric yaw or antisymmetric pitch and antisymmetric yaw. Note that CW denotes for the clockwise direction and CCW denotes for the counter-clockwise direction.

VFR and HFR are not changeable. The values are set according the ground vibration test results or guessingly. The typical ratios are ranging from 1.12 to 1.18. CFR is an analogy of TFR shown above.

We assume the typical frequency order (i.e. pitch frequency lower compare to the yaw frequency) in the following description. The first preparatory step is intended to set the initial design variables for the main optimization.

The design constraint includes the requirement to keep the frequency of the selected engine vibration mode, typically the highest one ($f_{A\Psi}$), at the selected value ($f_{A\Psi T}$) using the ±2% band as:

$$ABS\left(\frac{(f_{A\Psi}-f_{A\Psi T})}{f_{A\Psi T}}\right)<0.02 \tag{17}$$

The objective function is defined as the minimization of the frequency ratio error expressed as:

$$OBJ=min\left\{SSQ\left[\left(\frac{f_{A\Theta}}{f_{S\Theta}}-VFR\right),\left(\frac{f_{A\Psi}}{f_{S\Psi}}-HFR\right),\left(\frac{f_{A\Psi}}{f_{S\Theta}}-CFR\right)\right]\right\} \tag{18}$$

or as:

$$OBJ=min\left\{SSQ\left[\left(\frac{f_{A\Theta}}{f_{S\Theta}}-VFR\right),\left(\frac{f_{A\Psi}}{f_{S\Psi}}-HFR\right),\left(\frac{f_{S\Psi}}{f_{S\Theta}}-CFR\right)\right]\right\} \tag{19}$$

where SSQ denotes for sum of squares. Eq. (18) is applicable for the case of identical directions of propeller rotations while eq. (19) is applicable for the case of inverse directions of propeller rotations. This preparatory analysis gives the initial values of design variables for the main optimization. Frequency ratios are equal to VFR, HFR and CFR values.

The main optimization is performed similarly as for the half-span model. The design constraints include the requirement to keep the frequency ratios, again, with the ±2% band as:

$$ABS\left(\frac{\left(\frac{f_{A\Theta}}{f_{S\Theta}}-VFR\right)}{VFR}\right)<0.02 \tag{20}$$

$$ABS\left(\frac{\left(\frac{f_{A\Psi}}{f_{S\Psi}}-HFR\right)}{HFR}\right)<0.02 \tag{21}$$

and

$$ABS\left(\frac{\left(\frac{f_{A\Psi}}{f_{S\Theta}}-CFR\right)}{CFR}\right)<0.02 \tag{22}$$

or

$$ABS\left(\frac{\left(\frac{f_{S\Psi}}{f_{S\Theta}}-CFR\right)}{CFR}\right)<0.02 \tag{23}$$

Again, eq. (22) is applicable for the case of identical directions of propeller rotations while eq. (23) is applicable for the case of inverse directions of propeller rotations.

Another constraint including the requirement to keep the flutter stability (i.e., negative damping) at the selected certification speed V_{cert} is expressed in the same way as for the half-span model, i.e. by eq. (14).

The objective function is defined as minimization of frequency sum, here expressed as:

$$OBJ = min\left[\, SUM\left(f_{S\Theta}, f_{A\Theta}, f_{S\Psi}, f_{A\Psi}\right)\right] \tag{24}$$

As the output, we will obtain the values of design variables ($(K_{\delta 1}$; $K_{\delta 2}$; $(GI_k)_1$; $(GI_k)_2$), for which the flutter speed is equal to the specified certification speed and three specified frequency ratios are equal to the specified target values. Similarly as for the half-span model, the optimization is then repeated for a several CFR values, typically ranging from 1.05 to 2.0 to get enough points to construct a stability boundary curve. Also, the notes regarding the divergence, downwash effect and regarding the mode switches mentioned in section 6.1 are valid also for the full-span model.

7 APPLICATION EXAMPLE

The reference model to test the described methodology was derived from the model of the new Czech twin wing-mounted engine commuter aircraft for 19 passengers with the maximal take-off weight of 7000 kg.

For the purpose of the first test analyses, the simplified model with four degrees-of-freedom, which are represented by both the symmetric and antisymmetric engine pitch and yaw vibrations, was used. Stiffness characteristics of residual structure, which were modeled using beam elements, were replaced by rigid connections and control surface and tab actuation drives were blocked.

Aerodynamic model included only the wing, nacelles and tip tanks. This simplification was made with regard to the fact, that the aerodynamics of the fuselage and tail surfaces has the negligible effect to the whirl flutter phenomenon. Wing was modeled as Doublet-Lattice Panels and nacelles and tip tanks were modeled as Slender and Interference Bodies. The aerodynamic model included also correction factors for the propeller slipstream applied to the appropriate aerodynamic elements of the wing and nacelles. Furthermore, there was also a correction in the aerodynamic forces and moments at the nose part of the control surfaces. The aerodynamic model is shown in figure 7.

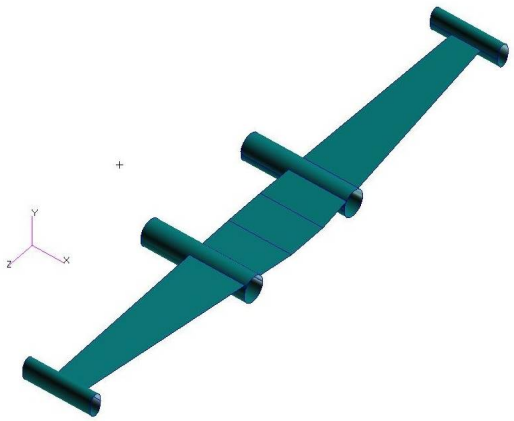


Figure 7. Aerodynamic full-span model of reference aircraft - reduced model (wing, nacelles and tip tanks), only interference tubes are depicted for bodies.

Flight parameters for test analyses were chosen according the aircraft flight envelope. Certification speed was V_{cert} = 191.4 m.s^{-1}, air density ρ = 0.7963 kg.m^{-3} (altitude H = 4267 m) and reference Mach number M = 0.493. Structural damping was neglected. First set of analyses, which is presented here, include the symmetric propeller revolutions, i.e. same revolutions for both left and right propeller (Ω = 2080 rpm). Tested directions of propeller revolutions include CW-CW and CW-CCW combinations.

For the case of identical directions of propeller revolutions (CW-CW), two mechanisms for the whirl flutter appear: 1) a combination of symmetric pitch and antisymmetric yaw modes (SΘ/AΨ) and 2) a combination of antisymmetric pitch and symmetric yaw modes (AΘ/SΨ). The stability margins were calculated with respect to both of the mechanisms of whirl flutter and both are presented in figure 8. As is apparent from the figure, the required engine pitch and yaw frequency is higher for the former mechanism of whirl flutter. Therefore, this mechanism of flutter is more critical compare to the latter one. Values of VFR and HFR were considered at the three levels (1.00; 1.05; 1.10). As the influence of VFR and HFR on the stability margin is negligible, the influence of remaining modes on the stability is negligible as well.

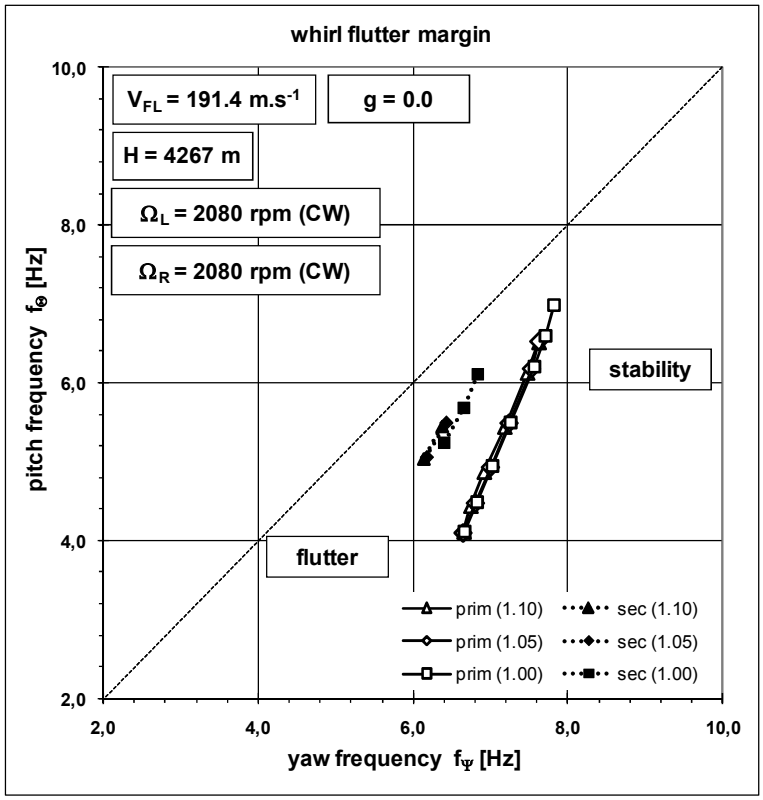


Figure 8. Whirl flutter stability margins - identical directions of revolutions (CW-CW), legend: prim = primary flutter (SΘ/AΨ), sec = secondary flutter (AΘ/SΨ), (1.xx) = VFR, HFR

Figure 9 shows the example of a V-g-f (velocity - damping - frequency) diagram calculated by the standard approach. The crossing of the V-g curve from the negative to the positive damping values represents a flutter state. The mode nr.1 (SΘ mode) crossing represents the primary flutter mechanism (SΘ/AΨ). The flutter speed is here equal to the certification speed. The mode nr.2 (AΘ mode) crossing represents the secondary flutter mechanism (AΘ/SΨ). The flutter speed is here above the certification speed, i.e. above the stability margin, as this type of instability is less critical.

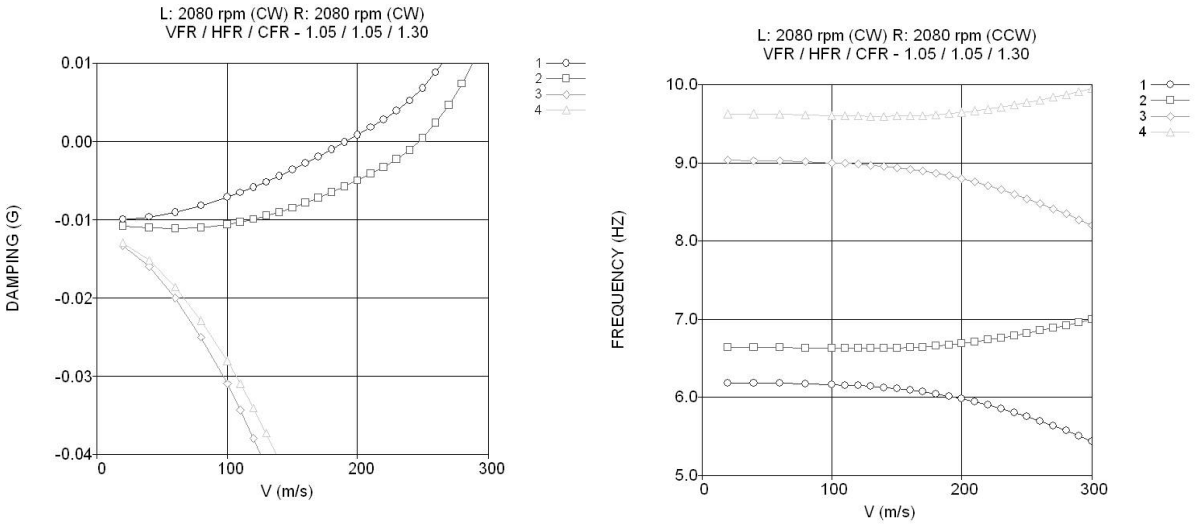


Figure 9. Example of V-g-f diagram - identical directions of revolutions (CW-CW)

For the case of inverse directions of propeller revolutions (CW-CCW), the character of the whirl flutter is different. The instability is caused by the combination of antisymmetric pitch and antisymmetric yaw modes (AΘ/AΨ). Stability margins are presented in figure 10. Compared to the CW-CW case, the required engine pitch and yaw frequencies are considerably higher. Furthermore, this type of instability is influenced also by the remaining modes (SΘ and SΨ) as the influence of VFR and HFR on the stability margin is remarkable.

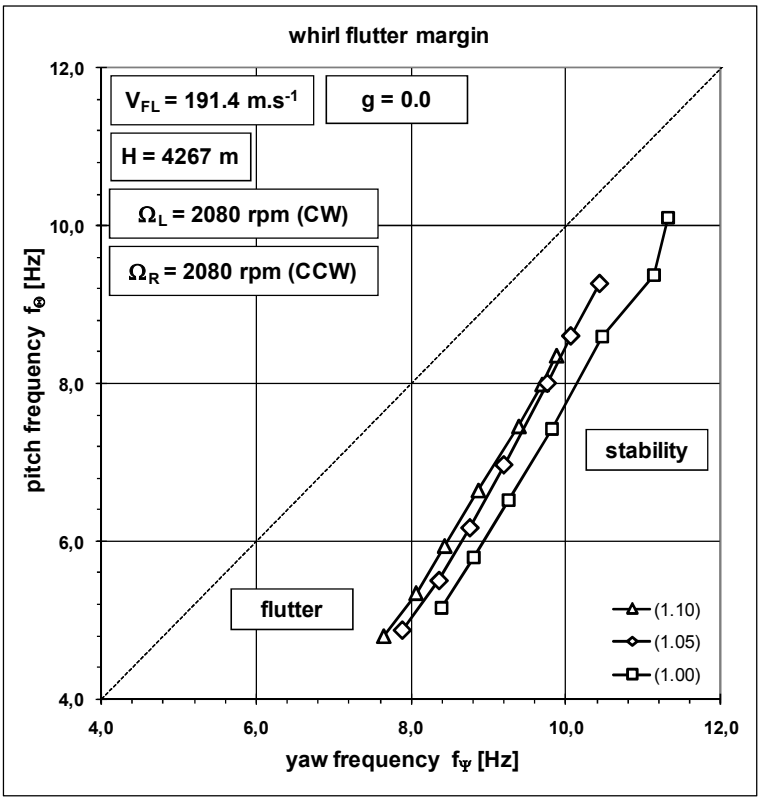


Figure 10. Whirl flutter stability margins - inverse directions of revolutions (CW-CCW), legend: (1.xx) = VFR, HFR

Similarly to the previous case, figure 11 shows the example of a V-g-f diagram for the inverse directions of propeller revolutions. The mode nr.2 (AΘ mode) represent the flutter mechanism (AΘ/AΨ). The flutter speed is here equal to the certification speed.

8 CONCLUSION

This paper presents the optimization-based approach to whirl flutter analysis and the application of the method to the full-span model of the aircraft structure. The necessary adjustments of a stick computational model to make it applicable as a full-span and also the modification of the optimization solution are described. The methodology is demonstrated on the reference model of a twin-engine turboprop commuter aircraft. First test analyses including the symmetric propeller revolutions with both available choices of both propeller rotation directions, i.e. CW-CW and CW-CCW are presented. The results include whirl flutter stability margins for several flutter mechanisms. The most critical one is the case of inverse directions of propeller revolutions (CW-CCW), for which, the critical flutter modes are engine antisymmetric pitch and antisymmetric yaw. The future work will be focused on additional choices of propeller rotations, including unsymmetrical revolutions representing the failure cases of a single propeller overspeed, and the absence of a single propeller rotation representing a failure case of a single propeller feathering.

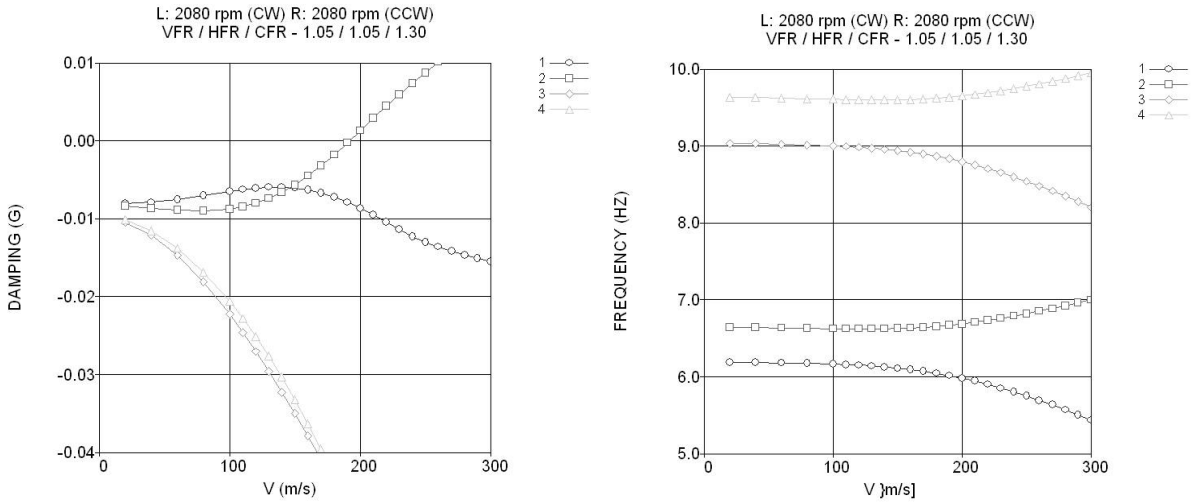


Figure 11. Example of V-g-f diagram - identical directions of revolutions (CW-CCW)

REFERENCES

[1] *Certification Specifications for Normal, Utility, Aerobatic, and Commuter Category Aeroplanes, CS-23, Amendment 3*, European Aviation Safety Agency, 20.7.2012

[2] *Federal Aviation Regulations, Part 23: Airworthiness Standards: Normal, Utility, Acrobatic and Commuter Category Airplanes, FAR-23, Amendment 62*, US. Department of Transportation, Federal Aviation Administration, 31.1.2012

[3] H.S. Ribner, *Propellers in Yaw*, NACA Report 820, 1945

[4] H.S. Ribner, *Formulas for Propellers in Yaw and charts of the Side – Force Derivatives*, NACA Report 819, 1945

[5] J.C. Houbolt, W.H. Reed, Propeller – Nacelle Whirl Flutter, *Journal of the Aerospace Sciences*, 29, 333 – 346, 1962

[6] H.W. Forsching, *Grundlagen der Aeroelastik*, Springer - Verlag, 1974, ISBN 0387065407, 9780387065403

[7] E.S. Taylor, K.A. Browne, Vibration Isolation of Aircraft Power Plants, *Journal of Aerospace Sciences*, 6 (2), 43-49, 1938

[8] W.H. Reed, R.M. Bennett, Propeller Whirl Flutter Considerations for \V/STOL Aircraft, *CAL/TRECOM Symposium*, Buffalo, NY, USA, 1963

[9] F. Kiessling, Some Problems in Research on Whirl Flutter in V/STOL Aircraft,. *European Space Research Organisation*, ESRO-TT-60, May 1975

[10] J. Cecrdle, *Whirl Flutter of Turboprop Aircraft Structures*, Elsevier Science, 2015, ISBN 978-1-782421-85-6, eISBN 978-1-782421-86-3

[11] T. Theodorsen, *General Theory of Aerodynamic Instability and the Mechanism of Flutter*, NACA Report 496, 1935

[12] J.P. Giessing, T.P. Kalman, W.P. Rodden, Subsonic Steady and Oscillatory Aerodynamics for Multiple Interfering Wings and Bodies, *Journal of Aircraft*, 9, 693-702, 1972

[13] W.P. Rodden, E.D. Bellinger, Aerodynamic Lag Functions, Divergence, and the British Flutter Method, *Journal of Aircraft*, 19, 596-598, 1982

[14] G.J. Moore, *MSC/NASTRAN Design Sensitivity and Optimization*, The MacNeal Schwendler Corporation, Los Angeles, CA, USA, May 1994

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON THE CONSIDERATION OF UNCERTAINTY IN DESIGN: OPTIMIZATION – RELIABILITY – ROBUSTNESS

N. Lelièvre[1,3], P. Beaurepaire[2,3], C. Mattrand[2,3], N. Gayton[2,3], A. Otsmane[4]

[1] Université Clermont Auvergne, Institut Pascal
BP 10448, F-63000 CLERMONT-FERRAND
e-mail: nicolas.lelievre@sigma-clermont.fr

[2] Université Clermont Auvergne, SIGMA Clermont, Institut Pascal
BP 10448, F-63171 AUBIERE
e-mail: {pierre.beaurepaire, cecile.mattrand, nicolas.gayton }@sigma-clermont.fr

[3] CNRS, UMR 6602, Institut Pascal
F-63171 AUBIERE

[4] SNECMA Site Villaroche
Rond-Point René Ravaud 77550 MOISSY-CRAMAYEL CEDEX
e-mail: abdelkader.otsmane@snecma.fr

Keywords: Optimization, Design, Robustness, Reliability

Abstract. *Structural designs aspire to ensure the functionality of a system respecting specific requirements defined by engineers. Structural optimization techniques are widely used to optimize the system performances while garanteeing that specific requirements are fulfilled. However, it is worth remembering that uncertainties might affect all design quantities which can make the design problem much more arduous to solve. This paper aims to discuss the formulation of design problems under uncertainty. Several strategies might be identified when considering uncertainties in the objective and/or constraint functions. Moreover, this paper aspires to clarify the notions of robustness and reliability. After a presentation of each formulation, an academic example illustrates the different strategies and shows the main differences between results of each one. Thus, the choice of the formulation should be viewed as a crucial step of design procedure.*

1 Introduction

During the design of a structure, engineers can aim to two main strategies. On one hand, structures have to be able to resist the environmental loads. The structure is thus characterized by an analytical or numerical model which can be used to identify the admissible designs (see e.g. [1]). Designer can adapt the structure parameters to prevent failures and identify an appropriate design. The second strategy is the optimization of the structure which consists of maximize (resp. minimize) the performance (resp. cost) of the structure. The two strategies can obviously be coupled to constrained optimization which provides the best admissible design.

Parameters from structural design problems are in practice affected by uncertainties which may be caused by a lack of knowledge of the production condition for instance or by intrinsic uncertainties. To take into account these uncertainties in design, engineers often use simplifying hypothesis such as safety factors or considering only a reference value. Nevertheless, the variability of the uncertain parameters cannot be considered completely by these methods which can be perform accounting explicitly for uncertainties with probabilities or imprecise probabilities.

In the literature, many papers on design under uncertainty are relating to numerical methods to solve these problem (see e.g. [2]) whereas the objective of the present paper is the clarification of problem formulations. Indeed, stochastic formulations of constraint and objective functions are described. Two main notions are explained: robustness which is habitually associated with the consideration of uncertainties in objective functions and reliability which is based on the introduction in the constraint functions. The proposed classification presented in Table 1 is based on an overview of the literature on the topic (see e.g. [3, 4]).

The paper is structured as follows: section 2 discusses the general concept of uncertainties and also proposes our classification of design formulations. Based on this classification, section 3 presents three deterministic design formulations. Section 4 introduces five formulations for design under uncertainty. An academic example then compares the results obtained from each formulation in section 5. Finally, some conclusions are drawn in section 6.

2 Design under uncertainty

2.1 Sources of uncertainty and their classification

Engineers face uncertainties during the design process, which might affect any parameter. Two catergories are commonly identify in the structural engineering community, namely *aleatory or random* uncertainties and *epistemic* uncertainties, see e.g. [5, 6]. On one hand, aleatory or random uncertainties refer to those that cannot be reduced by introducing additional data or improving the modeling process. They therefore should be viewed as inherent or intrinsic to the considered phenomenon and, as Beyer and Sendhoff [5] pointed out, "the designer has to "live with them" and optimize his design according to this reality". On the other hand, epistemic uncertainties are only due to a lack of knowledge of the studied phenomenon or system behavior, and could therefore be reduced if some conceivable efforts are undertaken.

In a design framework, another classification for uncertainties might be proposed, as suggested by several authors [5, 7]. From their standpoint, uncertainties can be distinguished depending on whether they can be acted upon or not in the design phase:

- *Type I uncertainties* are primitively linked to the environment and conditions of use. The variables that show this type of uncertainty are hereafter noted in the series $P_j(\omega), j = 1, ..., m$ and stored in vector $\mathrm{P}(\omega)$. They do not play a role in the design procedure, i.e.

they are independent from it, and as such they are not design variables.

- *System function uncertainties* are those linked to the evaluation of the performance (or output) of the system. In this paper, these uncertainties are also gathered in vector $\mathrm{P}(\omega)$.

- *Feasibility uncertainties* are associated with uncertainties on the constraint functionsand have hence an influence on the definition of the design space. These uncertainties are also considered as model uncertainties and grouped in vector $\mathrm{P}(\omega)$.

- *Type II uncertainties* are those connected with the production/ manufacturing process. Geometrical variables noted $X_i(\omega), i = 1, ..., n$ (and stored in vector $\mathrm{X}(\omega)$) are usually linked to this type of uncertainty. They are part of the design variables which might affect the performance of the system.

2.2 Consideration of uncertainties in engineering

In the design process, uncertainties can be considered by two approaches. The first approach is the conversion of each uncertain parameter into a reference value. The stochastic problem thus becomes a deterministic problem. The so-called worst-case method permits to identify this reference value. Indeed, in this method, uncertain parameters are assumed to be bounded, and the bound of each variable associated with the worst scenario is identified. In tolerance analysis, this method is commonly used (see e.g. [8]) because each dimensions are uncertain and the tolerance intervals provide suitable bounds. However, this strategy is not applicable to general problems and the identification of the bounds may be a complex task. The second method to convert a random variable into a deterministic parameter consists of weighting each uncertain parameter by a safety factor commonly, which depends on the importance of the variable. With this method, unbounded variable can also be treated.

The second approach is the explicit consideration of uncertainties which are characterized by a mathematical representation. The variation of uncertain parameters can thus be taken entirely into account and it allows the propagation uncertainties to the response of the structure. Several approaches may be used to realize this characterization. The probabilistic approach defines each uncertain variable by a distribution function (see e.g. [9]) and it is commonly used in mechanical design (see e.g. [10]). An alternative approach is imprecise probabilities. The possibilistic approach defines uncertain parameters by their possibility distributions, as detailed e.g. by Zadeh [**?**]. Another method is using fuzzy sets to characterize uncertainties (see e.g. [**?**]).

2.3 Design requirements

In design processes, the state of mechanical system's performances can be characterized by two different functions. First, the *objective-type functions* $f(\mathrm{X}(\omega), \mathrm{P}(\omega))$ aim to quantify the performances of the mechanical system. Hence, optimization consist of maximizing the performances of the system (e.g. quality level), minimizing the cost or targeting a reference value. Other design-related functions are the *constraint-type functions* $g(\mathrm{X}(\omega), \mathrm{P}(\omega))$ which must be satisfied in all operating conditions to ensure the functionality of the system. These constraint functions must be achieved as closely as possible in order not to degrade the objective functions. The constraint functions are used to define the admissible space which is defined as $\{\mathrm{X} \in \mathbb{R}^n \mid g(\mathrm{X}(\omega), \mathrm{P}(\omega)) \geq 0\}$.

2.4 Proposed classification of design formulations

During design process, uncertainties can be accouted by several strategies, which involves multiple design problem formulations. Concepts such as reliability and robustness need to be introduced. A classification is proposed by Gang [11] who brings forward five design problem formulations. Our classification is introduced in Table 1, where eight design formulations are considered making the distinction by taking into account uncertainties in the objective and/or the constraint functions. Robustness is defined by the faculty of a system's response to be insensitive to small variations in system parameters, and is thus associated with the objective function. When uncertainties are taken into consideration in the constraint functions, non-admissible solutions are tolerated as long as they remain rare. Reliability is therefore associated with the constraint functions; it characterizes the ability of a system to ensure its functions in a given context.

Three possible states can be defined for each output function: no function is used, a deterministic function is employed or a function subject to uncertainties is considered. Nine combinations are thus introduced according to the state of objective and constraint functions in Table 1 and are detailed in the next sections.

	Robustness ——————————>			
Reliability ↓		No function f (objective)	Function f X, P deterministic	Function f X, P uncertain
	No function g (constraint)		Optimal design (optimization without constraint) (Section 3.2)	Robust design (Section 4.3)
	Function g X, P deterministic	Admissible design (sizing) (Section 3.1)	Optimal and admissible design (optimization under constraint) (Section 3.3)	Robust and admissible design (Section 4.4)
	Function g X, P uncertain	Reliable design (reliability) (Section 4.1)	Optimal and reliable design (RBDO) (Section 4.2)	Robust and reliable design (RBRDO) (Section 4.5)

Table 1: Different design approaches. Grey boxes take into account some uncertainties

3 Deterministic design formulation

3.1 Admissible design - Sizing

Sizing is the most often used scheme in design because of it ease of implementation. Indeed, the admissible design corresponds to the set of solutions that respect problem constraints and thus provides the solution space of design variables. This design formulation (Equation (1)) is

completely deterministic.

$$\text{Find } \bar{\mathrm{X}}_{Adm} \text{ such that : } g(\bar{\mathrm{X}}_{Adm}, \mathrm{P}^{(k)}) \geq 0 \tag{1}$$

However, dealing with uncertainties in an implicit manner can be realized by transforming the initial stochastic problem into a deterministic problem using the worst-case approach or the safety factors

3.2 Optimal design - Optimization without constraint

The principle of optimal design is to optimize an objective function which is not subject to constraints. This problem formulated in Equation (2) is entirely deterministic.

$$\text{Find } \bar{\mathrm{X}}_{Opt} \text{ such that : } \bar{\mathrm{X}}_{Opt} = \underset{\bar{\mathrm{X}}}{\mathrm{Arg\,max}} \ f(\bar{\mathrm{X}}, \mathrm{P}^{(k)}) \tag{2}$$

Industrial problems cannot be treated with this design formulation since this is an unconstrained optimization and can thus lead to non-admissible solutions (see section 3.1). However, this design formulation can be applied for mathematical applications such as for instance linear and non-linear regressions solved by the least-square method.

3.3 Optimal and admissible design - Optimization under constraint

The optimal and admissible design in the combination of the two previous formulations since using only one of the two formulations is generally insufficient for designers. Hence, the obtained design is both optimal with respect to the objective function and admissible with respect to the constraint functions (Equation (3)).

$$\begin{aligned} &\text{Find } \bar{\mathrm{X}}_{OptAdm} \text{ such that :} \\ &\bar{\mathrm{X}}_{OptAdm} = \underset{\bar{\mathrm{X}}}{\mathrm{Arg\,max}} \ f(\bar{\mathrm{X}}, \mathrm{P}^{(k)}) \\ &\text{Subject to (s.t.) } g(\bar{\mathrm{X}}, \mathrm{P}^{(k)}) \geq 0 \end{aligned} \tag{3}$$

This is a classic problem of optimization under constraint without considering uncertainties, which is also called deterministic constrained optimization.

4 Formulation of design in an uncertain context

4.1 Reliable design

Reliable design formulation can be viewed as an extension of the admissible design problem introduced in section 3.1. In this approach, uncertainties are explicitly taken into account. They may be modeled by the probabilistic approach, as briefly recalled in section 2.2, thereby leading to an assessment of the failure probability of a system or structure and guaranteeing that it is below a given threshold value P_{target}. $\bar{\mathrm{X}}$ denotes therefore the mean of the probability law of X.

$$\begin{aligned} &\text{Find } \bar{\mathrm{X}}_{Rel} \text{ such that :} \\ &\mathrm{Prob}\left(g(\mathrm{X}(\bar{\mathrm{X}}_{Rel}, \omega), \mathrm{P}(\omega)) \leq 0\right) \leq P_{target} \end{aligned} \tag{4}$$

The failure probability threshold value may be fixed using a risk-based approach (e.g. [12]) for which consequences of failure are evaluated. In the case of major failure events P_{target} needs to be very low, whereas its value can be higher if failure events have minor consequences.

4.2 Optimal and reliable design - Reliability-Based Design Optimization

Optimal and reliable design formulation consists of optimizing a deterministic objective function subject to stochastic constraints. This design, commonly named reliability-based design optimization (RBDO), is formulated in Equation (5) as suggested by [13, 14, 15]. In this paper, RBDO is always associated with deterministic objective and probabilistic constraints.

$$\begin{aligned} &\text{Find } \bar{\mathrm{X}}_{OptRel} \text{ such that :} \\ &\bar{\mathrm{X}}_{OptRel} = \underset{\bar{\mathrm{X}}}{\mathrm{Arg\,max}} \; f(\bar{\mathrm{X}}, \mathrm{P}^{(k)}) \\ &\text{s.t. : } \mathrm{Prob}\left(g(\mathrm{X}(\bar{\mathrm{X}}, \omega), \mathrm{P}(\omega)) \leq 0\right) \leq P_{target} \end{aligned} \tag{5}$$

Reliability-based design optimization has been widely discussed in the literature and alternative formulations might be found where for instance uncertainties are considered in both objective and constraint functions [16, 17].

4.3 Robust design

Such as the optimal design, robust design formulation consists of optimizing an objective function without taken any constraints into account (Equation (6)). The main difference is that uncertainties are taken into account in the objective function (Ψ) and which is thus named robust objective function.

$$\begin{aligned} &\text{Find } \bar{\mathrm{X}}_{Rob} \text{ such that :} \\ &\bar{\mathrm{X}}_{Rob} = \underset{\bar{\mathrm{X}}}{\mathrm{Arg\,max}} \; \Psi(\mathrm{X}(\bar{\mathrm{X}}, \omega), \mathrm{P}(\omega)) \end{aligned} \tag{6}$$

The choice of formulation for the robust function is essential, since it can significantly modify the results of the optimization. Many formulations can be used for the robust function, from basic ones to more elaborate formulas such as linear combination ([18]), Taguchi's Mean Square Deviation ([5, 19]) or the Mean Square Error ([20]). This formulation is generally not relevant for industrial problem because of the lack of constraints.

4.4 Robust and admissible design

Robust and admissible design consists of optimizing the robust objective function (Ψ) subject to deterministic constraint functions (Equation (7)).

$$\begin{aligned} &\text{Find } \bar{\mathrm{X}}_{RobAdm} \text{ such that :} \\ &\bar{\mathrm{X}}_{RobAdm} = \underset{\bar{\mathrm{X}}}{\mathrm{Arg\,max}} \; \Psi(\mathrm{X}(\bar{\mathrm{X}}, \omega), \mathrm{P}(\omega)) \\ &\text{s.t. : } g(\bar{\mathrm{X}}, \mathrm{P}^{(k)}) \geq 0 \end{aligned} \tag{7}$$

In this formulation, uncertainties are considered only in the objective functions, the constraint functions remain deterministic. The obtained optimum performance is also insensitive with respect to moderate variations of the variables.

4.5 Robust and reliable design

The last formulation is the robust and reliable design where uncertainties are explicitly taken into account in both objective and constraint functions (Equation (8)). Solutions are thus both

robust and reliable. This formulation is sometimes called reliability-based robust design optimization (RBRDO) ([20, 21]).

$$\begin{aligned} &\text{Find } \bar{\mathrm{X}}_{RobRel} \text{ such that :} \\ &\bar{\mathrm{X}}_{RobRel} = \underset{\bar{\mathrm{X}}}{\mathrm{Arg\,max}}\ \Psi(\mathrm{X}(\bar{\mathrm{X}}, \omega), \mathrm{P}(\omega)) \\ &\text{s.t. : } \mathrm{Prob}\left(g(\mathrm{X}(\bar{\mathrm{X}}, \omega), \mathrm{P}(\omega)) \leq 0\right) \leq P_{target} \end{aligned} \tag{8}$$

This formulation might be viewed as the most complete formulation which provides the surest solutions.

5 Application example

This section aims to illustrate each design formulation on an academic example. The application used to highlight the effect of design formulation on the solution is the study of a cylindrical container. This container is defined by a radius R and a height h and must be able to contain a minimum of $33\ \mathrm{cm}^3$ whilst using a minimum of material. The container is manufactured using metal sheet with the same thickness for each side. Hence, the constraint function is defined by Equation (9) and the objective function, which represents the area of material, is defined by Equation (10).

$$g(R, h) = V(R, h) - 33 = \pi R^2 h - 33 \tag{9}$$

$$f(R, h) = 2\pi Rh + 2\pi R^2 \tag{10}$$

The constraint function must be positive to ensure the sufficient volume of the container. The objective function must be minimized to obtain the design minimizing the quantity of material. The search of each design is bounded to $1 \leq R \leq 4$ and $1 \leq h \leq 10$ (cm) and results are given in Table 2.

Deterministic design

The non-admissible domain is defined by the constraint function (Equation (9)) and is represented by a hatched area in Figure 1. Without any constraint, the minimal quantity of material is zero, the optimal design is thus the zero solution.To best respect the constraint, the constraint function must be as close to the equality as possible. The optimal and admissible solution is thus at the limit between the admissible and non-admissible domains (see continuous curve Figure 1).

Design in an uncertain context

The variables R and h are modeled by Gaussian distributions with the same standard deviation of $0.5\ \mathrm{mm}$, independently of their nominal value. The failure probability if set at $P_{target} = 0.1$. The reliable design domain is bounded by this limit and the non-reliable domain is represented by the hatched area in Figure 2.

For robust optimization, a new objective function must be defined considering uncertainties. Here, a linear combination (Equation (11)) of the expectation ($E(\circ)$) and the standard deviation ($\sigma(\circ)$) of the initial objective function is chosen.

$$\Psi(R, h) = 0.25 \times E(f(R, h)) + 0.75 \times \sigma(f(R, h)) \tag{11}$$

Since no constraints are taken into account, the robust design solution is the zero solution (as it happens for optimal design).

As discussed previously, to best respect constraints, the equality mus be targeted by the constraint function. The obtained RBDO solution, thus, appears at the limit between the reliable and non-reliable domains (see dotted curve in Figure 1) whereas the robust and admissible design (resp. robust and reliable design) is at the limit between the admissible and non-admissible domains (resp. reliable and non-reliable design). The main difference between each solution is that the robust solutions are more insensitive to variations in the problem parameters.

This application shows the main differences between the formulations, which therefore should be considered as an essential component of the design process in order to obtain the desired design.

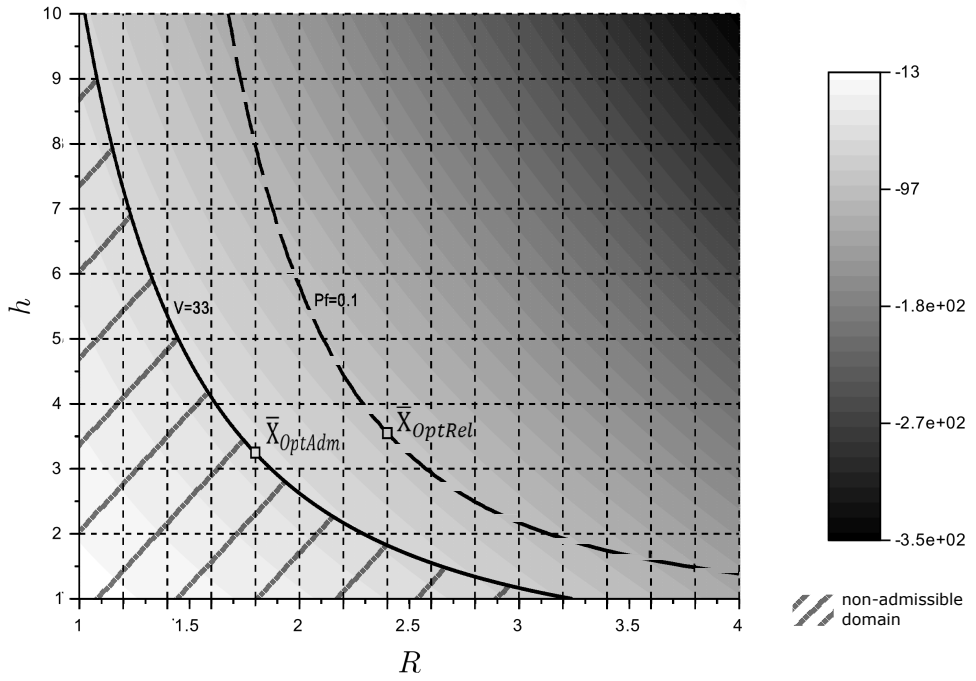


Figure 1: Example results – objective function ($-f(R,h)$ to be maximized) in grayscale

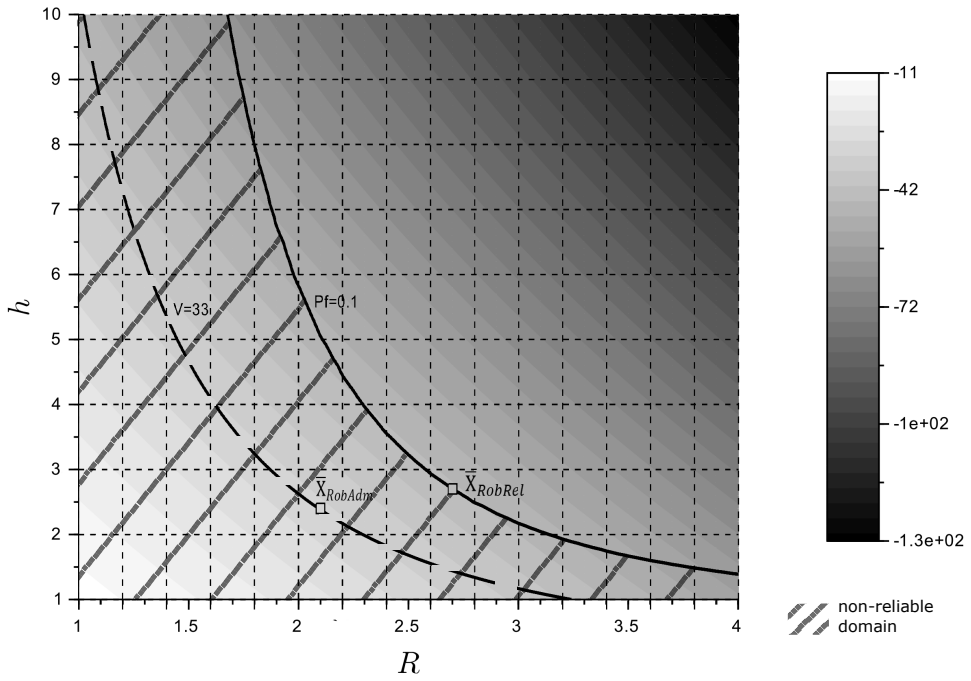


Figure 2: Example results – objective function ($-\Psi(R,h)$ to be maximized) in grayscale

	Robustness →			
Reliability ↓		No function f (objective)	Function f X, P deterministic	Function f X, P uncertain
	No function g (constraint)		Optimal design $\bar{\mathrm{X}}_{Opt} = [0;0]$	Robust design $\bar{\mathrm{X}}_{Rob} = [0;0]$
	Function g X, P deterministic	Admissible design see Figure 1	Optimal and admissible design $\bar{\mathrm{X}}_{OptAdm} = [1.8;3.25]$	Robust and admissible design $\bar{\mathrm{X}}_{RobAdm} = [2.1;2.4]$
	Function g X, P uncertain	Reliable design see Figure 1	Optimal and reliable design (RBDO) $\bar{\mathrm{X}}_{OptRel} = [2.4;3.55]$	Robust and reliable design (RBRDO) $\bar{\mathrm{X}}_{RobRel} = [2.4;2.7]$

Table 2: Summary of results for the container application.

6 Conclusion

In this paper, a comprehensive summary of design formulations is discussed by proposing a classification of design problems (Table 1). The different definitions and formulations are clarified. Designing aspires to identify admissible solutions subject to constraints, and optimizing

aims to maximize performance metrics. Robustness is linked to the objective whereas reliability is associated with the constraints. The application shows that significant differences exist between each formulations. The choice of design formulation is thus a crucial step in the design procedure, and requires careful consideration.

acknowledgements

The first author was funded by a ministerial grant from Blaise Pascal University. The contribution of SNECMA Villaroche is also gratefully acknowledged.

REFERENCES

[1] H. Al Nageim, F. Durka, W. Morgan, D.T. Williams, Structural mechanics: loads, analysis, materials and design of structural elements, *7th edition*, Pearson Education, London, 2010.

[2] M.A. Valdebenito, G.I. Schuëller, A survey on approaches for reliability-based optimization, *Structural and Multidisciplinary Optimization*, 42(5), 1-19, 2010a.

[3] I. Doltsinis, Z. Kan, Robust design of structures using optimization methods, *Computer Methods in Applied Mechanics and Engineering*, 193(23-26), 2221-2237, 2004.

[4] H.A. Jensen, M.A. Valdebenito, G.I. Schuëller, Kusanovic D.S. , Reliability-based optimization of stochastic systems using line search, *Computer Methods in Applied Mechanics and Engineering*, 198(49–52),3915-3924, 2009.

[5] H.G. Beyer, B. Sendhoff, Robust optimization - A comprehensive survey, *Computer Methods in Applied Mechanics and Engineering*, 196, 3190-3218, 2007.

[6] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic ? Does it matter ?, *Structural Safety*, 31, 105-112, 2009.

[7] W. Chen, J.K. Allen, K.L. Tsui, F. Mistree, A Procedure for Robust Design: Minimizing Variations Caused by Noise Factors and Control Factors, *ASME Journal of Mechanical Design*, 118, n° 4, 478-485, 1996.

[8] W.H. Greenwood, K.W. Chase, A new tolerance analysis method for designers and manufacturers, *Transaction of the ASME*, 109, 112-116, 1987.

[9] A.S. Nowak, K.R. Collins, *Reliability of Structures*, McGraw-Hill Higher Education, 2000.

[10] Z. Meng, G. Li, B.P. Wang, P. Hao, A hybrid chaos control approach of the performance measure functions for reliability-based design optimization, *Computers and Structures*, 146, 32-43, 2015.

[11] W. Gang, S. Wang, F. Xiao, D. Gao, Robust optimal design of building cooling systems considering cooling load uncertainty and equipment reliability, *Applied Energy*, 159, 265-275, 2015.

[12] R. Rackwitz, Optimization and risk acceptability based on the Life Quality Index, *Structural Safety*, 24, 397-331, 2002.

[13] Z. Meng, P. Hao, G. Li, B. Wang, K. Zhang, Non-probabilistic reliability-based design optimization of stiffened shells under buckling constraint, *Thin-Walled Structures*, 94, 325-333, 2015.

[14] Z.P. Mourelatos, J. Liang, A Methodology for Trading-Off Performance and Robustness Under Uncertainty, *Journal of Mechanical Design*, 128, 856-863, 2006.

[15] E.H. Vanmarcke, Matrix formulation of reliability analysis and reliability-based design, *Computers and Structures*, 3, 757-770, 1973.

[16] I. Enevoldsen, J.D. Sørensen, Reliability-based optimization in structural engineering, *Structural Safety*, 15, 169-196, 1994.

[17] G.I. Schuëller, H.A. Jensen, Computational methods in optimization considering uncertainties -– An overview, *Computer methods in applied mechanics and engineering*, 198, 2-13, 2008.

[18] M. Papadrakakis, N.D. Lagaros, V. Plevris, Design optimization of steel structures considering uncertainties, E*ngineering Structure*, 27, no. 9, 1408-1418, 2005.

[19] M.W. Trosset, Taguchi and Robust Optimization, 1997.

[20] A.F. Shahraki, R. Noorossama, Reliability-based robust design optimization: A general methodology using genetic algorithm, *Computers and Industrial Engineering*, 74, 199-207, 2014.

[21] I. Lee, K.K. Choi, L. Du, D. Gorsich, Dimension reduction method for reliability-based robust design optimization, *Computers and Structures*, 86, 1550-1562, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A NEW OPTIMIZED ANISOTROPIC MOVING LEAST SQUARES SURROGATE MODEL WITH MAXIMIZED PROGNOSIS

Kevin Cremanns[1] **& Dirk Roos**[2]

Institute of Modelling and
High-Performance Computing
Niederrhein University of Applied Sciences, Germany
[1] Kevin.Cremanns@hs-niederrhein.de
[2] Dirk.Roos@hs-niederrhein.de

Keywords: Anisotropic Moving Least Squares, Weighted Least Squares, Local Regression, Surrogate Model, Metamodel, Maximized Prognosis Quality

Abstract. *Computational simulations or experiments in engineering applications are often time and resource consuming. This is especially so when performing optimization, robustness or reliability analyses, because they need hundreds or even thousands of design evaluations. A common solution for this are surrogate models, which are mathematical approximations of the existent data. Nevertheless, these surrogate models also need some design evaluations as a base to train them and are usually obtained by performing a design of experiments with an intelligent design plan to reduce the number of needed points.*

Moving least squares is a very simple but powerful surrogate model method. It performs a least squares regression with a subset of the data in the local area of the point to be approximated. This allows the method to approximate complex functions with a low order polynomial base (first or second order). This basically works with the help of a weighting function, which gives a high weight to points near the target point and decreasing weights to points with an increasing distance to this point. Therefore, in the standard moving least squares method, a free model parameter has to be estimated. This parameter describes the influence radius, in which points are considered to approximate the target point. In this standard formulation, the weights depend on the distances of all variables. So, variables with a minor or even non-influence, contribute to these weights and so to the accuracy of the approximation.

The proposed method works anisotropically and is similar to methods like the anisotropic Kriging, so that the free model parameters need to be defined for each variable individually and the possibility of suppressing unimportant variables and thus increasing the prognosis quality of the standard moving least squares method is given.

This work presents the mathematical description of this method and demonstrates its performance for different analytical test functions compared to the standard moving least squares method.

NOMENCLATURE

AMLS	anisotropic moving least squares
CFD	computational fluid dynamics
DOE	design of experiment
FSI	fluid structure interaction
LHS	latin hypercube sampling
MLS	moving least squares
CV	cross validation
α	constant for weighting function
μ	mean value
σ	variance
x_n	n-th variable of prediction point $\mathbf{x}$
n	number of input variables
N	number of samples
$\widetilde{y}$	approximated output
R^2_{pred}	prediction quality based on CV
w_n	Gaussian weighting function
r	radius of points used for approximation
w_i	weight for input point i
$\mathbf{x}$	prediction point
$\mathbf{v}_n$	training values for n-th variable
$\mathbf{p}(\mathbf{x})$	vector of basis polynomials
$\mathbf{a}(\mathbf{X}, \mathbf{x})$	vector of approximation coefficients
$\mathbf{y}(\mathbf{X})$	vector of output training values
$\mathbf{w}_{N,n}$	weighting vectors for the anisotropic formulation
$\widetilde{\mathbf{y}}$	vector of approximated output values
$\mathbf{X}$	matrix of training points
$\mathbf{P}(\mathbf{X})$	matrix of basis polynomials $N \times n$
$\mathbf{W}(\mathbf{X}, \mathbf{x})$	weighting matrix $N \times N$
$\mathbf{X}_{\text{test}}$	points used to calculate R^2_{pred} not used for training

SUBSCRIPTS

n	n-th variable
m	$(n-1)$-th variable
N	N-th design point
test	test point not used for training
iso	isotropic
aniso	anisotropic

1 INTRODUCTION

In engineering applications, a lot of time and resource consuming simulations (e.g. CFD / FSI - calculations) or experiments take place. Therefore, in most cases, optimization or robustness evaluations are too expensive in terms of time and resources. One outcome of this is the usage of surrogate models. These are functions based on given training data points to approximate new unknown function evaluations within the initial training data range.

Especially in the field of aerodynamic optimization, surrogate models are used, because of the high calculation time. In [26] used for example a simulated annealing optimization algorithm to optimize the shape of a three-finned, stepped labyrinth seal. This optimization took about 830 design evaluations with a calculation time of 25 minutes per design. Hence, the whole optimization took more than 14 days. Therefore, others like [25] and [22] showed optimizations of turbo machinery parts by using surrogate models trained on a specific number of CFD-calculations. Also in other fields, surrogate models are used for optimization such as aircraft design ([11]), engine cycle design ([3]) or structural reliability analysis ([18]). A general overview of surrogate based optimization is also given in [9].

A wide range of different methods to create such surrogate models exist. Some of the most sophisticated are radial basis functions, Kriging (also known as the Gaussian process), support vector regression, neural networks and moving least squares. Because of this variety of methods, a lot of benchmarks for these different methods also exist, as described in [6], [10], [15] and [4]. In most of these comparisons Kriging ([14]) was identified as one of the most accurate methods and it has been shown that this method works well for a whole range of problems. However, in some cases other methods, such as moving least squares (MLS), ([16], [17]) can be more sophisticated then Kriging, as [1] shows for non-convex functions. Especially in arbitrarily spaced data ([7]) or in 3D surface reconstruction presented in [27], [2] and [8], MLS showed good performance.

MLS regression has the advantage compared to methods like Kriging, that it does not experience difficulties when there are a large number of design points to train on, because of numerical instabilities. Also, if the design points show different local behaviour, such as clusters, MLS performs well. The main obstacle for local regression methods like MLS is to determine the points to consider for the local approximation, especially if the points have a large variance in a defined window around the prediction point.

Another point is the dimensionality of the problem to be approximated. A lot of surrogate methods experience problems with high dimensional data. In the standard isotropic MLS all dimensions have an influence on the variance of the data points used for the approximation. Therefore, with increasing dimensionality the approximation accuracy reduces. Two possible methods exist for working with high dimensional problems. The first method is to reduce the variables to only the most important parameters through the use of a sensitivity analysis such as the variance based sensitivity analysis from [12]. An example of the successful utilization of this type of procedure is presented in [21] or in [5]. In many engineering applications, the number of most important parameters for an individual output is up to 20 input variables. Therefore, these methods work very well for finding these few important parameters and thus dramatically increase the performance of the surrogate model used. The other possible method of working with high dimensional data is anisotropy. In a similar way to anisotropic Kriging, the surrogate method provides the opportunity to consider the different importance influence of the input variables in the prediction. In Kriging, different correlation lengths will be defined for each input parameter, leading to higher accuracy of the Kriging surrogate model, especially

if the variance is different for each input parameter. [24] showed this for the MLS method. Also, an anisotropy was introduced through the definition of different bandwidths for each input parameter. Then, only the points which fall in the intersection of the different bandwidths which define the prediction window, are used for the prediction. As will be shown in section 4, different bandwidths for each input parameter will be defined, but a different method of calculating the overall weights for each prediction point will be used.

In this paper, the concept of anisotropy is used and applied to the moving least squares method different to [24], which in most cases leads to higher prognosis quality. First, the standard MLS is described, followed by the extension to anisotropy. Subsequently, examples of different analytical test functions are shown to compare both methods.

2 DESIGN SAMPLING

To create the training data for the surrogate models, some simulation or experiments have to be evaluated as base information for different designs. These designs should be part of a sampling plan which covers the range of interest for the specific problem. There are many possible methods for creating such a design plan, such as full factorial, D-optimal or orthogonal arrays. Latin hypercube sampling is commonly used (LHS) ([20]), which creates *n* hypercubes for each variable. Within each of these hypercubes, a point is chosen randomly, or in the case of mid-point LHS the point is always set in the center of this hypercube. Optimized LHS also exist, with optimized distance between the points and minimal unwanted correlations as [13] shows. For the test problems shown here, an optimized LHS will be used.

3 REVIEW OF THE MLS APPROXIMATION METHOD

The basis for the anisotropic moving least squares (AMLS) approximation is the standard MLS. The formulation to approximate the value $\widetilde{y}$ of the prediction point $\mathbf{x}$ is:

$$\widetilde{y}(\mathbf{x}) = \mathbf{p}^T(\mathbf{x})\mathbf{a}(\mathbf{X}, \mathbf{x}) \tag{1}$$

The prediction point $\mathbf{x}$ consists of the *n* input variables of the problem to approximate (e.g. a simulation model, mathematical function or experiment):

$$\mathbf{x} = [x_1, x_2, ..., x_n] \tag{2}$$

The vector $\mathbf{p}(\mathbf{x})^T$ contains the basis polynomial for the approximation depending on $\mathbf{x}$. For example a second order polynomial with cross terms:

$$\mathbf{p}^T(\mathbf{x}) = [1, x_1, x_2, \ldots, x_n, x_1^2, x_2^2, \ldots, x_n^2, x_1x_2, \ldots, x_nx_m]^T \tag{3}$$

The vector $\mathbf{a}(\mathbf{X}, \mathbf{x})$ contains the approximation coefficients and is calculated through:

$$\mathbf{a}(\mathbf{X}, \mathbf{x}) = (\mathbf{P}^T(\mathbf{X})\mathbf{W}(\mathbf{X}, \mathbf{x})\mathbf{P}(\mathbf{X}))^{-1}\mathbf{P}^T(\mathbf{X})\mathbf{W}(\mathbf{X}, \mathbf{x})\mathbf{y}(\mathbf{X}) \tag{4}$$

with the matrix $\mathbf{P}^T(\mathbf{X})$ containing the basis polynomials for the $N \times n$ training points matrix $\mathbf{X}$ of *n* samples:

$$\mathbf{X} = \begin{Bmatrix} x_{1,1} & \cdots & x_{n,1} \\ \vdots & \ddots & \vdots \\ x_{1,N} & \cdots & x_{n,N} \end{Bmatrix} \tag{5}$$

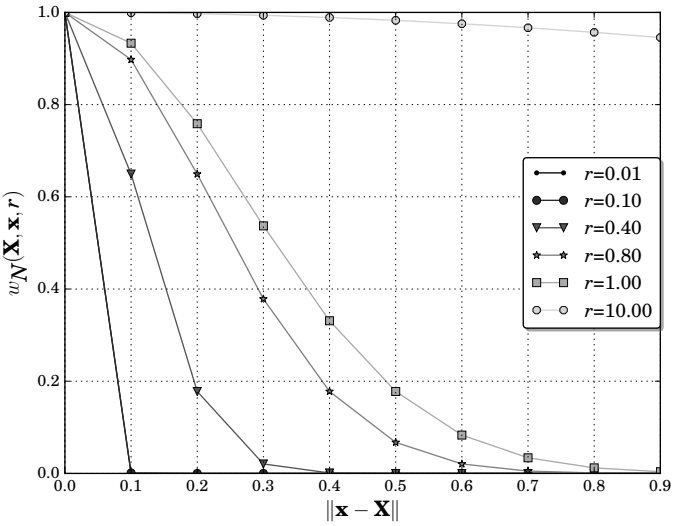


Figure 1: Gaussian weighting function depending on *r*

with $\mathbf{P}^T(\mathbf{X})$:

$$\mathbf{P}^T(\mathbf{X}) = \begin{Bmatrix} 1 & x_{1,1} & \dots & x_{n,1} & x_{1,1}^2 & \dots & x_{n,1}^2 & x_{1,1}x_{2,1} & \dots & x_{n,1}x_{n+1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,N} & \dots & x_{n,N} & x_{1,N}^2 & \dots & x_{n,N}^2 & x_{1,N}x_{2,N} & \dots & x_{n,N}x_{n+1,N} \end{Bmatrix} \tag{6}$$

The matrix $\mathbf{W}(\mathbf{X}, \mathbf{x})$ is the $N \times N$ weighting diagonal matrix for the prediction point $\mathbf{x}$ and the main mathematical difference to normal polynomial regression. $\mathbf{W}(\mathbf{X}, \mathbf{x})$ contains the weights for each element of $\mathbf{X}$ regarding their distance to $\mathbf{x}$:

$$\mathbf{W}(\mathbf{X}, \mathbf{x}) = \begin{Bmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_N \end{Bmatrix} \tag{7}$$

To estimate these weights, different types of weighting functions exist. Some of them can even force the MLS to interpolate through the training points ([19]). In this investigation a Gaussian weighting function is used like proposed in [17]:

$$w_N(\mathbf{X}, \mathbf{x}, r) = e^{-\left(\frac{\sum_1^N \|\mathbf{x} - \mathbf{X}\|}{r\alpha}\right)^2} \tag{8}$$

with the constant:

$$\alpha = \frac{1}{\sqrt{-log(0.001)}} \tag{9}$$

In this formulation the $\|\mathbf{x} - \mathbf{X}\|$ stands for the Euclidean distance calculated over all variable dimensions between the prediction point $\mathbf{x}$ and the trainings points $\mathbf{X}$. There is one free model parameter *r*, which represent the influence radius around the prediction point $\mathbf{x}$, in which the training points $\mathbf{X}$ are considered to approximate $\mathbf{x}$. As shown in Fig. 1 the weights tend to 0 if the training points lie outside of *r*. The last part to estimate $\mathbf{a}(\mathbf{X}, \mathbf{x})$ is the output vector $\mathbf{y}(\mathbf{X})$,

which contains the output values of the training points $\mathbf{X}$.

$$\mathbf{y}(\mathbf{X}) = [y_1, \ldots, y_N] \tag{10}$$

To estimate the free model parameter r different kinds of possibilities exist. A straightforward solution is to use an optimizer in combination with a stratified k-fold cross validation (CV). k-fold CV involves splitting the training points in k equal sized packages. Then the model is trained with $k-1$ packages and the remaining package is used as prediction points (test points). After each package has been evaluated the corresponding prognosis quality can be estimated with:

$$R^2_{\text{pred}} = \left(\frac{\sum_1^{N_{\text{test}}} (\mathbf{y}(\mathbf{X}_{\text{test}}) - \mu_{\mathbf{y}_{\text{test}}})(\widetilde{\mathbf{y}}(\mathbf{X}_{\text{test}}) - \mu_{\widetilde{\mathbf{y}}_{\text{test}}})}{(N_{\text{test}} - 1)\sigma_{\mathbf{y}_{\text{test}}}\sigma_{\widetilde{\mathbf{y}}_{\text{test}}}} \right)^2 \tag{11}$$

The objective of the optimization would be to maximize the prognosis quality $R^2_{\text{pred}} \in [0, 1]$.

4 EXTENSION TO ANISOTROPY

Having explained the standard MLS method, this section describes the extension of the method to the anisotropy. The key point of the MLS method is the weighting of training points, so that each prediction point is locally approximated. In the standard formulation, the Euclidean distance is calculated over all variable dimensions, as shown in 8. To give each variable an individual influence on the weights for the training points used, the anisotropic version calculates the weights for each variable separately. This leads to n weighting vectors $\mathbf{w}_{N,n}(\mathbf{v}_n, x_n, r_n)$ (the diagonal of the weighting matrix $\mathbf{W}(\mathbf{X}, \mathbf{x})$) with n elements and vector $\mathbf{v}_n$ describing the training data of the n-th variable:

$$w_{N,n}(\mathbf{v}_n, x_n, r_n) = e^{-\left(\frac{\sum_1^N \|\mathbf{v}_n - x_n\|}{r_n \alpha} \right)^2} \tag{12}$$

To combine these weighting vectors to one global and so to a corresponding weighting diagonal matrix $\mathbf{W}(\mathbf{X}, \mathbf{x})$, the geometric mean over all n weighting vectors $\mathbf{w}_{N,n}(\mathbf{v}_n, x_n, r_n)$ is calculated:

$$\mathbf{W}(\mathbf{X}, \mathbf{x}) = \text{diag}\left(\sqrt[n]{\prod_{i=1}^{n} \mathbf{w}_{N,n}(\mathbf{v}_n, x_n, r_n)} \right) \tag{13}$$

Compared to the standard MLS, one r_n per variable exists now. Because the geometric mean is used to combine the weighting vectors $\mathbf{w}_{N,n}(\mathbf{v}_n, x_n, r_n)$, variables with a minor or no influence on the output $\mathbf{y}(\mathbf{X})$, their r_n will tend to:

$$r_n \to \infty \tag{14}$$

because $r_n \to \infty$ gives $\mathbf{w}_{N,n}(\mathbf{v}_n, x_n, \infty) = \mathbf{1}$, meaning that all training points are considered with a weight of 1 regarding the non-important variables (see Fig. 1 with $r = 10$). Whereas variable with high influence will tend to:

$$r_n \to 0 \tag{15}$$

and so the weights are $\neq 0$ for training points in the local area of the prediction point $\mathbf{x}$. This means that only the r_n of the important variables have a significant influence on the final weighting matrix $\mathbf{W}(\mathbf{X}, \mathbf{x})$. Even though with this anisotropic approach of the MLS method it is not

necessary to scale the input parameters to equal ranges as is necessary for the isotropic MLS, but it is recommended that this scaling is done for numerical stability.

To estimate the free model parameters r_n the Powell optimization algorithm is used ([23]). This algorithm is especially suitable for this kind of optimization problem, because it iterates in a double loop over each optimization variable separately. The inner loop iterates over each free parameter and changes it individually until no further improvement is archived. If each free parameter has been optimized once, a new iteration begins (the outer loop). Furthermore, this algorithm does not need any gradients, just a start vector.

Additional as mentioned in 1, it is suggested that a sensitivity analysis should be used in applications with very high dimensions, to further increase the accuracy of the model and to decrease the optimization effort. So overall, it is recommended that this anisotropic MLS method is used in combination with a variance based sensitivity analysis ([12]) to receive the most accurate surrogate model.

5 EXAMPLES

In the following section, examples of comparing the standard MLS method to the anisotropic version are shown. For the comparison, the same Powell optimization algorithm is used for the isotropic MLS as for the anisotropic MLS. The optimization criterion is the already introduced R^2_{pred} (11) evaluated with k-fold CV. To train the surrogate model, 50 uniformly distributed samples were created, with optimized latin hypercube sampling as mentioned in 2. For these examples, no scaling was used, because all parameters are in the same value range. Further, each test function was approximated with a first order and second order polynomial base and the best result was taken. The k-fold CV was done with $k = 5$, with the same packages structure (same designs in each package) for both methods. For the comparisons 4 test functions assist in demonstrating how the AMLS works and how the free parameters r_n behave. They are defined as:

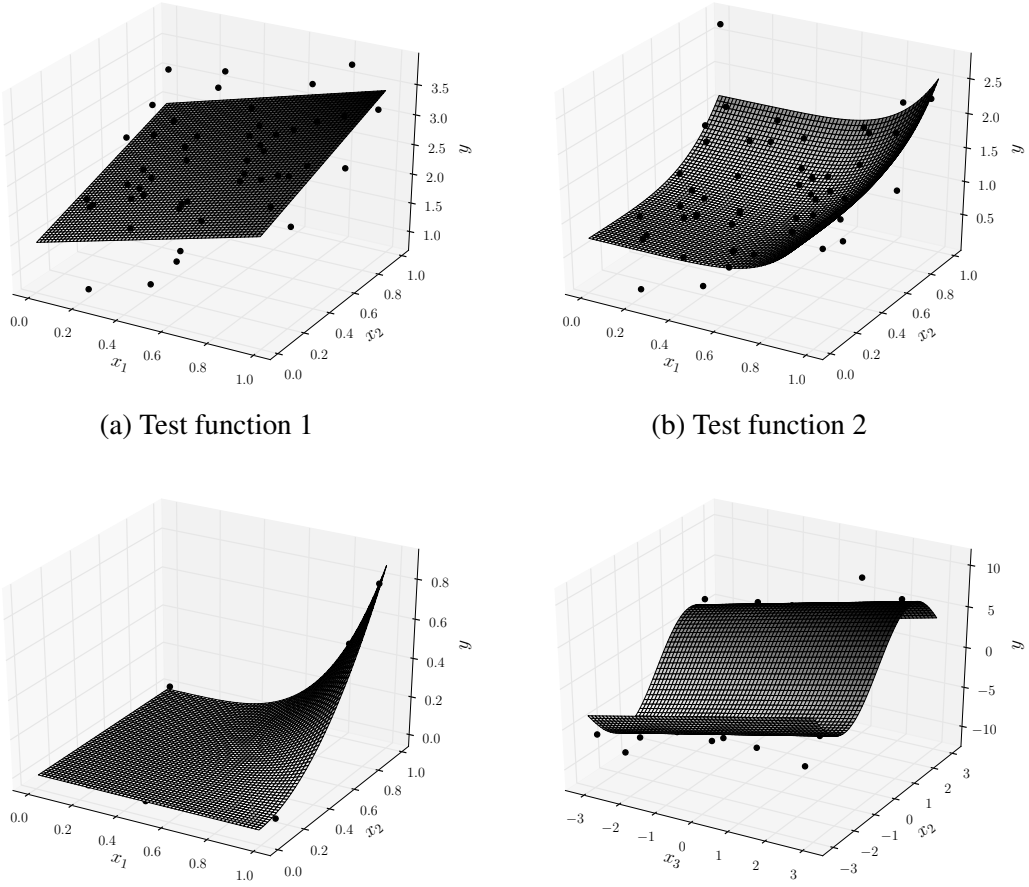
$$y(\mathbf{x}) = x_1 + x_2 + x_3 + x_4 + x_5, x_{1-5} \in [0, 1] \tag{16}$$

$$y(\mathbf{x}) = x_1^6 + x_2^4 + x_3^2 + x_4 + 0.001x_5, x_{1-5} \in [0, 1] \tag{17}$$

$$y(\mathbf{x}) = x_1^4 x_2^2 + 0.01x_3x_4 + 0.001x_5, x_{1-5} \in [0, 1] \tag{18}$$

$$y(\mathbf{x}) = 0.5x_1 + x_2 + 0.5x_1x_2 + 5sin(x_3) + 0.2x_4 + 0.1x_5, x_{1-5} \in [-\pi, \pi] \tag{19}$$

In Eq. 16 all variables have the same influence on $y(\mathbf{x})$ and should show equal results for the isotropic MLS and AMLS. Furthermore, all r_n of the AMLS should be equal because all variables have equal influence on the output $y(\mathbf{x})$. Eq. 17 includes decreasing influence from x_1 the highest and x_4 the lowest. Here it is intended, that the free parameter r_n should also show a decreasing value for the corresponding input parameters. The variance of each variable is different and so the power of the anisotropic version should result in a higher R^2_{pred} value for the AMLS. Eq. 18 contains influence of cross terms with a high variance in one corner of the data range and Eq. 19 has one dominant variable x_3 and 4 'noise' terms. The test functions 16 to 19 are illustrated in Fig. 2a to Fig. 2d. Tab. 1 shows the results for the free parameters of the MLS r_{iso} and for AMLS r_{1-5}. Additional the prognosis quality R^2_{pred} for both methods is given. For the first test function 16 the free model parameter of the AMLS were all equal as expected. The prognosis quality of both methods is maximal with $R^2_{\text{pred}} = 1$ because the test



(a) Test function 1

(b) Test function 2

(c) Test function 3

(d) Test function 4

Figure 2: Overview test functions

function is an additive simple example. The values of the free model parameters r_n of the AMLS method is higher than parameter the range of $[0, 1]$, which means all points are considered in the weighting and there is no local regression effect, in other words it works like a normal least squares approximation. The same applies to the MLS method for this specific example with parameter ranges of $[0, 1]$ with a maximum of $\sqrt{\sum_1^5 1^2} = \sqrt{5}$. The results for the second test function show the first difference between the MLS and the AMLS method regarding the prognosis quality R^2_{pred} of 5%. The free model parameters r_n of the AMLS are very low for the most important parameters x_1, x_2, x_3 and are very high for the unimportant parameters x_4, x_5 so that their weighting does not influence the overall weighting in the geometric mean described in 13. The values of the important parameters suggest that the approximation works locally. Additionally, there is a ranking of the parameters x_1, x_2, x_3 corresponding to the real ranking in the test function. The third test function 18 shows the highest difference of R^2_{pred} with 72% for all test functions. Here the AMLS seems to use all parameters, but with a focus on the first two parameters x_1, x_2 as important parameters for the overall weighting, corresponding to the test function part with $x_1^4 x_2^2$ which clearly dominate the values of the test function. The MLS is not capable to approximate this function with one single free parameter r. The last test function 19 yields a difference for R^2_{pred} of 34%. The dominant parameter x_3 has been identified in the AMLS, which results in the lowest free model parameter of $r_3 = 0.99$.

r_{iso}	r_1	r_2	r_3	r_4	r_5	$R^2_{\text{pred}_{\text{iso}}}$	$R^2_{\text{pred}_{\text{aniso}}}$
7.06	3.07	3.07	3.07	3.07	3.07	1.0	1.0
1.73	0.33	0.67	0.72	5494.53	2314.16	0.92	0.97
1.97	0.07	0.14	0.88	0.35	0.44	0.09	0.81
7.56	4762.98	5.17	0.99	3.71	2.85	0.54	0.88

Table 1: Overview of free model parameters r_n for test functions 16 to 19

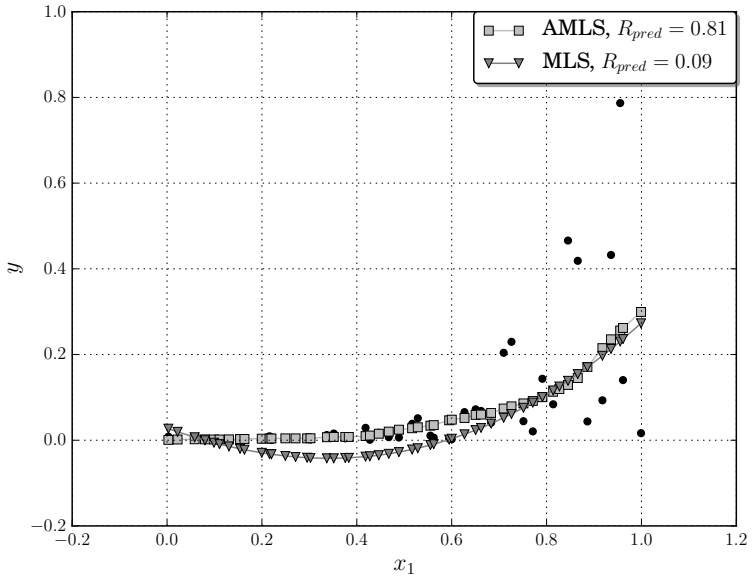


Figure 3: Surrogate model for test function 3, 1D view for parameter x_1

In Fig. 3 and 4 the 1D view of the surrogate models for the most important parameters of test function 18 and 19 is presented. For both pictures, it can be seen, that the trend is captured much better by the AMLS corresponding to the higher R^2_{pred}, especially for x_3 of test function 19, where it captured the $sin()$ term compared to the MLS, which seems to average through the data points.

6 CONCLUSION

The anisotropic moving least squares (AMLS) surrogate model method presented here uses, in contrast to the standard isotropic moving least squares (MLS), a free model parameter r_n for each input parameter and further the geometric mean to calculate the overall weighting (see 4). The AMLS method is more accurate than the MLS method for the examples with input parameters of differing strong influence as demonstrated with the test functions Eq. 17 to Eq. 19. Large differences could be seen in the example with high multivariate dependencies.

This is accomplished by setting the free model parameter r_n corresponding to the influence of the input parameter. Input parameters with a high influence result in lower values for r_n and vise versa. To find the right value for r_n, the Powel optimization algorithm was used. A possible start solution vector for this optimizer can be obtained by previously performing a sensitivity analysis. Otherwise, it is recommended that the mean value of the input parameter bounds for each r_n are used as a starting point. Additionally, if the input parameters are of different

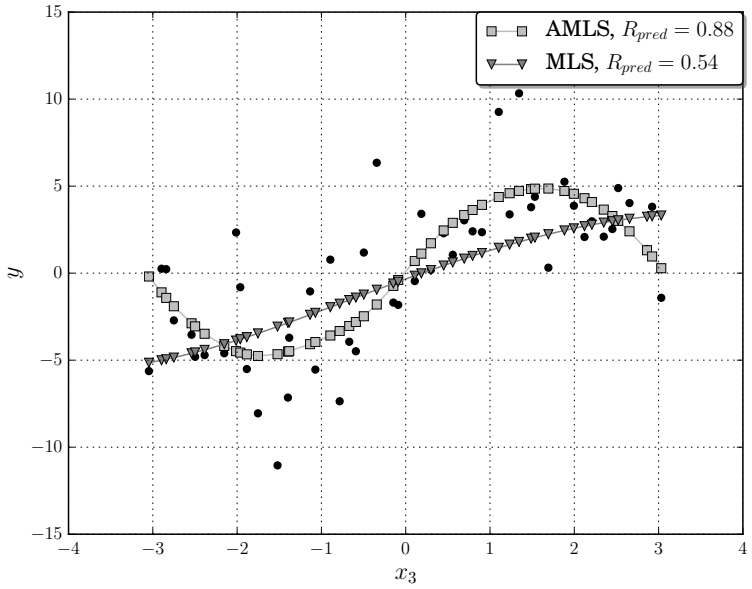


Figure 4: Surrogate model for test function 4, 1D view for parameter x_3

magnitudes it is recommended that they are rescaled.

If no sensitivity analysis has been performed, the AMLS optimization can also work as a simple kind of sensitivity analysis. Unimportant parameters can be identified through high values for the r_n parameters (similar to high correlation lengths in the anisotropic Kriging method). For all other parameters, the optimization also delivers a ranking of importance with the smallest r_n value correspond to the most important parameter etc.

Further improvements might be archived by testing different weighting functions. For example an interpolating weighting function or a function which avoids the high numerical numbers in the optimization for unimportant parameters by normalizing the maximum r_n with the maximum 1D distance for each specific input variable.

ACKNOWLEDGMENT

The authors would like to express their thanks to the German Federal Ministry for Economic Affairs and Energy, which supported this project, support identification no.: 03ET7071D.

References

[1] K. Beyong-Soo, L. Yong-Bin, and C. Dong-Hoon. Comparison study on the accuracy of metmetamodel technique for non-convex functions. *Journal of Mechanical Science and Technology*, 23:1175–1181, 2009.

[2] W. Y. Chang. Surface Reconstruction from Points. Technical report, Department of Computer Science and Engineering University of California, San Diego, 2009.

[3] J. A. Corman and B. J. German. A Comparison of Metamodeling Techniques for Engine Cycle Design Data. In *13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference*, 2010.

[4] J.-P. Costa, L. Pronzato, and E. Thierry. A comparison between Kriging and radial basis function network for nonlinear prediction. In *Conference: Proceedings of the IEEE-*

EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), Antalya, Turkey, June 20-23, 1999.

[5] K. Cremanns, D. Roos, and A. Grassmann. Sequential vs. Multidisciplinary Coupled Optimization and Efficient Surrogate Modelling of a Last Stage and the Successive Axial-radial Diffusor in a Low Pressure Steam Turbine. In *ASME Turbo Expo 2014: Turbomachinery Technical Conference and Exposition*, 2014.

[6] H. Fang, M. Rais-Rohani, Z. Liu, and M. F. Horstmeyer. A comparative study of metmetamodel methods for multiobjective crashworthiness optimiztaion. *Computers and Structures*, 83:2121–2136, 2005.

[7] R. Farwig. Multivariate interpolation of arbitrarily spaced data by moving least squares methods. *Journal of Computational and Applied Mathematics*, 16:79–93, 1986.

[8] S. Fleishman, D. Cohen-Or, and C. T. Silva. Robust moving least-squares fitting with sharp features. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2005*, 24:544–552, 2005.

[9] A. I. J. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45:50–79, 2009.

[10] A. A. Giuanta and L. T. Watson. A comparison of approximation modeling techniques: polynomial versus interpolating models. Technical report, Multidisciplinary Analysis and Design (MAD) Center for Advanced Vehicles, Virginia Polytechnic Institute and State University, 1998.

[11] Z.-H. Han and K.-S. Zhang. Surrogate-Based Optimization. Technical report, School of Aeronautics, Northwestern Polytechnical University, Xian, P.R. China, 2012.

[12] T. Homma and A. Saltelli. Importance Measures in Global Sensitivity Analysis of Nonlinear Models. *Reliability Engineering and System Safety*, 52(1):1 – 17, 1996.

[13] R. V. Joseph and Y. Hung. Orthogonal-Maxmin Latin Hypercube Designs. *Statistica Sinica*, 18:171–186, 2008.

[14] G. D. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chem., Metal. and Mining Socity of South Africa*, 52:119–139, 1951.

[15] T. Krishnamurty. Comparison of Response Surface Construction Methods for Derivative Estimation Using Moving Least Squares, Kriging and Radial Basis Functions. In *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference*, 2005.

[16] P. Lancaster and K. Salkauskas. Surfaces Generated by Moving Least Squares Methods. *Mathematics of Computation*, 37:141–158, 1981.

[17] D. Levin. The approximation power of moving least squares. *Mathematics of Computation*, 67:1517–1531, 1998.

[18] J. Li, H. Wang, and N. H. Kim. Doubly weighted moving least squares and its application to structural reliability analysis. *Structural and Multidisciplinary Optimization*, 46:69–82, 2012.

[19] G. G. Maisuradze and D. L. Thompson. Interpolating Moving Least Squares Methods for Fitting Potential Energy Surfaces. *The Journal of pyhsical chemistry A*, 37:7118–7124, 2003.

[20] M. McKay, R. Beckman, and W. Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 21:239–245, 1979.

[21] T. Most and J. Will. Metamodel of optimal prognosis - an automatic approach for variable reduction and optimal metamodel selection. In *Proceedings of the 5th Weimar Optimization and Stochastic Days*, Weimar, Germany, November 20-21, 2008. DYNARDO GmbH.

[22] D. I. Papadimitriou, I. C. Kampolis, and K. C. Giannakoglou. Stochastic and deterministic optimization for turbomachinery applications based on the adjoint formulation. In *Design Optimization International Conference*, 2004.

[23] M. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, 7:155–162, 1964.

[24] P. Salemi, B. L. Nelson, and J. Staum. Moving Least Squares Regression for High Dimensional Simulation Metamodelling. In *Proceedings of 2012 Winter Simulation Conference*, 2012.

[25] A. Samad and K. Kwang-Yong. Surrogate based optimiztaion techniques for aerodynamic design of turbomachinery. *International Journal of Fluid Machinery and Systems*, 2:179–188, 2009.

[26] V. Schramm, J. Denecke, S. Kim, and S. Wittig. Shape Optimization of a Labyrinth Seal Applying the Simulated Annealing Method. *International Journal of Rotating Machinery*, 10:365–371, 2004.

[27] A. Soon-Jeong, Y. Jaechil, L. Byung-Gook, and L. Joon-Jae. 3D surface reconstruction from scattered data using moving least square method. In *Image Analysis and Processing (ICIAP)*, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

LATIN HYPERCUBE SAMPLING BASED ON ADAPTIVE ORTHOGONAL DECOMPOSITION

Dirk Roos

Institute of Modelling and High-Performance Computing
Niederrhein University of Applied Sciences, Germany
dirk.roos@hs-niederrhein.de

Keywords: Experimental and Simulation Design, Latin Hypercube Sampling, Proper Orthogonal Decomposition, Singular Value Decomposition.

Abstract. *In the field of real world or simulation experiments, latin hypercube designs are very useful to approximate implicit given functions based on experimental designs of rectangular grid of points used e.g. for an efficient optimization of black-box experiments. The so called standard latin hypercube sampling is based on an idea of [8] to simulate correlated multivariate relationships based on uncorrelated variables. Hereby, the linear or rank correlations between the random variables can be reduced using a Cholesky decomposition of the covariance matrix. However, this algorithm requires that the number of samples is higher than the number of input variables to ensure a positive definite and non-singular covariance matrix. Furthermore, a rearrangement of the latin hypercubes so that they will have the same ordering to increase the multidimensional uniformity introduces an unwanted pseudo linear correlation of the simulated variables.*

In this paper, the new method of proper orthogonal decomposition methods in adaptive combination with re-sorting the rank-order of the original marginal distributions not only eliminates the drawbacks of the standard latin hypercube sampling, but also reduces the number of experimental designs necessary. In contrast to the standard latin hypercube sampling, the adaptive approach results in a further improvement of the independency and the multidimensional uniformity of the variables and shows an improved efficiency in comparison to optimization-based methods. Furthermore, a lower-rank approximation of the design matrix based on singular value decomposition ensures a positive definite and non-singular correlation matrix to reduce the correlation error in a stepwise manner during the adaptive procedure.

1 INTRODUCTION

Physical or computer experiments are widely used for the design and optimization of products and processes. In simulation experiments, rather than carrying out physical experiments, numerical models describing the performance of the product or the process are only implicitly given through numerical methods. When modelling complicated physical phenomena we are searching for a mathematical model

$$y = f(x_1, ..., x_i, ..., x_n) = f(\mathbf{x}), \quad \mathbf{x} = (x_1, ..., x_i, ..., x_n)^T \in \mathbb{D}^n \subset \mathbb{R}^n$$

where the vector $\mathbf{x}$ consists of the n design (input) variables or factors and y defines a specific output variable (response). The function $f(\mathbf{x})$ is defined within the design space $\mathbb{D}$ with the lower and upper bounds x_i^l, x_i^u of the variables.

In general, efficient design exploration techniques, such as sensitivity analysis and optimization are based on surrogate (or meta) models $\tilde{y}(x_i)$ of response values using evaluation points or the so-called designs. A design of experiments with N points and n variables is usually written as a $N \times n$ design matrix

$$\mathbf{X} = \left[\mathbf{x}^1, ..., \mathbf{x}^k, ..., \mathbf{x}^N\right]^T \in \mathbb{D}^{N,n} \subset \mathbb{R}^{N,n}$$

where each row

$$\mathbf{x}^k = \left[x_1^k, ..., x_i^k, ..., x_n^k\right]$$

represents a sample k and each column represents a variable i. Obviously, in a real world design of experiment, it is not possible to enlarge the necessary number of experimental designs n to obtain acceptable space-filling and independency of the simulated design space which is important for an accurate sensitivity analysis and for design space exploration. So one of the main aims in the field of experiments is therefore to obtain efficient designs of experiments.

A commonly used stratified sampling technique is the latin hypercube sampling, especially if the number of design variables is greater than $n = 5$ and general nonlinearities of the model responses are possible. Numerous modifications have been proposed for latin hypercube sampling to minimize correlation error and to ensure the unidimensional uniformity. The pseudo linear correlations can be minimized by an internal optimization procedure using a column pairwise switch algorithm to increase uniformity or space-fillingness of a sampling plan. Several methods are proposed in the literature, e.g.[7, 13] use a simulated annealing and in [9] an enhanced stochastic evolutionary is presented. Commonly used optimization criteria are the maximization of the minimum distance (maximin metric) according to [13] introduced by [10, 5] and mean of the square of the pairwise correlations, proposed by [8]. A Matlab implementation of an optimization algorithm to construct these maximin latin hypercubes is given in [5].

Other methods are developed to create small orthogonal or nearly-orthogonal latin hypercube matrices [12], which are perfectly uniform in the multivariate space. These methods are applicable only for small sets of variables and design evaluations. In [2] a method is presented for calculating nearly orthogonal latin hypercubes and calculating the best space-filling designs which is computationally expensive. The method presented in [4] does not necessarily calculate the best space-filling designs but is applicable to high dimensional problems, preserving the integrity of the latin hypercube sampling estimator by increasing the statistical distance between designs using sequentially eliminating design evaluations that are near to each other in the multidimensional space. In [14] another method for reducing correlations of a latin hypercube using the Gram-Schmidt orthogonalization is proposed. An extended overview of space-filling designs can be found in [3].

Because of the huge combinatorial effort of the assigned optimization problem, finding an optimized latin hypercube sampling is a very difficult and time-consuming task and restricted to a relative small number of variables. Otherwise, several decomposition methods of the covariance matrix exist to reduce the unwanted pseudo correlation of the variables in a high dimensional space, which can be enhanced to improve unidimensional uniformity and independency of the design points, as shown in the following.

2 STANDARD LATIN HYPERCUBE SAMPLING

2.1 Plain latin hypercube sampling

A plain Monte Carlo simulation can be used to create $i = 1, ..., n$ nearly-uniformly distributed samples $U_i^k(0.5; 1/(2\sqrt{3}))$ with $k = 1, ..., N$ designs with normalization into a $[0, 1]^n$ box. Unfortunately, a plain Monte Carlo simulation does not consider that every class of all variables should be simulated with the same probability and results in non-uniformity of the marginal distributions.

The basic idea of [15] assumes that each sample out of N is unique with high probability. Of course, sorting these samples in ascending order gives random numbers too, and returns a row vector of $\mathbf{u}$ for every variable $i = 1, ..., n$ containing a random permutation of the integers from $k = 1, .., N$. The permutation matrix $\pi_{U_i}^k$ of U_i^k ensures that the bins have their centers in the center of the latin hypercubes with

$$H_i^k = \frac{\pi_{U_i}^k - 0.5}{N}$$

The splitting of the design space $\mathbb{D}^n$ into N classes $\mathbb{D}_k^i$ with the same probability

$$P\left[h_i^k \in \mathbb{D}_k^i\right] = \frac{1}{N}, \qquad i = 1, ..., n, \quad k = 1, ..., N$$

gives N^n hypercubes (bins) with the probability N^{-n}. Hereby, all portions of the random variables' range are represented and therefore we obtain a univariate uniformity of the marginal distribution of the variables. Unfortunately, this procedure leads to an unwanted correlated sample matrix

$$\mathbf{H} = \begin{bmatrix} h_1^1 & h_2^1 & \cdots & h_n^1 \\ h_1^2 & h_2^2 & ... & h_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ h_1^N & h_2^N & \cdots & h_n^N \end{bmatrix}$$

especially for a small number of samples N. In some cases the maximal error of the correlation matrix of a plain latin hypercube sampling

$$\epsilon_{\boldsymbol{\rho}_{\mathbf{HH}}} = \max_{1 \le i,j \le n} \left(|\boldsymbol{\rho}_{\mathbf{HH}} - \mathbf{I}|\right)$$

can be larger than the maximal correlation error $\epsilon_{\boldsymbol{\rho}_{\mathbf{UU}}}$ of a plain Monte Carlo sampling.

2.2 Correlation transformation using factorization of the covariance matrix

According to [8], the unwanted linear or rank correlations between the random variables can be eliminated using a Cholesky decomposition of the covariance matrix

$$\mathbf{C}_{\mathbf{HH}} = E\left[\left(\mathbf{H} - \bar{\mathbf{H}}\right)\left(\mathbf{H} - \bar{\mathbf{H}}\right)^T\right]$$

whereby $E[.]$ denotes the expectation value and $\bar{\mathbf{H}}$ the vector of the means of the latin hypercube samples. Using the standardization $Y_i(0;1) \sim H_i$ by

$$Y_i = \frac{H_i - E[H_i]}{\sigma_{H_i}} \tag{1}$$

the covariance matrix of Y_i is equal to the assigned correlation matrix

$$\boldsymbol{\rho}_{\mathbf{YY}} = \mathbf{C}_{\mathbf{YY}} = E\left[\mathbf{Y}\mathbf{Y}^T\right] \tag{2}$$

In the general case a variable transformation $Z_i(0;1) \sim Y_i$ is to find so that the correlation matrix of all variables Z_i represents the target correlation in terms of the identity matrix, thus

$$\boldsymbol{\rho}_{\mathbf{ZZ}} = E\left[\mathbf{Z}\mathbf{Z}^T\right] \stackrel{!}{=} \mathbf{I} \tag{3}$$

Inserting Eq. (2) and Eq. (3) in

$$\boldsymbol{\rho}_{\mathbf{ZZ}}\, \boldsymbol{\rho}_{\mathbf{ZZ}}^{-1} = \mathbf{I} = \boldsymbol{\rho}_{\mathbf{YY}}\, \boldsymbol{\rho}_{\mathbf{YY}}^{-1} \tag{4}$$

and using the Cholesky factorization of the matrix of the pseudo correlations

$$\boldsymbol{\rho}_{\mathbf{YY}} = \mathbf{L}\mathbf{L}^T \tag{5}$$

we obtain

$$E\left[\mathbf{Z}\mathbf{Z}^T\right] = \mathbf{I}\,\mathbf{I}^T = \left(\mathbf{L}^{-1}\mathbf{L}\right)\left(\mathbf{L}^{-1}\mathbf{L}\right)^T = \mathbf{L}^{-1}\mathbf{L}\,\mathbf{L}^T\mathbf{L}^{-1^T} = \mathbf{L}^{-1}\boldsymbol{\rho}_{\mathbf{YY}}\mathbf{L}^{-1^T}$$

and furthermore

$$E\left[\mathbf{Z}\mathbf{Z}^T\right] = \mathbf{L}^{-1}E\left[\mathbf{Y}\mathbf{Y}^T\right]\mathbf{L}^{-1^T} = E\left[\mathbf{L}^{-1}\mathbf{Y}\;\mathbf{Y}^T\mathbf{L}^{-1^T}\right] = E\left[\underbrace{\left(\mathbf{L}^{-1}\mathbf{Y}\right)}_{\mathbf{Z}}\underbrace{\left(\mathbf{L}^{-1}\mathbf{Y}\right)^T}_{\mathbf{Z}^T}\right]$$

In the case of the non-singular lower triangle matrix $\mathbf{L}$ the transformation results in

$$\mathbf{Z} = \mathbf{L}^{-1}\mathbf{Y} \tag{6}$$

and deals in this way numerically without correlation error. The back substitution of Eq. (1)

$$U_i = Z_i \cdot \sigma_{H_i} + E[H_i]$$

gives normalized, nearly-uniformly distributed variables U_i. Finally, the uniform distributed variables $X_i \sim U_i$ with respect to the vectors of lower and upper bounds $\mathbf{x}^l$ and $\mathbf{x}^u$ are simply given by

$$\mathbf{X} = \mathbf{x}^l + \mathbf{U} \circ \left(\mathbf{x}^u - \mathbf{x}^l\right) = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^N & x_2^N & \cdots & x_n^N \end{bmatrix} \in \mathbb{D}^{N,n} \subset \mathbb{R}^{N,n} \tag{7}$$

whereby, the symbol $\circ$ denotes the entrywise product.

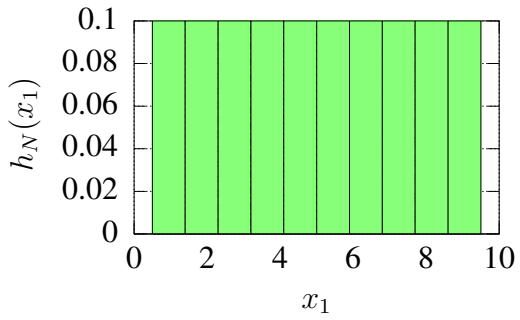


Figure 1: Histogram of the first variable x_1, simulated by Eq. (7).

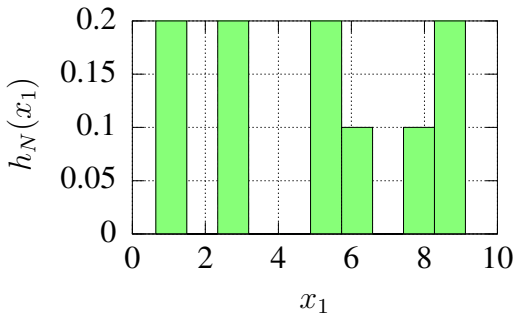


Figure 2: Histogram of the second variable x_2, simulated by Eq. (7).

2.3 Rearrangement of the latin hypercubes

On the other hand, the linear transformation in Eq. (6) moves the sample points coordinates for all variables $i > 1$. And as a result the samples may be moved beyond the desired bounds and marginal distributions may be effected from the change of correlation, as shown in Fig. (1) and Fig. (2). To solve this problem, in [8] a resorting of the design matrix $\mathbf{Z}$ according to the rank correlation of the exact marginal distributions of $\mathbf{H}$

$$H_i^k = \frac{\pi_{Z_i}^k[\pi_{Z_i}^k] - 0.5}{N}$$

is presented. Whereby, the inner permutation matrix describes the rank-order sorting of the N samples of n variables of the design matrix $\mathbf{Z}$ and the outer permutation matrix describes the rank-order sorting of the sorted indices to create modified center-based latin hypercubes $\mathbf{H}$ with accurate marginal distributions to ensure univariate uniformity. Unfortunately, this rearrangement within the so-called standard latin hypercube sampling achieves the multidimensional uniformity but introduces an unwanted pseudo linear correlation of the simulated variables

$$\mathbf{X} = \mathbf{x}^l + \mathbf{H} \circ \left(\mathbf{x}^u - \mathbf{x}^l\right)$$

again. But in summary, the maximal correlation error of the simulated variables

$$\epsilon_{\boldsymbol{\rho}_{\mathbf{XX}}} = \max_{1 \le i,j \le n} \left(|\boldsymbol{\rho}_{\mathbf{XX}} - \mathbf{I}|\right) \tag{8}$$

is significantly smaller than the maximal correlation error $\epsilon_{\boldsymbol{\rho}_{\mathbf{HH}}}$ of the first latin hypercube sampling.

3 LATIN HYPERCUBE SAMPLING BASED ON ADAPTIVE ORTHOGONAL DECOMPOSITION

3.1 Adaptive rearrangement of the latin hypercubes

Particularly, in high-dimensional design spaces and with a reduced number of design evaluations $n \lesssim N$ the correlation error in Eq. (8) increases as a result of the rank-order sorting. Otherwise, the linear transformation in Eq. (6) achieves an exact correlation and maintain the uniformity of the first variable. So, an essential enhancement of the standard latin hypercube sampling is an adaptive rearrangement of the columns of the current given latin hypercube matrix $\mathbf{H}$ in each iteration step, as shown in Fig. (3).

$$[|\boldsymbol{\rho}_{\mathbf{XX}} - \mathbf{I}|] = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & \dots & \dots & j & \dots & n \end{array} \\ \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \epsilon_{\boldsymbol{\rho}_{\mathbf{XX}}} & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \end{pmatrix} \end{array} \quad \mathbf{H} = \begin{bmatrix} h_1^1 & \cdots & h_i^1 & \cdots & h_j^1 & \cdots & h_n^1 \\ h_1^2 & \cdots & h_i^2 & \cdots & h_j^2 & ... & h_n^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ h_1^N & \cdots & h_i^N & \cdots & h_j^N & \cdots & h_n^N \end{bmatrix}$$

Figure 3: Iterative rearrangement of the columns of the latin hypercube matrix $\mathbf{H}$, according to the rows number i and the columns number j of the maximal correlation error and the first column with accurate uniformity.

3.2 Orthogonal transformation using proper orthogonal decomposition

The Cholesky factorization Eq. (5) requires that the number of samples is greater than the number of input variables, thus $N > n$ to ensure a positive definite pseudo correlation matrix. Otherwise, this kind of decomposition is not applicable. Eq. (5) can be replaced by principal component analysis, e.g. using the eigenvalue decomposition

$$\boldsymbol{\rho}_{\mathbf{YY}}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$$

whereby $\mathbf{V}$ denotes the matrix of the eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix of the eigenvalues. Inserting the orthogonality condition $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ we obtain

$$\mathbf{V}^T\boldsymbol{\rho}_{\mathbf{YY}}\mathbf{V} = \boldsymbol{\Lambda}$$

and furthermore the spectral decomposition of the pseudo correlation matrix

$$\boldsymbol{\rho}_{\mathbf{YY}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T = \mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}{\boldsymbol{\Lambda}^{\frac{1}{2}}}^T\mathbf{V}^T$$

which is a symmetric decomposition and can be used instead of Eq. (6) for the transformation of the design matrix

$$\mathbf{Z} = \left(\mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}\mathbf{Y} \tag{9}$$

This transformation is recommended in case of a non-positive definite pseudo correlation matrix. However, in the next step, the case $N \leq n$ results in a singular correlation matrix within Eq. (9), again. Indeed, this is solvable using the Moore-Penrose pseudo inverse but this procedure would occur with the same order of error magnitude as using plain Monte Carlo simulation.

The proper orthogonal decomposition, also known as Karhunen-Loève decomposition, may serve to ensure the correlation transformation in case of $N < n$ within two steps, namely the order reduction of the design matrix by projecting high-dimensional data into a lower-dimensional space and orthogonal transformation of the sample covariance or correlation matrix to the basis of the largest eigenvalues. A bibliographical review and applications of the proper orthogonal decomposition is given in [11]. The discrete modelling of the proper orthogonal decomposition is achieved by the singular value decomposition. The singular value decomposition of the correlation matrix can be written as

$$\boldsymbol{\rho}_{\mathbf{YY}} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{V}^T$$

where $\mathbf{U}$ is an $n \times n$ orthonormal matrix containing the left singular vectors of the correlation matrix; $\mathbf{S}$ is an $n \times n$ diagonal and semi-positive definite matrix with diagonal entries containing

the singular values σ_i and $\mathbf{V}$ is an $n \times n$ orthonormal matrix containing the right singular vectors. Since, the diagonal matrix $\mathbf{S}$ contains non-negative real numbers only, we obtain a non-negative definite transformation matrix in Eq. (6) using

$$\mathbf{Z} = \left(\mathbf{U}\mathbf{S}^{\frac{1}{2}}\right)^{-1}\mathbf{Y} \tag{10}$$

There are several numerical methods for calculating the singular value decomposition, e.g. presented in [6].

3.3 Low-rank approximation of the design matrix

Unfortunately, the application of Eq. (10) for the singular case $N \leq n$ results in a numerically singular transformation matrix $\mathbf{U}\mathbf{S}^{\frac{1}{2}}$. This problem can be eliminated by projecting the design matrix $\mathbf{Y}$ from the n-dimensional variable space into a lower-dimensional space, see e.g. [1]. For the non-singular case $N > n$ a singular value decomposition of the design matrix is given by

$$\mathbf{Y} = \mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Phi}^T$$

where $\mathbf{\Psi}$ is an $N \times N$ unitary matrix, $\mathbf{\Sigma}$ is a $N \times n$ pseudodiagonal matrix with non-negative numbers on the diagonal also known as the singular values of the design matrix $\mathbf{Y}$ and $\mathbf{\Phi}$ is an $n \times n$ unitary matrix. The columns of $\mathbf{\Psi}$ and the columns of $\mathbf{\Phi}$ are called the left-singular vectors and right-singular vectors of $\mathbf{Y}$, respectively.

For the singular case we use any $m < N < n$ column-reduced and row-reduced matrices to calculate an optimal rank m approximation of the data matrix $\mathbf{Y}$ as follow

$$\mathbf{Y}_m = \mathbf{\Psi}_m\mathbf{\Sigma}_m\mathbf{\Phi}_m^T; \quad \mathbf{Y}_m \in \mathbb{R}^{N,n}, \mathbf{\Psi}_m \in \mathbb{R}^{N,m}, \mathbf{\Sigma}_m \in \mathbb{R}^{m,m}, \mathbf{\Phi}_m \in \mathbb{R}^{n,m}$$

with the leading $m \times m$ principal minor of $\mathbf{\Sigma}_m$ and the m proper orthogonal modes of the m columns of $\mathbf{\Phi}_m$. The rank m can be chosen stepwise afterwards to reach the optimality of the approximation in every iteration step of the adaptive orthogonal decomposition and rearrangement of the latin hypercubes.

4 SIMULATION EXAMPLES OF UNCORRELATED VARIABLES WITH MULTIDIMENSIONAL UNIFORMITY

4.1 Non-singular, positive definite covariance matrix

In the following example the convergence of the proposed adaptive procedure in comparison with other methods is presented to simulate uniform distributed variables $X_i^k(5; 10/(2\sqrt{3}))$ with $i = 1, ..., n = 100$ variables and $k = 1, ..., N = 120$ latin hypercube samples.

The initial error $\epsilon_{\boldsymbol{\rho}_{\mathbf{HH}}}$ of plain latin hypercube sampling of 0.323 is near the error of the Monte Carlo simulation $\epsilon_{\boldsymbol{\rho}_{\mathbf{UU}}}$ of 0.3228. The adaptive procedure converged continuously to an error of $\epsilon_{\boldsymbol{\rho}_{\mathbf{XX}}} = 0.0045$ after 100 iterations. For comparison the errors of other methods are given in the Fig. (4). In particular, optimized latin hypercube sampling[1], according to the Matlab implementation of [5], Matlab implementation lhsdesign, based on the Gram-Schmidt orthogonalization[2], as proposed in [14] and standard latin hypercube sampling algorithm[3], as shown in section (2) with rearrangement of the latin hypercubes, introduced in [8].

4.2 Singular covariance matrix and low-rank approximation of the design matrix

For the singular case $N < n$ of the simulation of $i = 1, ..., n = 100$ uniform distributed variables $X_i^k(5; 10/(2\sqrt{3}))$ and $k = 1, ..., N = 80$ latin hypercube samples it is essential to

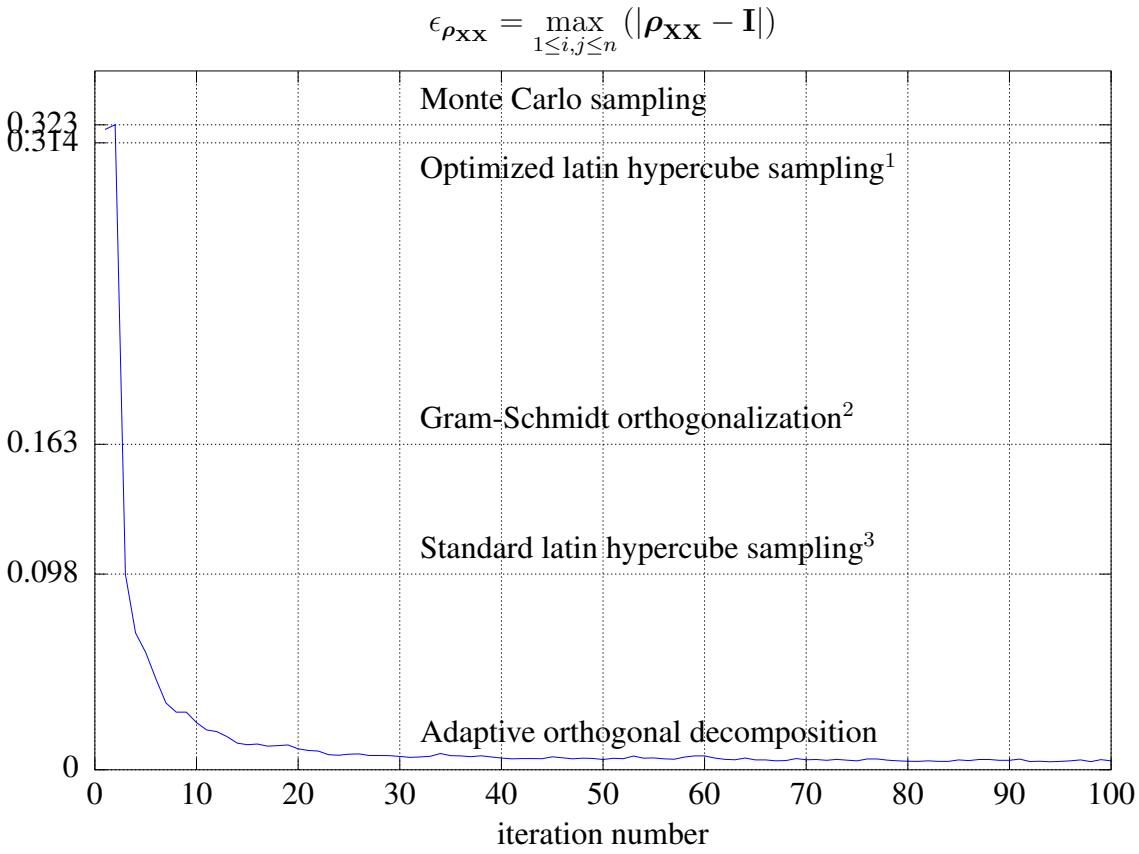


Figure 4: Convergence of the maximal correlation error $\epsilon_{\boldsymbol{\rho}_{\mathbf{XX}}}$ in regular case of $n = 100$ and $N = 120$.

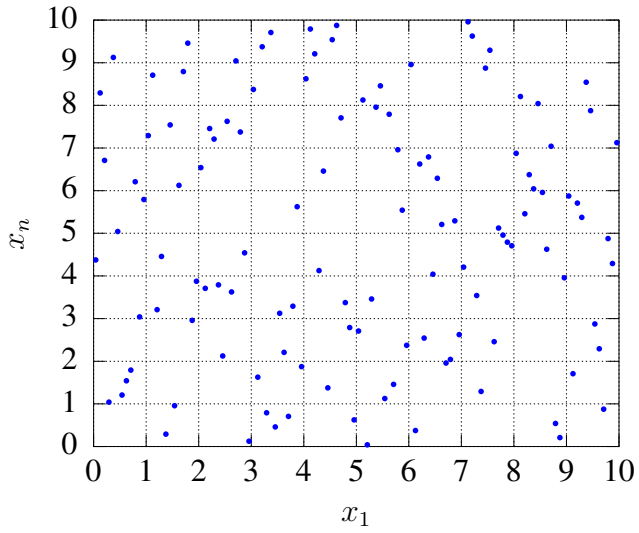


Figure 5: Anthill plot of the simulated variables x_1 and x_n in singular case of $n = 100$ and $N = 80$

calculate a low-rank approximation of the design matrix and perform the proper orthogonal decomposition in every iteration step. The initial error of plain latin hypercube sampling of $\epsilon_{\boldsymbol{\rho}_{\mathbf{HH}}} = 0.422$ is near the error of the Monte Carlo simulation. The adaptive procedure con-

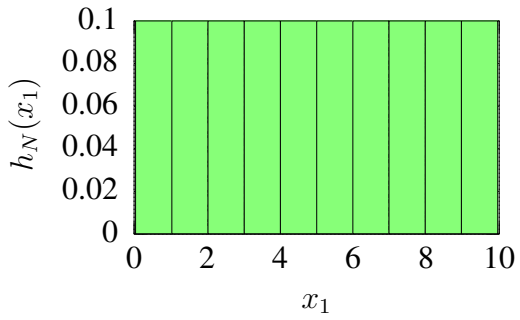


Figure 6: Histogram of the first variable x_1 in singular case of $n = 100$ and $N = 80$.

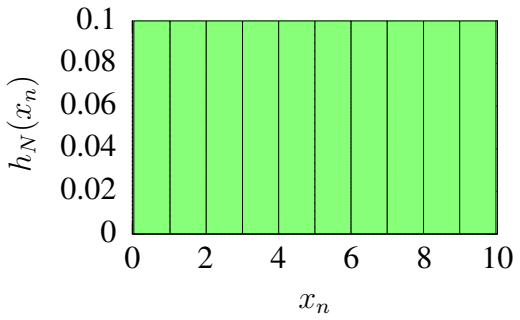


Figure 7: Histogram of the last variable x_n in singular case of $n = 100$ and $N = 80$.

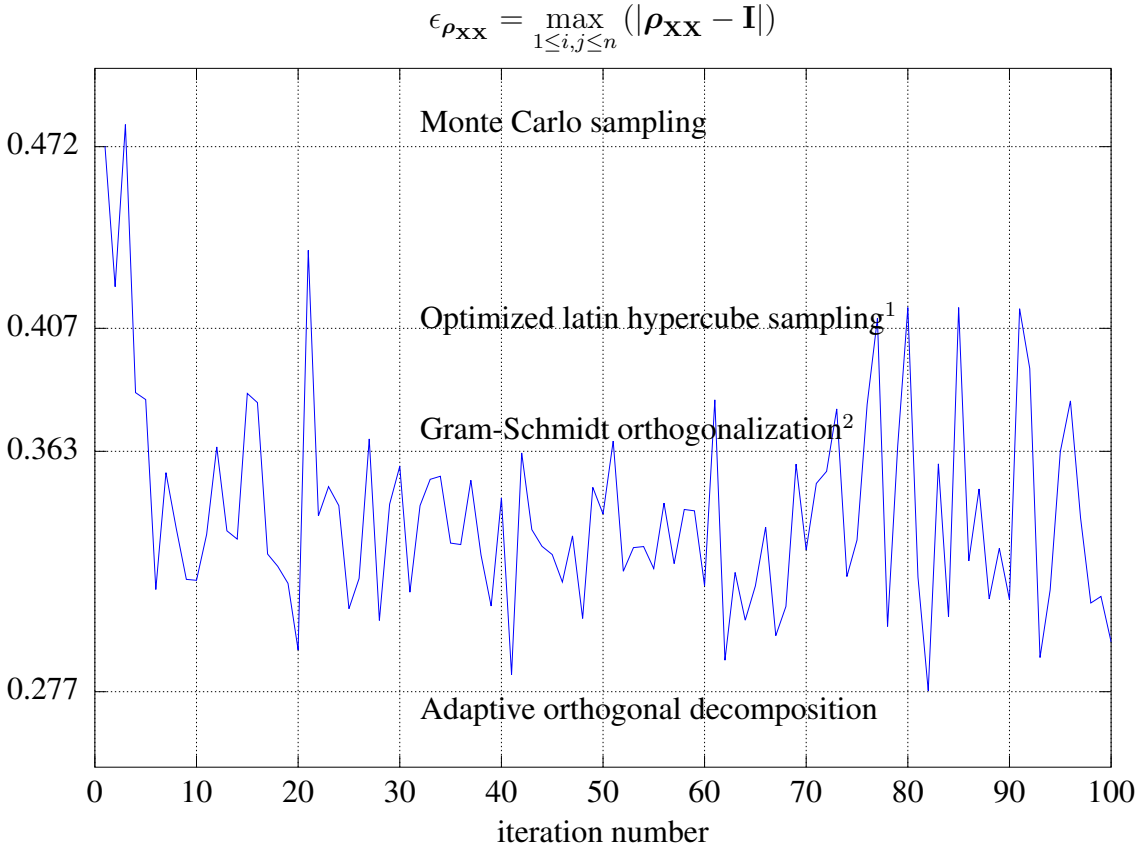


Figure 8: Convergence of the maximal correlation error $\epsilon_{\boldsymbol{\rho}_{\mathbf{XX}}}$ in singular case of $n = 100$ and $N = 80$.

verged continuously to an error of $\epsilon_{\boldsymbol{\rho}_{\mathbf{XX}}} = 0.277$ after 100 iterations, as shown in Fig. (8). The anthill plot Fig. (5) and the histograms Fig. (6) and Fig. (7) of the first and last simulated variable shows the applicability of the proposed method.

5 SUMMARY

The proposed method of proper orthogonal decomposition of the covariance matrix and low-rank approximation of the design matrix using the singular-value decomposition in combination with an enhanced adaptive rearrangement of the columns or variables of the latin hypercube de-

sign matrix and an adaptive rank-order sample resorting of the variables is a fast an accurate alternative simulation procedure to sample uncorrelated variables with multidimensional uniformity. In contrast to the standard latin hypercube sampling, the Gram-Schmidt orthogonalization and a chosen optimization-based procedure we obtain a stable convergence of the correlation error to zero, even in high-dimensional problems. For the singular case, which means a smaller number of design evaluations as the number of variables, the propose method is more stable and efficient than other tested optimization-based methods or orthogonalization methods. This is very important for an efficient design of experiment in industrial applications of the variance-based sensitivity analysis and the metamodel-based optimization.

REFERENCES

[1] A. Chatterjee. An introduction to the proper orthogonal decomposition. *Current Science*, 78(10):808–817, 2000.

[2] T.M. Cioppa and T.W. Lucas. Efficient nearly orthogonal and space-filling latin hypercubes. *Technometrics*, 49:45–55, 2007.

[3] K. Crombecq, E. Laermans, and T. Dhaene. Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. *European Journal of Operational Research*, 214(3):683 – 6965, 2011.

[4] J. L. Deutsch and C. V. Deutsch. Latin hypercube sampling with multidimensional uniformity. *Journal of Statistical Planning and Inference*, 142:763–772, 2012.

[5] Alexander I. J. Forrester, Andras Sobester, and Andy J. Keane. *Engineering Design via Surrogate Modelling - A Practical Guide.* Wiley, 2008.

[6] G. H. Golub and C. F Van Loan. *Matrix Computations*. The Johns Hopkins University Press, London, 1996.

[7] D. E. Huntington and C. S. Lyrintzis. Improvements to and limitations of latin hypercube sampling. *Probabilistic Engineering Mechanics*, 13(4):245 – 253, 1998.

[8] R. L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communication in Statistics*, B11(3):311–334, 1982.

[9] R. Jin, W. Chen, and A. Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134(1):268–287, 2005.

[10] M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

[11] G. Kerschen, J. Golinval, A. F. Vakakis, and L. A. Bergman. The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: An overview. *Nonlinear Dynamics*, 41(1):147–169, 2005.

[12] C.D. Lin, R. Mukerjee, and B. Tang. Construction of orthogonal and nearly orthogonal latin hypercubes. *Biometrika*, 96:243–247, 2009.

[13] Max D. Morris and Toby J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381 – 402, 1995.

[14] A.B. Owen. Controlling correlations in latin hypercube samples. *Journal of the American Statistical Association*, 89(428):1517–1522, 1994.

[15] M.L. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A BENCHMARK OF CONTEMPORARY METAMODELING ALGORITHMS

Can Bogoclu[1] **& Dirk Roos** [2]

Institute of Modeling and
High-Performance Computing
Niederrhein University of Applied Sciences, Germany
[1] can.bogoclu@hs-niederrhein.de
[2] dirk.roos@hs-niederrhein.de

Keywords: surrogate model, anisotropic Kriging, moving least squares, radial basis functions, support vector regression, artificial neural networks

Abstract. *Growing popularity of probabilistic and stochastic optimization methods in engineering applications has vastly increased the number of sampling points required to obtain a solution. Depending on the complexity of the underlying physical model, this often proves to be a computationally burdensome challenge. In order to overcome this challenge, one possible approach is to use surrogate models (metamodels), which approximate the responses of the physical model in a given variable subspace.*

In the past years, many different metamodeling algorithms such as Gaussian process (Kriging), moving (weighted) least squares, radial basis functions, regression neural networks and support vector regression have been suggested as an alternative to ordinary least squares (polynomial regression). The choice of the best metamodeling algorithm for any application is not a trivial task. Although there has been previous research to compare at least some of these methods to some extent ([1, 2, 3]), only a small number of publications compare all of these algorithms over a large number of multidimensional functions with varying characteristics.

In this paper, a comparison of the aforementioned methods is carried out. Using well-known analytical test functions for optimization, this paper aims to shed some light on the question of which algorithm performs best under which conditions. Apart from the structure of analytical test functions, the influence of the number of sampling points and the amount of noise in the observations on the performance of metamodeling algorithms is investigated.

1 INTRODUCTION

In physics and engineering, nature is understood and described by mathematical models, which are abstractions of reality. Although the underlying reality does not change, these models are updated constantly, whenever a newer theory proposes better results and a better understanding of our observations.

In this respect, a metamodel or a surrogate model is a mathematical abstraction of a physical model that tries to mimic the dependencies between the input and output parameters. This abstraction is useful if the computational cost of the physical model is too high and a large number of calculations are needed, as is the case for model optimization and reliability analysis. Furthermore, finding metamodels to describe the experimental data is also possible, even if a physical model is not available or is too expensive to compute.

On the other hand, some information loss is inevitable because of the abstraction. The amount of this information loss or error depends on a number of aspects. Besides the validity of the physical model, the choice of metamodeling algorithm, number of sample points used to train the algorithm and the amount of numerical or experimental noise all contribute to the amount of error in a metamodel. Therefore, this benchmark aims to test some of the existing algorithms under different conditions. There are also other sources of error, such as the inclusion of unimportant variables and the method for the design of experiments but these issues are not addressed within this benchmark.

2 METAMODELING

Before beginning with the benchmark, the algorithms to be tested will be introduced briefly. Our aim here is not to give instruction for building these algorithms but to address similarities between these algorithms by not conforming to individual choice of notations for each method and using similar symbols for similar elements of each metamodel. For more detailed instructions on how to build these algorithms, we refer to the corresponding citations.

2.1 Ordinary least squares

Ordinary least squares (OLS) or polynomial regression is one of the earliest and easiest methods for surrogate modeling. First, the input vector is mapped to a polynomial. After that, the coefficients of each monomial term are determined. The prediction function is the sum of the multiplication of coefficients with their corresponding monomials.

Let $\mathbf{x} \in \mathbb{R}^n$ be an n-dimensional vector with elements $\mathbf{x} = [x_1, x_2, x_3, \dots, x_n]$ and let $y = f(\mathbf{x})$ be the function that is trying to be approximated. The prediction function for a polynomial of degree d can be written as

$$\tilde{f}(\mathbf{x}) = \; c_0 + c_1 x_1 + \cdots + c_n x_n + c_{n+1} x_1^2 + c_{2n} x_1 x_n + \cdots + c_{\frac{n(n+1)}{2}} x_n^2 + \cdots + c_k x_n^d$$

or in matrix form

$$\begin{aligned} \tilde{f}(\mathbf{x}) &= \mathbf{p}(\mathbf{x}, d) \cdot \mathbf{c} \\ \mathbf{c} &= [c_0, c_1, c_2, ..., c_k]^T \end{aligned}$$

where $\mathbf{p}(\mathbf{x}, d)$ is the polynomial mapping function of order d for the n-dimensional input vector $\mathbf{x}$. Coefficient vector $\mathbf{c}$ must be determined by using m sampling points. The least squares problem is underdetermined for $m < k$. The relationship between the number of monomial

terms k, number of input parameters n and polynomial degree d is given by

$$k = \binom{n+d}{d} = \frac{(n+d)!}{d!n!} \tag{1}$$

where $(\cdot)!$ denotes the factorial operation. If the matrix of sample points $\mathbf{X}^0$ and the vector of responses $\mathbf{y}^0$ has more than k entries, coefficient vector $\mathbf{c}$ can be obtained by minimizing

$$\min_{\mathbf{c}} \sum_{i=0}^{m} (y_i^0 - \tilde{f}(\mathbf{x}_i^0))^2 \tag{2}$$

the sum of squared errors (SSE). This is a quadratic system of equations, which can be solved easily by the Newton's method. The only free parameter of OLS is the polynomial degree d. Figure 1a, shows that the quality of the surrogate model can increase for larger values of d. On the other hand, polynomial degree d has a large influence on the number of sampling points m needed for the approximation. Eq. 1 shows that the number of monomials increase rapidly with the increasing number of input parameters n and the polynomial degree d. This represents a great setback for this method. Further details about OLS and how to efficiently solve least squares problems can be found in [4]. In this benchmark, d is varied between 2 and 4 and the best result is recorded.

2.2 Moving least squares

Another method which has proved to be useful for metamodeling is the so called moving least squares (MLS). Similar to OLS, the input vector is mapped to a polynomial. Instead of determining global coefficients $\mathbf{c}$, local coefficients $\mathbf{c}(\mathbf{x})$ are obtained by solving a weighted least squares problem [5]. The prediction function

$$\tilde{f}(\mathbf{x}) = \mathbf{P}(\mathbf{x}, d)^T \cdot \mathbf{c}(\mathbf{x})$$

is similar to OLS, except the coefficients $\mathbf{c}(\mathbf{x})$ are not constant but functions of the approximation point $\mathbf{x}$.

Coefficients $\mathbf{c}(\mathbf{x})$ depend on the weight function $w(\mathbf{x})$ which determine the influence domain $D := \{\mathbf{x}_i^0 \in D \mid \|\mathbf{x} - \mathbf{x}_i^0\|_2 \leq \theta_r\}$ with the radius θ_r. If the sample points lie in this domain, the weight function $q(\mathbf{x})$ takes values between $0 < q(\mathbf{x}) \leq 1$, while it is set to $q(\mathbf{x}) = 0$ outside of the influence domain. Following Gaussian function [6]

$$q(r_i) = \frac{e^{-(r_i/\alpha\theta_r)^2} - e^{-\alpha^{-2}}}{1 - e^{-\alpha^{-2}}}$$

is chosen as the weight function with shape parameter $\alpha = 0.5$ for this study. Parameter r_i denotes the Euclidean distance between $\mathbf{x}$ and the sample point $\mathbf{x}_i^0$. The weighted least squares problem can be written as

$$\min_{\mathbf{c}} \sum_{i=0}^{m} q(r_i)(y_i^0 - \tilde{f}(\mathbf{x}_i^0))^2$$

similar to Eq 2. It is easier to formulate and solve this system of equations in matrix form. Local coefficients can be directly obtained by solving

$$\begin{aligned} \mathbf{c}(\mathbf{x}) &= \mathbf{P}_0^T \mathbf{Q}(\mathbf{x}) \cdot \mathbf{y}^0 \cdot (\mathbf{P}_0^T \mathbf{Q}(\mathbf{x}) \mathbf{P}_0)^{-1} \\ \mathbf{Q} &= \operatorname{diag}(q(r_1), q(r_2), \ldots, q(r_m)) \\ \mathbf{P}_0 &= \left[\mathbf{P}(\mathbf{x}_1^0, d), \ldots, \mathbf{P}(\mathbf{x}_m^0, d)\right]^T \end{aligned} \tag{3}$$

where r_i denotes the distance between the prediction point and sample point $\mathbf{x}_i^0$. This expression must be computed for each prediction point. This is illustrated in Figure 1b for different sample points. Each colour represents the local least squares approximation of a prediction point. Although the prediction function may seem to be more complicated, better approximations with lower degrees of polynomials can be achieved by this method. The polynomial degree in Figure 1b and during this benchmark is set to 2. Compared to Figure 1a, it can be seen that the quality of approximation is much better for the same polynomial degree. On the other hand, the radius of the influence domain θ_r is a free parameter and must be determined by optimization. This process is called training. Training becomes computationally more expensive with increasing number of free parameters. Nevertheless, this one dimensional optimization problem can often be solved quite fast even with stochastic optimization methods.

2.3 Kriging (Gaussian process)

Kriging is an interpolation method, which uses the spatial correlation between the data points to obtain weights for any prediction point. The prediction function for the Kriging model can be written as [7]

$$\tilde{f}(\mathbf{x}) = \mathbf{P}(\mathbf{x}, d)\hat{\mathbf{c}} + \mathbf{q}(\mathbf{x})\mathbf{Q}(\mathbf{X}^0)^{-1}(\mathbf{y}^0 - \mathbf{P}_{0,d}^T\hat{\mathbf{c}}) \tag{4}$$

and consists of two additive terms.

The first term of the summation is a polynomial regression. In ordinary Kriging, the polynomial degree is zero, hence the first term is a constant value. This can also be seen as the expectation value of a Gaussian process that this model represents. In the case of universal Kriging, polynomial degree may be greater than zero. In this case, the Gaussian process has different expected values at different locations.

The second term of the summation is the interpolating part. If the model is trained well, this part is equal to the error of the polynomial regression at each observed point. There are different variations of the spatial correlation functions $\mathbf{q}(\mathbf{x})$ and the choice of the correlation function has great impact on the surrogate model. Squared exponential

$$q_{i,sq}(\mathbf{x}) = \exp\left(-\sum_{j=1}^{n}\theta_j\left(x_j - X_{i,j}^0\right)^2\right) \tag{5}$$

and absolute exponential functions

$$q_{i,abs}(\mathbf{x}) = \exp\left(-\sum_{j=1}^{n}\theta_j\left|x_j - X_{i,j}^0\right|\right) \tag{6}$$

are used in this benchmark. Notice the index j in Eq. 5 and Eq. 6. It denotes that there may be different values for the free parameter θ_j for different input dimensions. This is the difference between anisotropic and the other types of Kriging mentioned above. For the latter, θ_j is constant for all dimensions. Theoretically, the set of possible solutions for anisotropic Kriging includes the solutions for other types of Kriging (i.e. setting all polynomial coefficients to zero except the constant term and setting all values of θ_k to some constant). Hence the global optimal solution for anisotropic Kriging should be at least as good as the other types of Kriging. On the other hand, there is no guarantee that the chosen optimization algorithm will always find the global optimum. Universal and anisotropic Kriging are therefore both trained with both correlation functions. The best result for each function is shown in the benchmark.

To determine θ, the maximum likelihood function is used as an alternative to SSE as the objective function of optimization. After θ is found, coefficients $\hat{\mathbf{c}}$ can be obtained through the solution of quadratic equation systems as in OLS 2 and in MLS 3.

Figure 1c shows the first part of the Eq. 4 and the final solution of Kriging compared to OLS for a second order polynomial. It is noticeable that the polynomial provides only a plane for the mean value, while the correlation function accounts for the variance on each data point and makes this method an interpolation. The polynomial degree d for this method is also set to 2 during this benchmark.

2.4 Radial basis functions

Another way of metamodeling is to interpolate the data points with radial basis functions (RBF) of distance. The prediction function

$$\tilde{f}(\mathbf{x}) = \sum_{i=1}^{m} c_i q\left(\left\|\mathbf{x} - \mathbf{x}_i^0\right\|_2, \theta\right) \tag{7}$$

is a linear combination of radial basis functions $q(\cdot)$. In the above equation, $\mathbf{x}_i^0$ denotes the i-th of m data points and $\|\cdot\|_2$ the Euclidean norm. Function $q(\cdot)$ is the chosen radial basis function which involves a free parameter θ. The choice of the right RBF is crucial and there are many possibilities. Some authors even suggest using anisotropic [8] (like in anisotropic Kriging) or locally bounded [1] (like in MLS) radial basis functions. In our preliminary tests, we found that the following three RBF outperformed others for our test functions. Hence, we chose to use the multiquadratic RBF

$$q(\mathbf{x}, \mathbf{x}_i^0) \quad = \sqrt{\|\mathbf{x} - \mathbf{x}_i^0\|_2^2 + \theta^2}$$

the cubic RBF

$$q(\mathbf{x}, \mathbf{x}_i^0) \quad = (\|\mathbf{x} - \mathbf{x}_i^0\|_2 + \theta)^3$$

and the Gaussian RBF

$$q(\mathbf{x}, \mathbf{x}_i^0) \quad = \exp(-\theta \|\mathbf{x} - \mathbf{x}_i^0\|_2^2)$$

for our benchmark. Like for Kriging, the best solution is chosen with a posteriori knowledge and shown in the benchmark. After choosing the radial basis function and obtaining the free parameter θ, coefficients c_i can be determined by minimizing SSE as before and the metamodel is trained. Since there are equal number of coefficients to the data points, the solution is an interpolation between these points through the RBF. This process is visualized in Figure 1d for Gaussian RBF.

2.5 Support vector regression

Originally a classification algorithm, support vector machines seek to find the hyperplane that separates the data with a maximized minimum distance to the margin [9]. In ϵ-support vector regression (SVR) this margin is used as the approximation function with a desired precision ϵ. This is demonstrated in Figure 1e for a data set with an exaggerated amount of error for visual purposes. Furthermore, the flattest plane is sought that achieves the above condition if more than one exist.

For linearly separable data, the prediction function

$$\tilde{f}(\mathbf{x}) = \hat{\mathbf{c}}^T\mathbf{x} + b$$

is similar to OLS. However, the coefficients are obtained not through minimizing the SSE but by solving the following optimization problem

$$\min \frac{1}{2}\|\mathbf{c}\|_2^2 + \theta^P \sum_{i=r}^{m}(\zeta_i + \zeta_i^*) \tag{8}$$
$$\text{s.t.} \begin{cases} y_i^0 - \hat{\mathbf{c}}^T\mathbf{x} - b \leq \epsilon + \zeta_i \\ -y_i^0 + \hat{\mathbf{c}}^T\mathbf{x} + b \leq \epsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \end{cases}$$

to ensure flatness. ζ_i and ζ_i^* are slack variables in Eq. 8 which relax the constraints and the free parameter θ^P is the penalty term for the violation of the constraints with ϵ. This problem can be solved for a given free parameter ϵ and θ^P. Since we seek to optimize the free parameters, training of this metamodel becomes a two step optimization problem as in Kriging, MLS and RBF.

For the solution of nonlinear problems, the input space is mapped to a higher dimensional space by the mapping operator $\phi(\mathbf{x})$ for linearization. The prediction function is modified as

$$\tilde{f}(\mathbf{x}) = \hat{\mathbf{c}}^T\phi(\mathbf{x}) + b$$

for this purpose. Obtaining the mapping function $\phi(\mathbf{X})$ explicitly is often very difficult. Instead, the so called kernel trick is generally used to define the mapping function implicitly by defining a kernel function $q_{i,j}(\mathbf{x}_i, \mathbf{x}_j)$ such that the kernel function

$$q(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$$

replaces the mapping operator $\phi(\mathbf{x})$ and the prediction function

$$\tilde{f}(\mathbf{x}) = \sum_{i=r}^{m} c_i q(\mathbf{x}_i^0, \mathbf{x}) + b$$

can be expressed without $\phi(\mathbf{x})$. It can be seen that only $m - r$ of the m data points are used for the prediction. In this benchmark, r is set to zero since the number of sample points are relatively small. As kernel functions, sigmoid function

$$q(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\theta\mathbf{x}_i^T\mathbf{x}_j)$$

polynomial of degree $d = 2$

$$q(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j + \theta)^d$$

and the Gaussian radial basis function

$$q(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\theta^2}\right)$$

are used. As before, the best results for each test function are shown but in practical applications, choosing the right kernel may become exhausting. The optimization problem in Eq. 8 must be reformulated to include the kernel mapping. This can be found in [9] and will not be expressed here.

2.6 Artificial neural networks

An artificial neural network (ANN) is a network of mathematical mapping operators, called neurons. Neurons take an input (a scalar, a vector or a matrix) that is passed to them and map it to an output with the activation function (kernel) that is stored in them. The Network consists of one or more layers. Neurons on each layer are connected to the neighboring layers. A more detailed description can be found in [10]. Among other purposes like classification and optimization, an ANN can also be used for metamodeling. In this benchmark two variations of radial basis neural networks are tested for this purpose. Such an ANN consists of one input layer, one or more hidden layers and one output layer. For each input dimension, there is a neuron on the input layer. The first hidden layer contains the radial basis neurons and the output layer contains the approximated value for the given input (Figure 1f).

The prediction function is given as

$$\tilde{f}(\mathbf{x}) = \sum_{j=1}^{m} c_j \phi_j(\mathbf{x}) + b_j$$
$$\phi_j(\mathbf{x}) = \exp\left(-\theta_j \left\|\mathbf{x} - \mathbf{x}_j^0\right\|_2\right)^2$$

where ϕ_j is the output of j-th neuron on the hidden layer and m is the number of neurons on the hidden layer which is set to the number of data points in this benchmark. Free parameter θ is the so called bias of the neurons in the hidden layer and must be obtained through optimization. Coefficient vectors $\mathbf{c}$ and $\mathbf{b}$ can be found linearly by minimizing SSE for a given θ.

The similarity between the Gaussian radial basis function and such a network is obvious. Apart from the additive coefficients b_j, the prediction functions are the same, although the formulations are quite different. Such a network also interpolates the data.

Another more generalized formulation of a regression neural network is also tested in this benchmark as described in [11]. The main difference between both formulations are in the second layer so only the first part or the second layer

$$\tilde{f}(\mathbf{x}) = \sum_{j=1}^{m} \frac{\mathbf{y}_0^T \phi_j(\mathbf{x})}{\|\phi_j(\mathbf{x})\|_2}$$

of the prediction function is modified. As before, both of these methods are tested and the best results for each test function are shown in the benchmark.

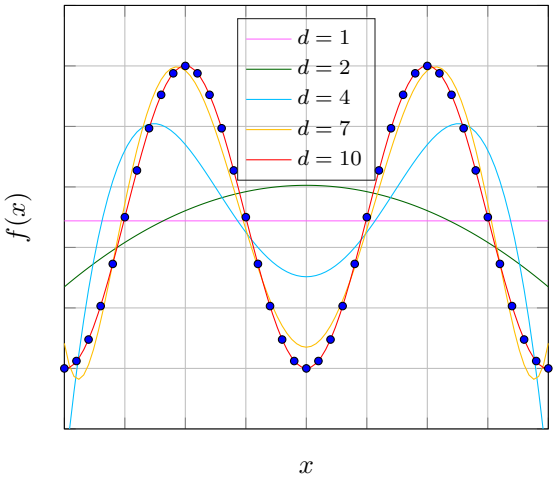
3 BENCHMARK

The benchmark consists of 15 test functions. DoEs are created for each function and each number of sample points by an optimized LHS method similar to [12], that takes both distance and correlation into account. First, the number of sample points is varied stepwise. Consecutively, Gaussian white noise is introduced to the model additively

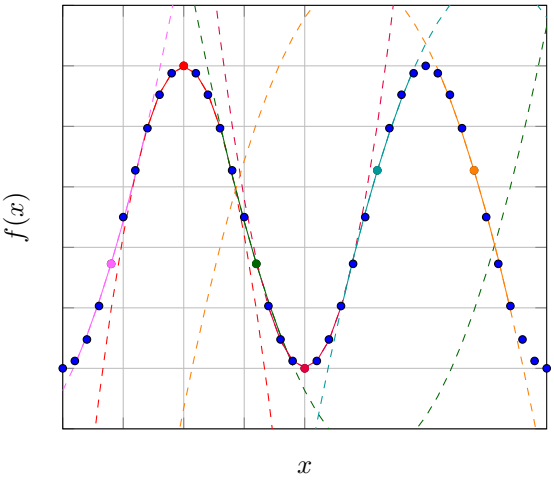
$$\hat{y}_i^0 = y_i^0(1 + \mathcal{N}(0, L_N))$$

and the level of noise L_N is varied stepwise for a fixed number of sample points.

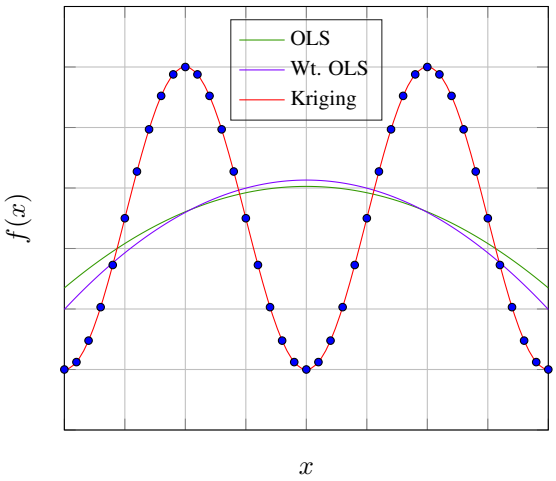
Furthermore, optimization of the metamodels is achieved by a particle swarm optimization algorithm to evade local minimums in combination with a local deterministic optimizer. A k-fold cross-validation is used to avoid overfitting. In this method, the training data is divided into



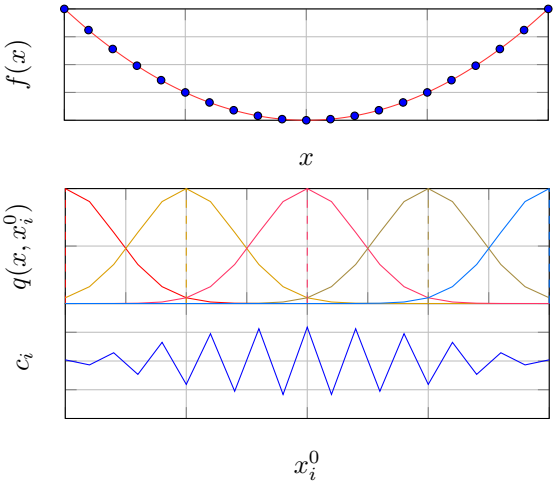
(a) OLS: Colours represent the polynomial order. Blue points represent the training data.



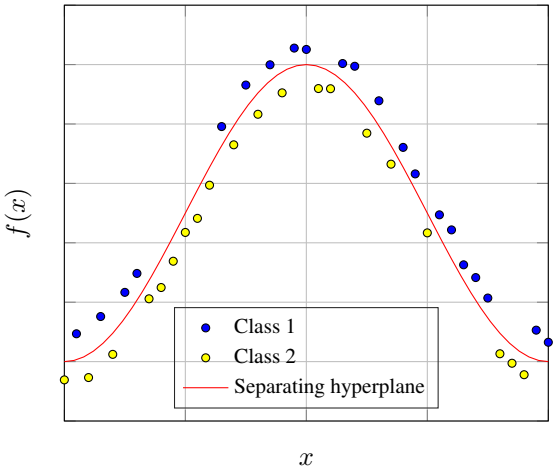
(b) MLS: Colours represent prediction points and their local approximation.



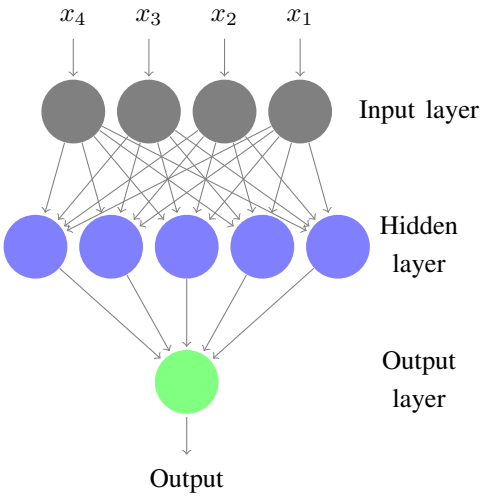
(c) Kriging (GP): The correlation matrix is used as weights for OLS. Errors of Wt. OLS are interpreted as variance and corrected by spatial correlation.



(d) RBF: The prediction function (top) is the linear combination of RBF (middle) and the coefficient vector c_i (bottom). Colours represent prediction points.



(e) SVR: Error on data points are exaggerated to visualize classification and the margin used as metamodel.



(f) ANN: All neurons are mathematical transformations that are interconnected through the layers.

Figure 1: Visualisation of metamodeling methods

k packages. $k-1$ packages are used for the training and one package for testing each iteration. The SSE is calculated by iterating over all test packages and used as the objective function for determining the free parameters. The number of folds k was set to 5.

Finally, metamodels are trained with the original and the standardized data to record the best result. Standardization of a variable $\mathbf{z}$ is done by

$$\overline{\mathbf{z}} = \frac{(\mathbf{z} - \mu_z)}{\sigma_z} \tag{9}$$

with mean vector μ_z and variation vector σ_z^2. Standardized data includes a standardized input matrix $\overline{\mathbf{x}}^0$ and a standardized output vector $\overline{\mathbf{y}}^0$.

3.1 Test functions

The following test functions were chosen for this benchmark:

1. Ackley function [13]:

$$f(\mathbf{x}) = -20\exp\left(-\frac{1}{5}\sqrt{\frac{1}{2}(x_1^2+x_2^2)}\right) - \exp\left(\frac{1}{2}(\cos(2\pi x_1)+\cos(2\pi x_2))\right) + e + 20$$
$$-5 \leq x_{1,2} \leq 5$$

2. Beale function [14]:

$$f(\mathbf{x}) = (1.5 - x_1 + x_1x_2)^2 + \left(2.25 - x_1 + x_1x_2^2\right)^2 + \left(2.625 - x_1 + x_1x_2^3\right)$$
$$-4.5 \leq x_{1,2} \leq 4.5$$

3. Booth function [14]:

$$f(\mathbf{x}) = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2$$
$$-10 \leq x_{1,2} \leq 10$$

4. Bukin function no. 6 [15]:

$$f(\mathbf{x}) = 100\sqrt{|x_2 - 0.01x_1^2|} + 0.01|x_1 + 10|$$
$$-15 \leq x_1 \leq -5;\ -3 \leq x_2 \leq 3$$

5. Cross-In-Tray function [16]:

$$f(\mathbf{x}) = -10^{-4}\left(\left|\sin(x_1)\sin(x_2)\exp\left(\left|100 - \frac{\sqrt{x_1^2+x_2^2}}{\pi}\right|\right)\right| + 1\right)^{0.1}$$
$$-5 \leq x_{1,2} \leq 5$$

6. Custom probability density function:

$$f(\mathbf{x}) = \left(1 + x_1^4 + 5x_2^4 + 2x_2^2\right)^{-1}$$
$$-3 \leq x_{1,2} \leq 3$$

7. Easom function [17]:

$$f(\mathbf{x}) = -\cos(x_1)\ \cos(x_2)\ \exp\left(-(x_1-\pi)^2-(x_2-\pi)^2\right)$$
$$0 \leq x_{1,2} \leq 6$$

8. Eggholder function [16]:

$$f(\mathbf{x}) = -(x_2+47)\sin\left(\sqrt{\left|x_2+\frac{x_1}{2}+47\right|}\right) - x_1\sin\left(\sqrt{|x_1-x_2-47|}\right)$$
$$-512 \leq x_{1,2} \leq 512$$

9. Goldstein-Price function [18]:

$$f(\mathbf{x})) = \left(1+(x_1+x_2+1)^2\left(19-14x_1+3x_1^2-14x_2+5x_1x_2+3x_2^2\right)\right)$$
$$\cdot\left(30+(2x_1-3x_2)^2\left(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2\right)\right)$$
$$-2 \leq x_{1,2} \leq 2$$

10. Holder Table function [16]:

$$f(\mathbf{x}) = -\left|\sin(x_1)\cos(x_2)\exp\left(\left|1-\pi^{-1}\sqrt{x_1^2+x_2^2}\right|\right)\right|$$
$$-3 \leq x_{1,2} \leq 3$$

11. Levy function No.13 [16]:

$$f(\mathbf{x}) = \sin^2(3\pi x_1)+(x_1-1)^2\left(1+\sin^2(3\pi x_2)\right)+(x_2-1)^2\left(1+\sin^2(2\pi x_2)\right)$$
$$-6 \leq x_{1,2} \leq 6$$

12. Michalewicz function [19]:

$$f(\mathbf{x}) = -\sum_{i=1}^{n}\sin(x_i)\sin^{20}\left(\frac{i\cdot x_i}{\pi}\right)$$
$$0 \leq x_{1,2} \leq \pi$$

13. Six hump camel function [20]:

$$f(\mathbf{x}) = 4x_1^2-2.1x_1^4+\frac{1}{3}x_1^6+x_1x_2-4x_2^2+4x_2^42$$
$$-2 \leq x_{1,2} \leq 2$$

14. Three hump camel function [16]:

$$f(\mathbf{x}) = 4x_1^2-2.1x_1^4+\frac{1}{3}x_1^6+x_1x_2-4x_2^2+4x_2^42$$
$$-5 \leq x_{1,2} \leq 5$$

15. Ursem function [21]:

$$f(\mathbf{x}) = -\sin(2x_1 - 0.5\pi) - 3\cos(x_2) - 0.5x_1$$
$$-2 \leq x_{1,2} \leq 2$$

Some of these functions are very easy to approximate (i.e. functions 2 and 3), while some are extremely challenging (i.e. functions 8 and 10). The majority of these functions should be moderately difficult to approximate for the tested metamodeling algorithms. In this way, some variance on the complexity of the structure of the test functions is achieved. The purpose of this variance is to reduce the bias of the benchmark such that no algorithm is favoured. Functions are expressed by means of equations and input bounds.

3.2 Results

Root mean squared error (RMSE) of 16384 validation points created from a Sobol sequence [22] is used as results in Table 1 and Table 2. Smaller values of RMSE represent a better model quality. The number of sample points are varied incrementally between 32, 64, 96 and 128 first. Consecutively, noise levels are varied between 1%, 2.5%, 5% and 10% incrementally while the DoEs with 128 points are reused. It can be argued that 10% noise is quite a lot but since we suggest using metamodels to model experimental data, such a high level of noise can still be interesting.

Besides the RMSE for each function, the mean value $mean_{Err}$ of RMSE and the standardized mean value $\overline{mean}_{Err}$ are displayed for each category (see Figure 2 and Figure 3). The mean value $mean_{Err}$ can be used as a measure to compare an algorithm with itself between the categories but it is biased since functions with greater prediction errors have a bigger impact on the results. Standardization is needed to equalize the weight of each function and make the score more independent of the magnitude of errors. Hence, each evaluation of a test function is standardized with the mean and the standard deviation of the root mean squared error of all algorithms for that evaluation. The mean value of the standardized error $\overline{mean}_{Err,k}$ of algorithm k can be expressed as shown in Eq. 9

$$\xi_{i,k} = \frac{RMSE_{i,k} - mean(RMSE_{i,all})}{std(RMSE_{i,all})}$$
$$\overline{mean}_{Err,k} = \frac{\sum_{i=1}^{15} \xi_{i,k}}{15}$$

and will be used for the categorical comparison of algorithms with each other. The variable $RMSE_{i,all}$ denotes the RMSE of all algorithms for the function i and does possess the same mean value as $mean_{Err}$ which denotes the mean of the RMSE along all functions in a category.

Samp.	Func.	ANN	GP	MLS	OLS	RBF	SVR
32	1	$7.17E-01$	$7.69E-01$	$7.83E-01$	$1.11E+00$	$\mathbf{6.95E-01}$	$7.83E-01$
	2	$8.48E+03$	$7.85E+03$	$1.19E+04$	$1.53E+04$	$\mathbf{3.30E+03}$	$1.82E+04$
	3	$2.41E-04$	$\mathbf{1.26E-13}$	$4.66E-13$	$6.24E-12$	$4.74E-04$	$2.37E+01$
	4	$1.37E+01$	$1.31E+01$	$1.49E+01$	$1.61E+01$	$1.31E+01$	$\mathbf{1.19E+01}$
	5	$2.03E-01$	$\mathbf{1.13E-01}$	$2.02E-01$	$2.11E-01$	$1.77E-01$	$2.04E-01$
	6	$\mathbf{6.11E-02}$	$6.58E-02$	$6.84E-02$	$1.51E-01$	$6.12E-02$	$1.04E-01$
	7	$\mathbf{5.40E-02}$	$6.43E-02$	$7.46E-02$	$1.40E-01$	$5.73E-02$	$6.09E-02$
	8	$\mathbf{2.91E+02}$	$3.23E+02$	$3.64E+02$	$3.51E+02$	$2.98E+02$	$3.18E+02$
	9	$\mathbf{1.41E+04}$	$1.92E+04$	$4.13E+04$	$5.13E+04$	$1.44E+04$	$3.74E+04$
	10	$2.72E+00$	$2.62E+00$	$4.92E+00$	$3.08E+00$	$2.69E+00$	$\mathbf{2.60E+00}$

Samp.	Func.	ANN	GP	MLS	OLS	RBF	SVR
64	11	$1.21E+01$	$1.10E+01$	$\mathbf{1.05E+01}$	$1.09E+01$	$1.11E+01$	$1.17E+01$
	12	$3.02E-01$	$\mathbf{1.28E-01}$	$2.40E-01$	$2.86E-01$	$2.33E-01$	$3.05E-01$
	13	$\mathbf{1.24E-01}$	$8.48E-01$	$2.84E+00$	$5.18E+00$	$2.16E-01$	$2.93E+00$
	14	$5.76E+01$	$\mathbf{6.32E-01}$	$2.15E+02$	$2.34E+02$	$5.60E+01$	$1.56E+02$
	15	$\mathbf{4.27E-02}$	$4.70E-02$	$1.46E-01$	$3.14E-01$	$3.21E-01$	$4.28E-02$
	1	$7.58E-01$	$\mathbf{6.58E-01}$	$7.17E-01$	$1.04E+00$	$6.73E-01$	$6.98E-01$
	2	$\mathbf{2.62E+02}$	$5.67E+03$	$6.29E+03$	$1.48E+04$	$5.48E+02$	$8.38E+03$
	3	$2.28E-04$	$\mathbf{1.17E-13}$	$4.95E-13$	$3.94E-12$	$3.05E-04$	$1.42E+01$
	4	$9.51E+00$	$8.18E+00$	$1.02E+01$	$1.54E+01$	$\mathbf{7.98E+00}$	$1.03E+01$
	5	$1.54E-01$	$\mathbf{8.48E-02}$	$2.02E-01$	$2.06E-01$	$1.48E-01$	$1.98E-01$
	6	$\mathbf{2.21E-02}$	$4.67E-02$	$3.24E-02$	$1.50E-01$	$2.25E-02$	$2.45E-02$
	7	$9.80E-03$	$3.34E-02$	$1.61E-02$	$1.39E-01$	$\mathbf{5.00E-03}$	$5.82E-03$
	8	$2.87E+02$	$2.80E+02$	$3.38E+02$	$3.45E+02$	$\mathbf{2.61E+02}$	$2.98E+02$
	9	$1.11E+03$	$5.79E+03$	$9.95E+03$	$4.83E+04$	$\mathbf{1.02E+03}$	$1.40E+04$
	10	$2.55E+00$	$\mathbf{2.12E+00}$	$2.88E+00$	$2.65E+00$	$2.59E+00$	$2.54E+00$
	11	$9.36E+00$	$\mathbf{8.47E+00}$	$9.86E+00$	$9.45E+00$	$1.05E+01$	$1.10E+01$
	12	$1.68E-01$	$\mathbf{4.15E-02}$	$1.84E-01$	$2.78E-01$	$1.61E-01$	$2.12E-01$
	13	$\mathbf{2.79E-03}$	$3.17E-01$	$1.40E+00$	$5.07E+00$	$1.66E-02$	$1.50E+00$
	14	$9.54E-01$	$\mathbf{4.58E-02}$	$8.06E+01$	$2.28E+02$	$1.24E+00$	$1.01E+02$
	15	$\mathbf{2.01E-02}$	$5.31E-02$	$5.41E-02$	$3.13E-01$	$2.14E-01$	$2.97E-02$
96	1	$7.55E-01$	$\mathbf{4.83E-01}$	$6.80E-01$	$1.03E+00$	$7.13E-01$	$6.78E-01$
	2	$\mathbf{9.80E+01}$	$1.52E+03$	$3.32E+03$	$1.43E+04$	$1.48E+02$	$4.69E+03$
	3	$1.22E-04$	$\mathbf{1.37E-13}$	$5.02E-13$	$5.05E-12$	$2.96E-04$	$8.21E+00$
	4	$8.52E+00$	$6.20E+00$	$1.20E+01$	$1.51E+01$	$\mathbf{5.18E+00}$	$9.10E+00$
	5	$1.39E-01$	$\mathbf{6.85E-02}$	$1.98E-01$	$1.99E-01$	$1.33E-01$	$1.49E-01$
	6	$1.57E-02$	$3.66E-02$	$2.42E-02$	$1.50E-01$	$\mathbf{1.17E-02}$	$1.44E-02$
	7	$\mathbf{6.07E-04}$	$1.22E-02$	$1.37E-02$	$1.39E-01$	$1.19E-03$	$1.91E-03$
	8	$2.99E+02$	$2.70E+02$	$2.90E+02$	$3.12E+02$	$\mathbf{2.53E+02}$	$2.99E+02$
	9	$\mathbf{1.43E+02}$	$1.11E+03$	$6.17E+03$	$4.83E+04$	$2.12E+02$	$1.01E+04$
	10	$2.36E+00$	$\mathbf{1.85E+00}$	$2.48E+00$	$2.53E+00$	$2.16E+00$	$2.46E+00$
	11	$9.27E+00$	$\mathbf{9.24E+00}$	$9.40E+00$	$9.78E+00$	$1.08E+01$	$9.50E+00$
	12	$1.36E-01$	$\mathbf{2.45E-02}$	$2.47E-01$	$2.78E-01$	$1.33E-01$	$1.37E-01$
	13	$\mathbf{2.30E-03}$	$1.93E-01$	$6.40E-01$	$4.97E+00$	$2.84E-03$	$3.92E-01$
	14	$6.18E-01$	$\mathbf{2.88E-03}$	$4.84E+01$	$2.34E+02$	$8.48E-01$	$8.11E+01$
	15	$1.47E-02$	$3.35E-02$	$1.87E-02$	$3.12E-01$	$2.15E-01$	$\mathbf{1.06E-02}$
128	1	$6.40E-01$	$\mathbf{3.63E-01}$	$6.54E-01$	$1.02E+00$	$6.88E-01$	$6.62E-01$
	2	$4.63E+01$	$7.29E+02$	$2.52E+03$	$1.43E+04$	$\mathbf{2.56E+01}$	$2.61E+03$
	3	$1.14E-04$	$\mathbf{1.18E-13}$	$5.34E-13$	$4.34E-12$	$2.14E-04$	$2.55E+00$
	4	$7.79E+00$	$6.03E+00$	$7.07E+00$	$1.51E+01$	$\mathbf{4.70E+00}$	$5.98E+00$
	5	$1.36E-01$	$\mathbf{6.34E-02}$	$1.51E-01$	$2.00E-01$	$1.32E-01$	$1.22E-01$
	6	$6.89E-03$	$2.29E-02$	$2.63E-02$	$1.51E-01$	$\mathbf{5.82E-03}$	$7.95E-03$
	7	$\mathbf{4.98E-05}$	$8.45E-03$	$1.43E-02$	$1.40E-01$	$2.04E-04$	$1.75E-03$
	8	$2.98E+02$	$2.45E+02$	$2.58E+02$	$3.06E+02$	$\mathbf{2.30E+02}$	$3.00E+02$
	9	$\mathbf{8.83E+01}$	$3.73E+02$	$9.43E+03$	$4.96E+04$	$9.26E+01$	$6.03E+03$
	10	$2.40E+00$	$\mathbf{1.78E+00}$	$2.18E+00$	$2.36E+00$	$1.98E+00$	$2.48E+00$
	11	$1.05E+01$	$9.79E+00$	$\mathbf{9.54E+00}$	$1.00E+01$	$1.15E+01$	$1.08E+01$
	12	$1.40E-01$	$\mathbf{1.40E-02}$	$7.25E-01$	$2.77E-01$	$1.12E-01$	$1.16E-01$
	13	$1.26E-03$	$1.33E-02$	$4.52E-01$	$4.98E+00$	$\mathbf{9.79E-04}$	$5.01E-01$
	14	$3.18E-01$	$\mathbf{2.41E-03}$	$4.88E+01$	$2.30E+02$	$3.16E-01$	$3.53E+01$
	15	$1.27E-02$	$2.57E-02$	$1.83E-02$	$3.11E-01$	$2.22E-01$	$\mathbf{1.25E-02}$

Table 1: Root mean squared error of prediction at 16384 validation points for various number of sample points

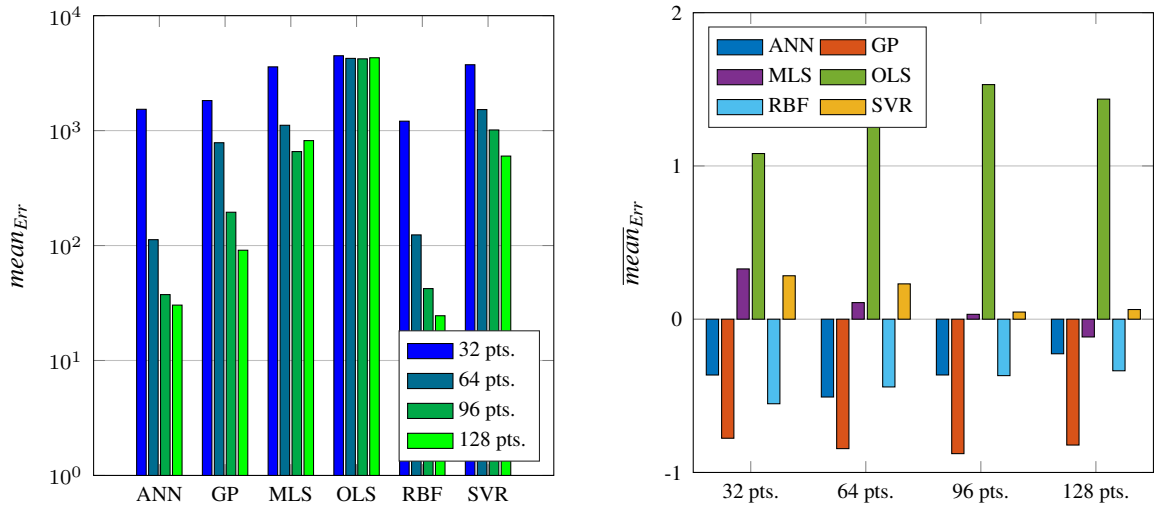


Figure 2: $mean_{Err}$ and $\overline{mean}_{Err}$ of the RMSE values for varied number of sample points

Table 1 shows the RMSE of 16384 validation points for a varying number of sample points. Smaller values of error are better and bold values represent the best result for each function. It can can be seen that the RMSE generally decreases with the increasing number of sample points. There are contradictions to this observation (i.e. Function 3), where the prediction error is already quite low for few samples and becomes larger because of oversampling.

Furthermore, it can be seen that different algorithms have the best score for different functions. Figure 2 shows that Kriging (GP) outperforms other algorithms on average, followed by ANN and RBF. OLS seems to have the poorest results in this part by a great margin.

Noise	Func.	ANN	GP	MLS	OLS	RBF	SVR
	1	$6.48E-01$	$\mathbf{3.98E-01}$	$6.63E-01$	$7.06E-01$	$6.98E-01$	$6.61E-01$
	2	$\mathbf{2.31E+02}$	$7.22E+02$	$2.56E+03$	$5.90E+03$	$2.79E+02$	$2.60E+03$
	3	$3.28E+00$	$\mathbf{2.72E+00}$	$2.86E+00$	$3.28E+00$	$3.26E+00$	$4.78E+00$
	4	$7.90E+00$	$6.05E+00$	$7.64E+00$	$1.49E+01$	$\mathbf{4.84E+00}$	$6.11E+00$
	5	$1.35E-01$	$\mathbf{6.56E-02}$	$1.53E-01$	$1.89E-01$	$1.33E-01$	$1.23E-01$
	6	$6.54E-03$	$7.19E-03$	$2.67E-02$	$1.19E-01$	$\mathbf{5.65E-03}$	$9.03E-03$
	7	$4.42E-03$	$5.92E-03$	$1.40E-02$	$1.14E-01$	$3.26E-03$	$\mathbf{1.85E-03}$
1 %	8	$2.98E+02$	$2.36E+02$	$2.58E+02$	$3.06E+02$	$\mathbf{2.30E+02}$	$3.00E+02$
	9	$\mathbf{2.17E+03}$	$7.25E+03$	$8.99E+03$	$3.36E+04$	$3.58E+03$	$5.69E+03$
	10	$2.26E+00$	$\mathbf{1.62E+00}$	$2.33E+00$	$2.26E+00$	$1.98E+00$	$2.48E+00$
	11	$1.05E+01$	$9.86E+00$	$\mathbf{9.54E+00}$	$1.02E+01$	$1.15E+01$	$1.09E+01$
	12	$1.25E-01$	$\mathbf{1.50E-02}$	$7.20E-01$	$2.68E-01$	$1.12E-01$	$1.17E-01$
	13	$\mathbf{1.21E-01}$	$2.42E-01$	$5.17E-01$	$4.57E-01$	$1.48E-01$	$6.32E-01$
	14	$5.12E+00$	$\mathbf{4.91E+00}$	$2.60E+02$	$5.21E+01$	$6.21E+00$	$5.82E+01$
	15	$4.08E-02$	$\mathbf{1.11E-02}$	$1.82E-02$	$1.31E-01$	$2.28E-01$	$1.18E-02$
	1	$7.21E-01$	$\mathbf{4.57E-01}$	$6.85E-01$	$7.18E-01$	$7.25E-01$	$6.63E-01$
	2	$\mathbf{5.21E+02}$	$1.35E+03$	$2.62E+03$	$5.90E+03$	$7.08E+02$	$2.72E+03$
	3	$9.06E+00$	$\mathbf{6.81E+00}$	$7.16E+00$	$8.19E+00$	$8.68E+00$	$1.20E+01$
	4	$9.04E+00$	$6.61E+00$	$8.58E+00$	$1.50E+01$	$\mathbf{5.50E+00}$	$6.31E+00$
	5	$1.40E-01$	$\mathbf{7.51E-02}$	$1.58E-01$	$1.90E-01$	$1.38E-01$	$1.28E-01$
	6	$7.25E-03$	$8.38E-03$	$2.75E-02$	$1.19E-01$	$\mathbf{6.69E-03}$	$1.16E-02$
	7	$9.11E-03$	$1.13E-02$	$1.36E-02$	$1.14E-01$	$6.10E-03$	$\mathbf{3.99E-03}$
2.5 %	8	$2.98E+02$	$2.36E+02$	$2.58E+02$	$3.06E+02$	$\mathbf{2.30E+02}$	$2.81E+02$
	9	$\mathbf{5.97E+03}$	$1.21E+04$	$8.58E+03$	$3.34E+04$	$6.07E+03$	$9.89E+03$
	10	$2.26E+00$	$\mathbf{1.63E+00}$	$2.33E+00$	$2.26E+00$	$1.98E+00$	$2.47E+00$

Noise	Func.	ANN	GP	MLS	OLS	RBF	SVR
	11	$1.02E+01$	$9.66E+00$	$\mathbf{9.54E+00}$	$1.02E+01$	$1.14E+01$	$1.01E+01$
	12	$1.25E-01$	$\mathbf{1.50E-02}$	$7.12E-01$	$2.68E-01$	$1.12E-01$	$1.18E-01$
	13	$\mathbf{2.95E-01}$	$3.64E-01$	$6.65E-01$	$5.14E-01$	$3.74E-01$	$7.50E-01$
	14	$\mathbf{1.10E+01}$	$1.13E+01$	$2.61E+02$	$5.23E+01$	$1.28E+01$	$3.58E+01$
	15	$4.05E-02$	$1.96E-02$	$1.85E-02$	$1.31E-01$	$2.42E-01$	$\mathbf{1.46E-02}$
	1	$7.43E-01$	$7.13E-01$	$7.26E-01$	$7.47E-01$	$7.99E-01$	$\mathbf{6.69E-01}$
	2	$\mathbf{1.02E+03}$	$2.13E+03$	$2.77E+03$	$5.91E+03$	$1.15E+03$	$2.60E+03$
	3	$1.56E+01$	$\mathbf{1.36E+01}$	$1.43E+01$	$1.64E+01$	$1.65E+01$	$1.49E+01$
	4	$9.08E+00$	$7.77E+00$	$1.03E+01$	$1.51E+01$	$\mathbf{7.06E+00}$	$8.36E+00$
	5	$1.48E-01$	$\mathbf{1.01E-01}$	$1.73E-01$	$1.93E-01$	$1.48E-01$	$1.45E-01$
	6	$1.13E-02$	$1.21E-02$	$2.97E-02$	$1.19E-01$	$\mathbf{1.03E-02}$	$1.93E-02$
	7	$1.48E-02$	$1.81E-02$	$1.34E-02$	$1.14E-01$	$1.04E-02$	$\mathbf{6.76E-03}$
5 %	8	$2.99E+02$	$2.36E+02$	$2.59E+02$	$3.07E+02$	$\mathbf{2.30E+02}$	$2.99E+02$
	9	$\mathbf{6.52E+03}$	$1.23E+04$	$1.48E+04$	$3.31E+04$	$7.75E+03$	$1.06E+04$
	10	$2.26E+00$	$\mathbf{1.63E+00}$	$2.18E+00$	$2.26E+00$	$1.98E+00$	$2.46E+00$
	11	$1.02E+01$	$\mathbf{6.76E+00}$	$9.54E+00$	$1.02E+01$	$1.15E+01$	$1.04E+01$
	12	$1.33E-01$	$\mathbf{2.02E-02}$	$6.99E-01$	$2.69E-01$	$1.12E-01$	$1.19E-01$
	13	$\mathbf{5.94E-01}$	$6.27E-01$	$9.63E-01$	$6.52E-01$	$7.51E-01$	$1.10E+00$
	14	$\mathbf{2.10E+01}$	$2.23E+01$	$6.25E+01$	$5.35E+01$	$2.74E+01$	$4.52E+01$
	15	$3.24E-02$	$2.99E-02$	$2.07E-02$	$1.30E-01$	$2.20E-01$	$\mathbf{1.93E-02}$
	1	$8.05E-01$	$9.19E-01$	$1.03E+00$	$8.38E-01$	$1.02E+00$	$\mathbf{6.98E-01}$
	2	$\mathbf{1.61E+03}$	$3.52E+03$	$3.14E+03$	$5.95E+03$	$3.29E+03$	$2.78E+03$
	3	$3.71E+01$	$2.72E+01$	$2.86E+01$	$3.28E+01$	$3.54E+01$	$\mathbf{2.31E+01}$
	4	$1.13E+01$	$1.11E+01$	$1.18E+01$	$1.56E+01$	$1.07E+01$	$\mathbf{1.03E+01}$
	5	$1.69E-01$	$\mathbf{1.65E-01}$	$2.29E-01$	$2.02E-01$	$1.83E-01$	$1.66E-01$
	6	$1.99E-02$	$2.09E-02$	$3.43E-02$	$1.19E-01$	$\mathbf{1.94E-02}$	$3.22E-02$
	7	$2.69E-02$	$2.53E-02$	$1.45E-02$	$1.14E-01$	$1.67E-02$	$\mathbf{1.21E-02}$
10 %	8	$2.98E+02$	$2.37E+02$	$2.61E+02$	$3.07E+02$	$\mathbf{2.31E+02}$	$2.94E+02$
	9	$\mathbf{9.79E+03}$	$1.61E+04$	$1.61E+04$	$3.28E+04$	$1.60E+04$	$1.63E+04$
	10	$2.27E+00$	$\mathbf{1.66E+00}$	$2.20E+00$	$2.27E+00$	$2.00E+00$	$2.45E+00$
	11	$1.07E+01$	$\mathbf{7.42E+00}$	$9.57E+00$	$1.03E+01$	$1.04E+01$	$1.13E+01$
	12	$1.28E-01$	$\mathbf{3.05E-02}$	$6.73E-01$	$2.69E-01$	$1.15E-01$	$1.22E-01$
	13	$\mathbf{1.00E+00}$	$1.21E+00$	$1.01E+00$	$1.47E+00$	$1.02E+00$	$1.80E+00$
	14	$\mathbf{4.12E+01}$	$4.44E+01$	$2.64E+02$	$5.90E+01$	$5.82E+01$	$5.71E+01$
	15	$5.80E-02$	$3.87E-02$	$2.73E-02$	$1.30E-01$	$2.15E-01$	$\mathbf{2.64E-02}$

Table 2: Root mean squared error of prediction at 16384 validation points for 128 sample points and varied levels of Gaussian noise

Results for varying noise levels are displayed in Table 2. Here it can be seen that increasing noise levels increase the prediction error and some algorithms are effected more than the others. Interestingly, the only exception is SVR; data with 1% noise seems to train a better metamodel than data with no noise for some test functions and on average.This can be explained by the fact that the results of SVR without noise were not as good on average. Furthermore SVR is the only tested algorithm with free parameters to define process error (ϵ and θ^P in Eq. 8).

As before, Kriging seems to perform better than other algorithms followed by ANN and RBF on average (Figure 3) . SVR becomes comparably good for higher levels of noise. Kriging, MLS, OLS and SVR also seem to be least affected from increasing levels of noise, since the relative increase in mean value is considerably smaller than other methods. OLS seems to consistently achieve the poorest results, although the difference gets smaller with increasing levels of noise.

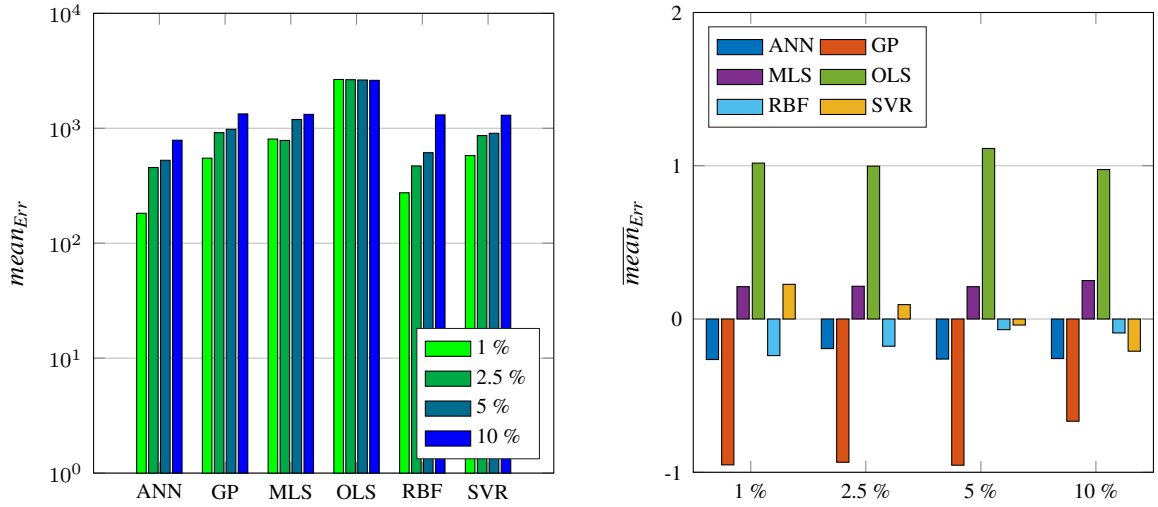


Figure 3: $mean_{Err}$ and $\overline{mean}_{Err}$ of the RMSE values for varied levels of Gaussian Noise

4 CONCLUSION

Six metamodeling algorithms are tested with 15 analytical functions in 8 different cases. A total of 720 results are included in this benchmark. The following observations can be made based on these results:

- Prediction error has a negative correlation with the number of sample points and a positive correlation with the noise level. On the other hand, it is observed that oversampling may worsen the quality of some metamodels. Furthermore, a small amount of Gaussian noise (1%) has a positive impact on the model quality of support vector regression for some test functions.
- On average, Kriging is the most accurate method followed by radial basis functions and neural networks independent of varying number of sample points and noise level. Conversely, every method except ordinary least squares delivers the best results for at least one of test functions. This shows that the axiom ”No free lunch theorem” [23] holds for metamodeling too.
- Kriging, moving least squares and support vector regression handle the noise best since relative increase of the prediction error is the smallest for these methods.
- Cross-validation helps avoid overfitting even for pure interpolating algorithms like Kriging and radial basis functions. This can be seen in the results for noisy data (Table 2).

Furthermore, the following observations were made during the benchmark:

- Standardizing data often but not always decreases the error.
- Finding the global optimum for anisotropic Kriging may sometimes be challenging for the optimizer. Therefore isotropic Kriging has better results for some functions.
- Gaussian radial basis function is often but not always the best of all tested radial basis functions.

- Radial basis function is the best of all tested kernels for support vector regression considering this benchmark.
- A better designed neural network for a specific problem may yield better results for that problem than a generalized network like in this benchmark.

We conclude that Kriging, neural networks and radial basis functions in this order are, on average, the best choices for unknown problems, at least when the constraints in this benchmark regarding the number of sample points and noise levels are fulfilled and the structural complexity is similar to those of the chosen test functions. Moreover, testing different algorithms may yield better results in some cases. Further work should be done regarding the effects of dimensionality on the quality of metamodels.

5 ACKNOWLEDGEMENT

The authors would like to express their thanks to the German Federal Ministry of Education and Research for their support for this project. Support identification no.: IN2013-671-221.

REFERENCES

[1] Krishnamurty, T., Comparison of Response Surface Construction Methods for Derivative Estimation Using Moving Least Squares, Kriging and Radial Basis Functions, *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference*, 2005

[2] Li, Y.F.; Ng, S.H.; Xie, M. and Goh, T.N., A systematic comparison of metamodeling techniques for simulation optimization in Decision Support Systems, *Applied Soft Computing*, vol.10, 1257 - 1273, 2010

[3] Luo, J. and Lu, W., Comparison of surrogate models with different methods in groundwater remediation process, *Journal of Earth System Science*, vol.123, 1579 - 1589, 2014

[4] Rao, C. R. and Toutenburg, H., Linear Models: Least Squares and Alternatives, Second Edition, *Springer Verlag*, 1999

[5] Levin, D., The Approximation Power of Moving Least Squares, *Mathematics of Computation*, vol. 67, no. 224, 1998

[6] Most, T. and Bucher, C., A Moving Least Squares weighting function for the Element-free Galerkin Method which almost fulfills essential boundary conditions, *Structural Engineering and Mechanics*, vol. 21, no. 3, 2005

[7] Krishnamurty, T. and Romero, V. J., Construction of Response Surface with Higher Order Continuity and its Application to Reliability Engineering, *43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2002

[8] Dinh, H. Q.; Turk, G. and Slabaugh, G., Reconstructing surfaces using anisotropic basis functions, *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol.2, 2001

[9] Smola, A. J. and Schlkopf,B., A Tutorial on Support Vector Regression, *Statistics and Computing*,2003

[10] Rojas ,R., Neural Networks - A Systematic Introduction, *Springer Verlag*, 1996

[11] Wasserman, P.D., Advanced Methods in Neural Computing, *Van Nostrand Reinhold*, 1993

[12] Joseph, R. V. and Hung. Y., Orthogonal-Maxmin Latin Hypercube Designs, *Statistica Sinica*, vol. 18, 2008

[13] Ackley, D. H., A Connectionist Machine for Genetic Hill-Climbing, *Kluwer Academic Publishers*, 1987

[14] Jamil, M. and Yang, X.S. A Literature Survey of Benchmark Functions For Global Optimization Problems, *Blekinge Institute of Technology*, 2013

[15] Bukin, A. D. , New Minimization Strategy For Non-Smooth Functions, *Budker Institute of Nuclear Physics*, 1997

[16] Mishra, S. K., Some New Test Functions for Global Optimization and Performance of Repulsive Particle Swarm Method, *Department of Economics, North-Eastern Hill University*, 2006

[17] Easom, E. E., A Survey of Global Optimization Techniques, *University of Louisville*, 1990

[18] Goldstein, A. A. and Price, I. F., On Descent From Local Minima, *Mathematics of Computation*, vol. 25, no. 115, 1971

[19] Michalewicz, Z., Genetich Algorithms + Data Structures = Evolution Programs, *Springer-Verlag*, 1992

[20] Dixon, L. C. and Szego, G. P., Towards Global Optimisation 2, *Elsevier Science Ltd. North Holland*, 1978

[21] Ursem, R., Multinational Evolutionary Algorithms, *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, vol.3, 1999

[22] Dalal, I.L.; Stefan, D. and Harwayne-Gidansky, J., Low Discrepancy Sequences for Monte Carlo Simulations on Reconfigurable Platforms, *Application-Specific Systems, Architectures and Processors, 2008. ASAP*, 2008

[23] Wolpert, D.H. and Macready, W.G., No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation* vol. 1, 1997

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOPOLOGY OPTIMIZATION USING A KRIGING-ASSISTED GENETIC ALGORITHM WITH A NOVEL LEVEL SET REPRESENTATION APPROACH

Mitsuo Yoshimura[1], Koji Shimoyama[1], Takashi Misaka[1], and Shigeru Obayashi[1]

[1] Institute of Fluid Science
Tohoku university
2-1-2 Katahira, Aoba-ku, Sendai, Japan
e-mail: {yoshimura, shimoyama, misaka, obayashi}@edge.ifs.tohoku.ac.jp

Keywords: Topology Optimization, Level-set Function, Computational Fluid Dynamics, Genetic Algorithm, Kriging Model.

Abstract. *Topology optimization is an optimization method which can modify connectivity of an object independently of its predefined topology. In this paper, a global optimization method for topology optimization of flow channels considering fluid and heat transfer using a genetic algorithm is presented. A genetic algorithm (GA) is assisted by the Kriging surrogate model to reduce computational cost required for function evaluation. In the present method, the boundary of a flow channel is represented by a level set function. Topology optimization seldom employs GA since topology optimization requires a large number of design variables for a high degree of freedom for shape and topology representation and GA is not effective to handle such a large scale problem. This paper presents a novel representation method to obtain the distribution of level set function with a reasonable number of design variables. The design variables are given at the scattered control points in the design domain, and the Helmholtz equation is solved in the entire domain. The proposed method is applied to a single-objective optimization problem to maximize heat transfer. As a result, GA found several flow channels, each of which has similar objective function values but with different topology. The result indicates that the objective function is a multi-modal function, which means that a method of population-based multipoint simultaneous exploration such as GA is essential for the present topology optimization problem. Considering minimizing pressure loss of a flow channel as the second objective function, the proposed method is applied to a multi-objective optimization problem. As a result, we confirm that the proposed representation method enables to represent flow channels that balance both objective functions and GA captures the trade-off between two objective functions.*

1 INTRODUCTION

Shape optimization has been attracting much attention in flow problems, which defines the boundary between fluid and solid regions. However, shape optimization cannot deal with the change of topology, *e.g.*, making new holes into an object. Topology optimization is the most flexible optimization method, which can modify connectivity of an object independently of its predefined topology in contrast to sizing optimization and shape optimization that keep their topology during the optimization.

Topology optimization has been applied to a variety of engineering optimization problems [1] such as structural mechanics problems [2], heat transfer problems [3], and acoustic problems [4] since Bendsøe and Kikuchi first proposed the so-called homogenization design method [5]. Although the topology optimization method for structural design has matured in the past few decades, the application to flow problems is still limited due to the non-linearity of the governing equations. The first application to flow problems was later than the aforementioned applications. It was performed for the Stokes flow by Borrvall and Petersson [6].

The basic concept of topology optimization is replacement of the optimization problem with a material distribution problem in a fixed design domain using a characteristic function that indicates whether material exists or not. However, conventional topology optimization tends to suffer from numerical instabilities such as checkerboard pattern [7]. The level set method [8-11] is one of the approaches to avoid such instabilities. The level set method introduces a signed scalar function called level set function and distinguishes solid and fluid regions according to the sign of the function. Thus, zero-contours of the function indicate the boundaries of the regions.

Topology optimization often contains a large design space due to a high degree of freedom for shape and topology representation (i.e., the number of design variables is often equal to the number of elements in the finite element mesh). Thus, conventional topology optimization generally has employed the gradient-based method according to the sensitivity of an objective function to explore the optimal solution and has been developed based on this approach. However, the gradient-based method tends to get stuck to the local optima rather than the global optimum. On the other hand, Evolutionary Algorithm (EA) such as Genetic Algorithm (GA) is one of the metaheuristic optimization methods, which is more capable to explore the global optimum. However, GA requires numerous function evaluations to realize population-based multipoint simultaneous exploration. Thus, GA is not efficient to solve the optimization problems with expensive calculations and requires much expensive computational cost (i.e., large population and many generations) to obtain competitive solutions. In this case, surrogate models are effective to reduce computational cost required for function evaluation. This model approximates the response of each objective or constraint function to design variables in an algebraic expression. This model is derived from several sample points with real values of the objective or constraint function given by expensive numerical simulations. Thus, it can promptly give estimates of function values at arbitrary design variable values.

The authors have proposed a non-gradient-based topology optimization method for single and multi-objective flow problems using GA assisted by the Kriging model [12]. Despite the results showed agreement with the results in previous studies [13] and revealed a novel insight into topology optimization problems, the representation method with the level set function obtained by solving a partial differential equation was not able to represent complex shapes. This study proposes a novel level set representation method using another partial differential equation. To validate the proposed representation method in a global topology optimization method applied to flow problems, this research focuses on a single-objective

optimization problems to maximize heat transfer, and a multi-objective optimization problem to minimize pressure loss and to maximize heat transfer of flow channels.

2 COMPUTATIONAL FORMULATIONS

Figure 1 shows a flowchart of the optimization process in this study. Details of each step in the flowchart are described as follows in this section. In Section 2.1 and 2.2, the representation method of flow channels and the Building-Cube Method (BCM), which is a Cartesian-mesh CFD approach to evaluate objective functions are stated. Section 2.3 presents NSGA II, a genetic algorithm employed in this study. Section 2.4 presents the Kriging model, a surrogate-model, and a criterion to decide additional sample points of the Kriging model.

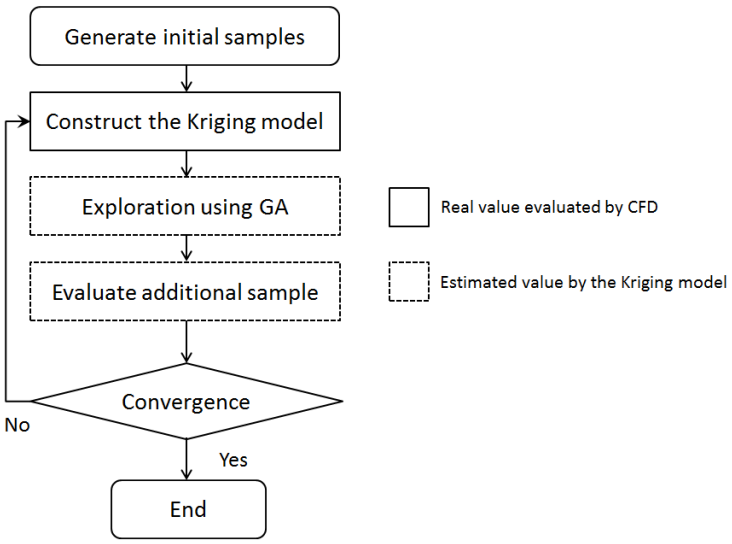


Figure 1: Optimization flowchart

2.1 Flow channel representation

In this section, a flow channel representation method proposed [12] and a novel method are stated. The boundaries between fluid and solid regions are represented by the level set representation that introduces a signed scalar function (level set function) $\phi(\boldsymbol{x})$ where $\boldsymbol{x}$ represents the location in the simulation domain. This research sets the range of $\phi(\boldsymbol{x})$ as $|\phi(\boldsymbol{x})| \leq 1$, and assumes $\boldsymbol{x}$ is in the fluid region if $\phi(\boldsymbol{x}) > 0$ or in the solid region if $\phi(\boldsymbol{x}) < 0$. A two-dimensional flow channel is described according to the sign of $\phi(\boldsymbol{x})$ given in the three-dimensional distribution. The example of the distribution of $\phi(\boldsymbol{x})$ is shown in Figure 2. As mentioned in Section 1, it is required to reduce the number of the design variables to employ GA for topology optimization. In this study, the distribution of $\phi(\boldsymbol{x})$ is obtained by several control points which are treated as the current design variables. Guirguis *et al.* employed a Kriging-interpolated level set (KLS) to obtain the distribution of $\phi(\boldsymbol{x})$ with a small number of the design variables and successfully applied GA to structural topology optimization [9]. The procedure to derive the distribution of $\phi(\boldsymbol{x})$ is stated below. First, at the outer boundary of the design domain, $\phi(\boldsymbol{x})$ is given as the step functions corresponding to the width of the inlet and outlet of the channel. Next, design variables ($\phi(\boldsymbol{x})$) are given at several discrete control points inside the design domain. Finally, the Laplace equation written in Eq. (1) is solved in the entire design domain to obtain the distribution of $\phi(\boldsymbol{x})$.

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0 \tag{1}$$

Note that the design variables are independent of the meshing size for CFD simulations in present method, which reduces the number of the design variables and enables the application of Genetic Algorithm for topology optimization problems.

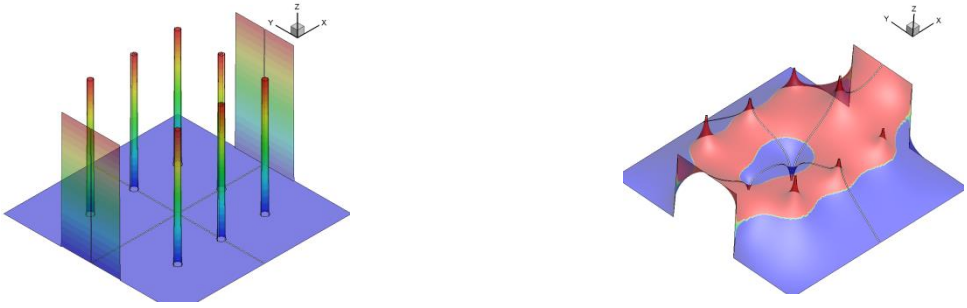
The distribution of $\phi(\boldsymbol{x})$ derived from solving the Laplace equation is steep and several shapes cannot be represented by any value of $\phi(\boldsymbol{x})$ given at the control points as the design variables. For example, the channels whose boundary is exactly straight cannot be represented. This issue must be solved to compare the proposed method with the previous study whose results include straight wall boundaries. Thus, it is required to develop an advanced method which can alleviate the steep distribution smoother.

Thus, in order to tackle this issue, we introduce another step to derive smooth level set distribution. Inspired by a filtering method in [14], the Helmholtz equation written in Eq. (2) is introduced.

$$-\nabla\left[R_x \frac{\partial \varphi}{\partial x}, R_y \frac{\partial \varphi}{\partial y}\right]^T + \varphi(\boldsymbol{x}) = \phi(\boldsymbol{x}) \tag{2}$$

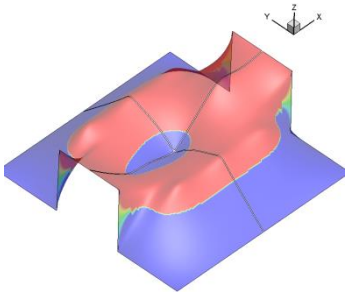
In Eq. (2), $\varphi(\boldsymbol{x})$ denotes a new distribution of level set function derived by solving the Helmholtz equation. $\phi(\boldsymbol{x})$, a solution of the Laplace equation, is substituted on the right hand side of Eq. (2). R_x and R_y are independent anisotropic weights of the Helmholtz equation that decide the alleviation strength of the level set distribution, $\phi(\boldsymbol{x})$. These weights are also involved as design variables and the present method enables to represent various shapes adaptively to the problems. Moreover, when both weights are set to 0, $\varphi(\boldsymbol{x})$ is identical to $\phi(\boldsymbol{x})$. This means that the present method provides a higher degree of freedom representation without sacrificing its original representation capability. It is important to note that although the flow channels evaluated by CFD simulation are represented based on the sign of $\varphi(\boldsymbol{x})$, the design variables considered in optimization are given at the control points in the $\phi(\boldsymbol{x})$ domain.

A conventional level set method for topology optimization [10, 11] employs the partial differential equation to update the value of level set function for a modified shape. The partial differential equation is solved with design sensitivities derived from the sensitivity analysis. In this study, on the other hand, the partial differential equation is not solved, and the value of $\phi(\boldsymbol{x})$ is updated based on the genetic algorithm stated in Section 2.3.

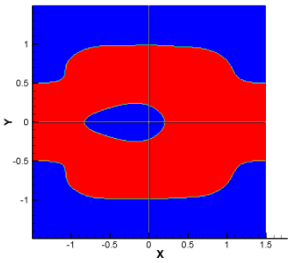


(a) Initial distribution (b) Final distribution

Figure 2: Distributions of the level set function obtained by solving the Laplace equation



(a) 3D final distribution



(b) 2D flow channel projected on the 2D design domain according to the sign of level set function of Figure 3 (a)

Figure 3: Distributions of the level set function obtained by solving theHelmholtz equation

2.2 Building-Cube Method

In order to evaluate the objective function of the channels, CFD simulations are conducted by the Building-Cube Method (BCM) [15], which is a Cartesian-mesh CFD approach. The governing equations of BCM are the 2D incompressible Navier-Stokes equations for unsteady state flow, and the 2D energy equation for unsteady state heat transfer. The convection terms are evaluated by a third-order upwind differencing [16], and the viscous terms are evaluated by a second-order central differencing. Time integration is conducted by the Crank-Nicolson method for the viscous terms and the Adams-Bashforth scheme for the convective terms, and the coupling of velocity and pressure is conducted by a fractional step method [17].

Since BCM employs a Cartesian-mesh CFD approach, it is easy to handle the complicated shapes of flow channels with topological change. However, in the Cartesian-mesh CFD approach, the object surface is represented by a staircase pattern, instead of smooth surface. For high accuracy computation, the Immersed Boundary Method (IBM) [18] using ghost cell and image point is employed at the wall boundary.

2.3 Genetic Algorithm

The genetic algorithm mimics the evolution of organisms, which selects individuals from the current generation as parents, generates new individuals as children by the crossover and mutation of the parents, and inherits better individuals to the next generation. In this study, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [19] proposed by Deb *et al.* is employed for exploration because this algorithm is effective and widespread employed for many optimization problems [9, 20]. Initially, a parent population $P_{t=1}$ with the size of N is created randomly. Here, t indicates the number of generation. Each feasible solution is assigned a rank (the solution with lower rank is better) according to its objective function value. On the other hand, each infeasible solution is assigned a rank which is higher than the minimum rank for the feasible solutions. Between two infeasible solutions, the solution with a smaller constraint violation has a better rank. Then, after choosing N solutions with lower rank in the parent population, recombination and mutation are conducted to create an offspring population Q_t with the size of N. In order to introduce elitism, first, a combined population $R_t = P_t \cup Q_t$ with the size of $2N$ is formed. Then, the solutions in R_t are sorted according to the ranks based on objective function values and constraint violation. Now, N solutions are chosen from R_t in the order of their ranks and make up a new population P_{t+1}. The procedure as described above

is for one generation. The non-dominated solutions with the lowest rank are explored by repeating this procedure for a certain number of generations.

2.4 Kriging model

Although GA is capable of finding the global optimum, it requires numerous function evaluations to realize population-based multipoint simultaneous exploration. For efficient global optimization, the Kriging surrogate model [21] is employed together with GA. The Kriging model is based on Bayesian statistics, and can adapt well to nonlinear functions. In addition, the Kriging model estimates not only the function values themselves but also their uncertainties. Based on these uncertainties, the expected improvement (EI) of an objective function, which may be achieved a new global optimum on the Kriging model by adding a new sample point, is estimated. In a single-objective optimization problem where $y(\boldsymbol{x})$ is minimized, the improvement value $I(\boldsymbol{x})$ and its expected value $E[I(\boldsymbol{x})]$ are defined as Eq. (3) and (4), respectively.

$$I(\boldsymbol{x}) = \begin{cases} [y_{min} - y] & \text{if } y < y_{min} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$E[I(\boldsymbol{x})] = \int_{-\infty}^{y_{min}} (y_{min} - y)\phi(y)dy \tag{4}$$

where ϕ is the probability density function denoted by $N[\hat{y}(\boldsymbol{x}), s^2(\boldsymbol{x})]$. Here, $\hat{y}(\boldsymbol{x})$ is the estimation of $y(\boldsymbol{x})$ and $s^2(\boldsymbol{x})$ is the mean square error at point $\boldsymbol{x}$ indicating the uncertainty of the estimated value. Maximizing the EI instead of the original objective function itself, the location of an additional sample point is determined for updating the Kriging model. Adding new samples to the Kriging model based on EI iteratively, these samples are expected to reach the global optima under the uncertainty of the Kriging model. This procedure is called Efficient Global Optimization (EGO) proposed by Jones *et al.* and widely employed for optimization [22].

Since the present optimization is capable of topological change, the flow channels may often become unconnected depending on design variable values. Since such unconnected channels make it difficult to evaluate the objective function values, they should not be considered as additional sample points for the Kriging model. Although GA has been successfully applied to topology optimization in several structural problems [9, 23, 24], it has not been accepted widely due in part to difficulty to ensure structural connectivity during the optimization procedure [25]. In order to ensure structural connectivity, the original EI value of the objective function, Eq. (4), is multiplied by the probability that the objective function value may be below a certain threshold estimated on the Kriging model. This probability $P(y(\boldsymbol{x}) < a)$, where a is a threshold, is formulated in

$$P(y(\boldsymbol{x}) < a) = \frac{1}{s\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-\frac{(y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x}))^2}{2s^2(\boldsymbol{x})} \right) dy \tag{5}$$

Maximizing the product of Eq. (4) and Eq. (5), the location of an additional sample point is determined for searching the global optima while assuring the connectivity of flow channels under the uncertainty of the Kriging model.

3 OPTIMIZATION PROBLEMS OF MAXIMIZING HEAT TRANSFER

In this section, two cases of the single-objective optimization to maximize heat transfer are presented. In this case, in addition to the 2D incompressible Navier-Stokes equation, the 2D

energy equation is also solved to evaluate the temperature field. In all case stated below, the design domain is discretized as a 240×240 uniform Cartesian grid. At the inlets, velocity is set as the Dirichlet boundary condition, parabolic profile given by Eq. (6) with the reference velocity of 1, and pressure is set as the Neumann condition.

$$g_n = \bar{g}\left(1 - (\frac{2s}{w})^2\right) \quad -\frac{w}{2} \le s \le \frac{w}{2} \tag{6}$$

where $\bar{g}$ is the prescribed velocity at the center of the flow profile, w is the width of the flow profile corresponding to the length of the inlet and the outlet, and s is location within the flow profile.

At the outlets, on the other hand, velocity is set as the Neumann condition and pressure is also set as the Dirichlet condition (zero pressure). Furthermore, no-slip boundary condition is given to the solid-fluid interface.

Temperature is set to be 0 at the inlet, and given by the Neumann condition at the outlet. Furthermore the temperature at the solid-fluid interface is expressed by the third type boundary condition as

$$-k\left(\frac{dT}{dx}\right)_{wall} = h(T_\infty - T_w) \tag{7}$$

where T_∞ is the temperature of the solid set to be 1, T_w is the fluid temperature on the wall, k and h correspond to the thermal conductivity and the heat transfer coefficient, respectively. In this study, h is given by the Nusselt number: $Nu = hl/k$. Here, the reference length l is the width of inlet set to be 1 and k is also set to be 1. Thus, the heat transfer coefficient is equal to the Nusselt number. The objective function is to maximize the bulk mean temperature T_m at the outlet given by

$$T_m = \frac{\int_{s_o} uTds}{\int_{s_o} uds} \tag{8}$$

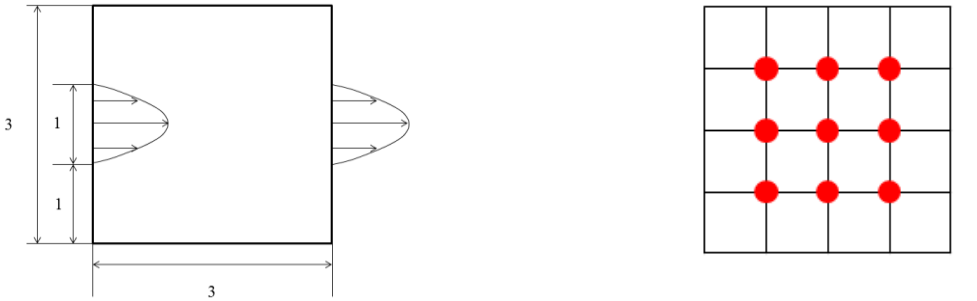
where S_o indicates the area of the outlet, u and T are the velocity and the temperature at the outlet, respectively. In this problem, since the possible range of the objective function value is known thermodynamically (i.e., the maximum is not greater than the wall temperature: 1 and the minimum is not less than the inlet temperature: 0). Thus, this case does not introduce a threshold into the objective function value as seen in Eq. (5) because it can reduce the diversity of the population in GA. However, in order to ensure connectivity during the optimization and explore the global optimum efficiently, the product of the original EI value of the objective function, Eq. (4), is multiplied by the probability written in Eq. (5) that the objective function value may be above a certain value is maximized by GA. In this problem, the objective function is multiplied by -1 and to be minimized. The threshold a in Eq. (5) is set to be 0, which is the largest value of the objective function.

3.1 Single pipe (Case 1)

3.1.1. Problem definition

First, a single pipe illustrated in Figure 4(a) is considered. The Reynolds number and the Nusselt number are set to be 5 and 50, respectively. The Prandtl number is set to be 6.78. Each dimensionless number is chosen to compare the present study with the previous study [12, 26]. The layout of the control points are illustrated as red points in Figure 4(b). The de-

sign variables are given at the control points vertically symmetric in the range of [-1, 1] and the range of the weights for the Helmholtz equation is set to [0, 0.4]. Thus, the number of the design variables is 8 in Case 1. Since too large weights make the flow channels discontinuous, the valid range of the weights keeping the flow channel connected are investigated beforehand.



(a) Design domain (b) Layout of control points

Figure 4: Geometry definition of a single pipe (Cases 1)

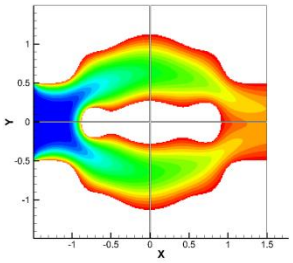
3.1.2. Results

In this case, 218 initial sample points satisfying the connectivity from the inlet to the outlet are used to construct the Kriging model. The Kriging model is updated 18 times. Figure 5 shows representative flow channels in the additional samples. In this case, several flow channels, each of which has similar objective function values as shown in Table 1 but with different topology, are found as the local optima. This result indicates that the objective function is a multi-modal function and it is required to employ a method of population-based multipoint simultaneous exploration such as GA. The result also indicates that, as the number of the solid islands in the channels increases, the size of each island becomes smaller. This is because the size of island is determined by the length of solid-fluid interface to achieve an equal amount of heat flux on the interface.

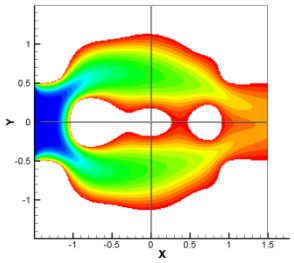
Matsumori *et al.* pointed out that the current optimization problem is non-convex and has many local optima [27]. The present method confirms the same issue and copes with that by GA. GA finds several solutions as a local optimum with the same topology of the results reported in previous study [12]. Intuitively, large islands should be put to enhance fluid-thermal interaction. However, flow separation does not occur easily since this case employs low Reynolds number, and basically any shapes of large islands contribute high performance of heat transfer. Thus, the Helmholtz weights are small (i.e. a flow channel represented by solving the Laplace equation is acceptable as local optima directly) and the global optimum shown in Figure 5 (b) is very similar to the global optimum shown in Figure 5 (d) and reported in [12].

	Temperature
(a) 11th sample	0.8351
(b) 13th sample	0.8512
(c) 18th sample	0.8426

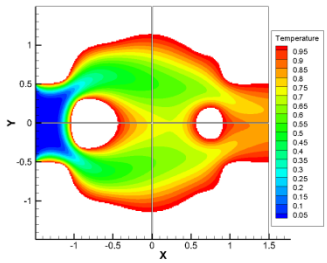
Table 1: Objective function values of the representative flow channels in the additional samples (Case 1).



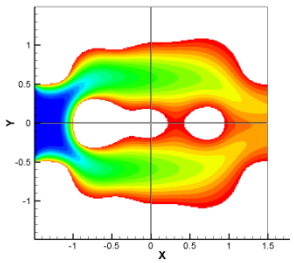
(a) 11th sample



(b) 13th sample (global optimum)



(c) 18th sample



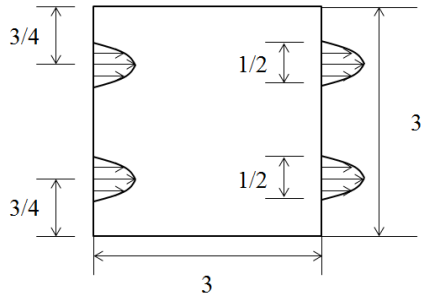
(d) The global optimum reported in [12]

Figure 5: Temperature distributions of the representative flow channels in the additional samples. (Case 1)

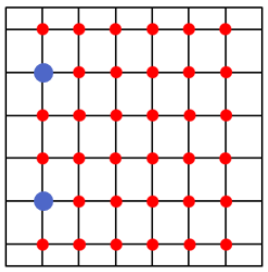
3.2 Double pipe (Case 2)

3.2.1. Problem definition

Second, a double pipe illustrated in Figure 6 (a) is considered. The Reynolds number and the Nusselt number are set to be 50 and 10, respectively. The layout of the control points are illustrated as red and blue points in Figure 6 (b). The design variables are given at the control points vertically symmetric in the ranges of [-1, 1] for red points and [0, 1] for blue points, and the range of the weights for the Helmholtz equation is set to [0, 0.4]. Thus, the number of the design variables is 20 in Case 2.



(a) Design domain



(b) Layout of control points

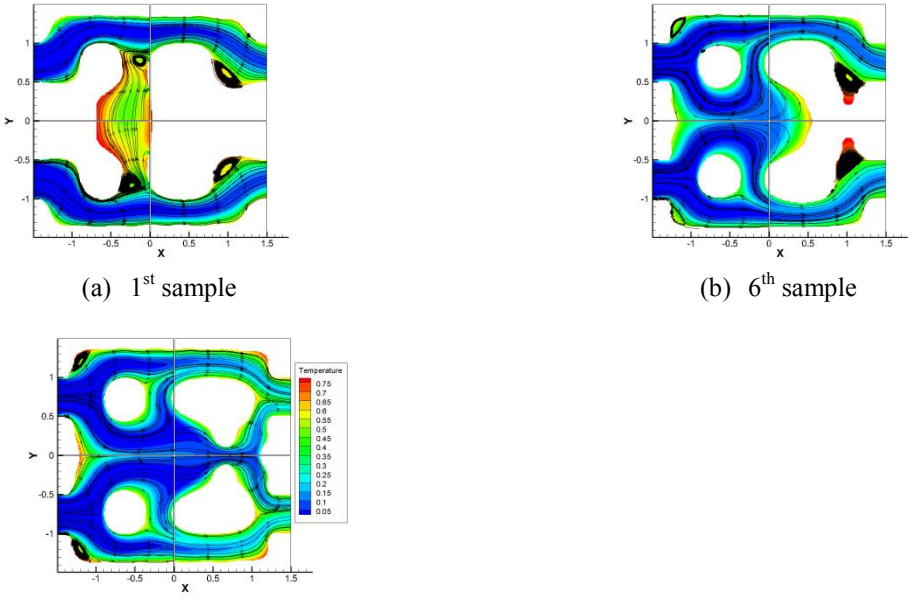
Figure 6: Geometry definition of a double pipe (Cases 2)

3.2.2. Results

In this case, 386 initial sample points satisfying the connectivity from the inlet to the outlet are used to construct the Kriging model. The Kriging model is updated 14 times. This case

has two differences from Case 1. First, since the Reynolds number in this case is higher than that in Case 1, flow separation occurs easily and several vortices are generated. Since these vortices affect heat transfer significantly, the shapes of islands inside the channels cannot be designed as arbitrarily as in Case 1. In Case 1, the flow channels with one or more islands have high temperature. In Case 2, on the other hand, putting islands arbitrarily does not always lead to high temperature. Second, since flow separation is a significant factor to increase temperature, the wall roughness and islands' shape are required to be smoothed by the Helmholtz weights whereas these weights are almost set to 0 in Case 1. It is important to note that the Helmholtz weights enable GA to explore efficiently because the design space is so complex that many details which do not contribute to the improvement of heat transfer are not removed during the optimization if the flow channels are represented only by the Laplace equation.

Figure 7 shows representative flow channels in the additional samples. Also in this case, several flow channels, each of which has similar objective function values but with different topology, are found as the local optima. However, in this case, as the number of islands increases, the objective function value also increases as shown in Table 2. In particular, the global optimum has much better objective function values compared with other local optima. As stated above, flow separation is a significant factor in a heat transfer problem, and we can confirm that the flow is not separated around the islands of the global optimum shown in Figure 7 (c). This result suggests that there is a dominant topology which shows high temperature. In this case, the present representation method always keeps the outer boundary of the flow channels smooth and lets GA explore efficient topology in the flow channels.



(a) 1st sample (b) 6th sample

(c) 11th sample (global optimum)

Figure 7: Temperature distributions of the representative flow channels in the additional samples. (Case 2)

	Temperature
(a) 1st sample	0.1913
(b) 6th sample	0.2423
(c) 11th sample	0.2798

Table 2: Objective function values of the representative flow channels in the additional samples (Case 2).

4 MULTI-OBJECTIVE OPTIMIZATION PROBLEM

Finally, two cases of multi-objective optimization problem to minimize pressure loss and maximize heat transfer of flow channels are considered. Minimizing pressure loss Δp written in Eq. (9) is considered as the second objective function.

$$\Delta p = \left(\frac{1}{2}\rho u^2 + p\right)_{in} - \left(\frac{1}{2}\rho u^2 + p\right)_{out} \tag{9}$$

Two cases employ the same geometry of the design domain, velocity, pressure, temperature boundary conditions, and mesh resolution as Cases 1 and 2. Cases 3 and 4 employ the same combinations of dimensionless numbers as Cases 1 and 2, respectively.

There are a number of non-dominated solutions, which are not worse than any other solution regarding all objective functions, in a multi-objective optimization problem whereas there is only one optimal solution in a single-objective optimization problem. Thus, it is important to ensure the diversity of the solutions in GA and capture the trade-off among objective functions. Hence, it should be careful to introduce a threshold of the pressure loss while keeping the diversity of the solutions in GA. However, without a threshold, initial sample points including huge pressure loss are used to construct the Kriging model, which may make the estimation accuracy of the Kriging model worse. Since the width of those flow channels is very small, these flow channels can be regarded as unconnected. Moreover, the bulk mean temperature of such unconnected or nearly unconnected flow channels is evaluated to be 0 in Cases 1 and 2. Thus, such solutions are hardly able to be the non-dominated solutions, and it will not lose the diversity of solutions even if such solutions are removed from initial sample points. Multi-objective optimization employs several additional sample points every time the Kriging model is updated whereas a single-objective optimization employs one additional sample point for every update. This study performs cluster analysis using the k-means method [28] to select representative sample points from many non-dominated solutions obtained by maximizing the EI value of each objective function on the Kriging model. In both cases stated below, 4 additional sample points are chosen for every update.

4.1 Single pipe (Case 3)

4.1.1. Problem definition

As the first example, a single pipe illustrated in Figure 4 (a) is considered. This case employs the same dimensionless numbers and the layout of control points as Case 1.

4.1.2. Results

In this case, 275 initial sample points satisfying the connectivity from the inlet to the outlet are used to construct the Kriging model. In this case, a threshold of the pressure loss is introduced and set to be 20. Since the value range of the objective function of pressure loss is so wide that the log transformed pressure loss is approximated by the Kriging model whereas the temperature is approximated directly. Figure 8 plots of the initial sample points (blue), the non-dominated solutions obtained after the 19th update of the Kriging model using the present method (red), and the non-dominated solutions obtained after the 19th update reported in [12] (green) in the objective space. The non-dominated solutions in the 19th update can be classified into 3 groups according to their characteristics.

First, in the yellow group, the flow channels have low pressure loss and low bulk mean temperature. These flow channels do not include any solid island inside the channels.

Second, all solutions in the blue group have one island and all but one solutions in the purple group have one island in the channel. Thus, the solutions in these two groups have the same characteristic with respect to their topology whereas the solutions reported in the previous study [12] can be classified based on their topology (i.e. the number of the solid islands in a channel). We classify the solutions in the blue and purple group based on the shape of the island in the channel. The island of the solutions in the blue group is a small round or streamlined. On the other hand, the island of the solutions in the purple group is a large round or blunt body. As illustrated in Figure 8, the channel with a streamlined island can improve its temperature whereas the pressure loss does not increase drastically compared with the channel with a round island. Four solutions in the purple group have a weak trade-off between two objective functions; they have almost the same bulk mean temperature to the optimum found in Case 1 whereas the pressure loss gets larger in proportion to the size of the island. This result indicates that the bulk mean temperature reaches its upper limit thermodynamically in these 4 solutions. This result indicates that the heat transfer is saturated and the pressure loss increases monotonically at this point, that is, the blue group is important from an engineering standpoint.

Since this case employs very low Reynolds number, the effect of the streamlined island which is expected to prevent flow separation does not appear clearly and the non-dominated front itself is not improved compared with that of the previous study [12]. However, the streamlined shapes show their capability to improve heat transfer and prevent the pressure loss getting worse drastically. Thus, the present method is promising to solve the problems with complex flow phenomena such as flow separation and indicates the possibility of designing a flow channel satisfying both high temperature and low pressure loss.

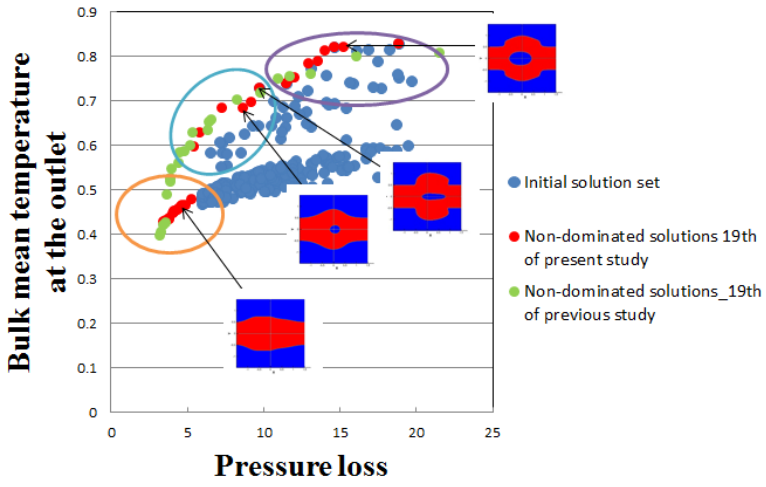


Figure 8: The solution set in the multi-objective problem (Case 3)

4.2 Double pipe (Case 4)

4.2.1. Problem definition

Finally, a multi-objective optimization problem of a double pipe illustrated in Figure 6 (a) is considered. This case employs the same dimensionless numbers and the layout of control points as Case 2.

4.2.2. Results

In this case, the initial sample points of the Kriging model are formed from 446 sample points satisfying the connectivity from the inlet to the outlet generated randomly and the global optimum of the single-objective optimization (Case 2). Hence, the number of the initial sample points is 447 in total. In this case, a threshold of the pressure loss is set to be 13. As well as Case 3, the log transformed pressure loss is approximated by the Kriging model whereas the temperature is approximated directly. Figure 9 plots the initial sample points (blue), the non-dominated solutions obtained after the 10th update of the Kriging model with the global optimum in the initial sample points (green), and the non-dominated solutions obtained after the 11th update without the global optimum in the initial sample points (red) in the objective space. As Figure 9 shows, the global optimum of the single-objective optimization facilitates GA to explore the non-dominated solutions that balance two objective functions.

In this case, low pressure loss can be achieved easily. As stated in Case 2, if the flow does not separate from an island put in a channel, the heat transfer performance does not worsen severely. It is obvious that the pressure loss gets larger if the flow separates in the channel. Thus, we anticipate that we can design a flow channel balancing both objective functions if the islands put in a channel can avoid flow separation. As shown in Figure 9, the red solutions (they are dominated solutions inherently since the true non-dominated solutions are green points) show improvement on heat transfer depending on the shape of the solid-fluid interface. Comparing two red solutions in Figure 9, we find that the backward shape of the left one is convex so that it can prevent flow separation whereas the backward of right one is flat. Thus, the left solution can balance two objective function whereas the pressure loss of the right one becomes larger despite its temperature is high since the solid-fluid interface is longer than that of the left one. However, as the green points in the blue circle indicate, two objective functions can be more improved if the backward is separated into two isolated islands. Flow separation does not occur in all of the green solutions and the values of the objective functions depend on the sizes of the backward islands; starting from ellipses (low pressure loss and low temperature), the backward islands become larger and transform into the same shape with the global optimum of the single-objective optimization in the channel with the highest temperature.

Thus, the proposed representation method enables to represent various shapes that contribute to improve both objective functions concurrently and GA reveals a dominant topology satisfying both objective functions.

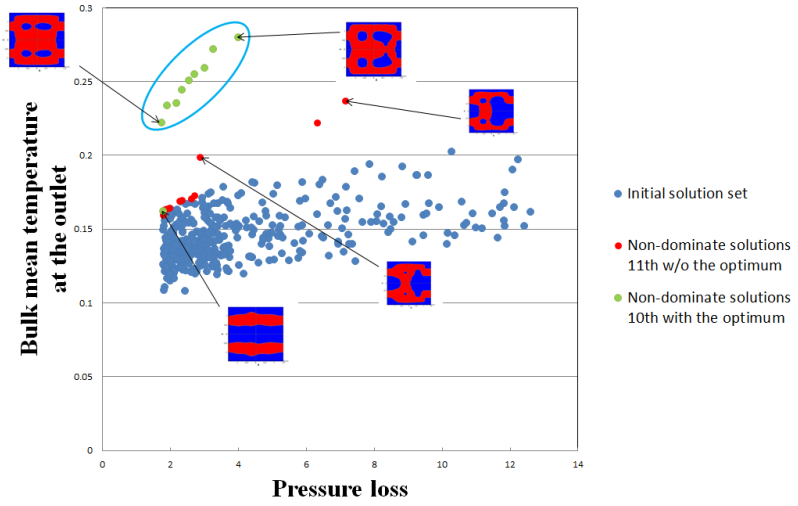


Figure 9: The solution set in the multi-objective problem (Case 4)

5 CONCLUSIONS

Topology optimization employing a Kriging-assisted GA was conducted in flow channel design problems that maximize heat transfer and/or minimize pressure loss using a novel level set representation approach.

In the single-objective optimization problems to maximize the bulk mean temperature, two cases with different layouts and dimensionless numbers were conducted. The GA found not only the optimal shape, but also several shapes that have quite similar objective function values but with different topologies from each other. In both cases, the objective function of temperature seems to be a multi-modal function and putting a solid island in a fluid region and increasing the length of the solid-fluid interface have significant effects to increase the temperature. However, in the case at the high Reynolds number, the flow separation severely affects heat transfer performance and the shape of islands should be considered so as not to provoke flow separation. The proposed representation method was able to represent various shapes which improve heat transfer and facilitated efficient exploration of GA in the complex design space.

Finally, considering minimizing pressure loss as the second objective function, multi-objective optimization problems were conducted in the same layout and numerical conditions with those of each case in the single-objective optimization problems. As a result, it was revealed that the size and the shape of a solid island put in a flow channel are the most important factors to increase the temperature and determine the pressure loss for both cases. Because the proposed representation method is able to represent streamlined shape, we could design flow channels with relatively large island without flow separation. Those flow channels with streamlined island were found as non-dominated solutions by GA. These solutions indicate that the proposed representation method is expected to help us design the flow channels that satisfy high temperature and low pressure loss at the same time.

Therefore, the proposed representation method and heuristic approach showed its capability to design promising flow channels and to explore global optima for both single-objective and multi-objective topology optimization in flow problems more efficiently.

REFERENCES

[1] M.P. Bendsøe, O. Sigmund, *Topology optimization: theory, methods and applications*, Springer, 2003.

[2] G.I.N. Rozvany, Aims, scope, methods, history and unified terminology of computer-aided topology optimization in structural mechanics. *Structural and Multidisciplinary Optimization*, **21**(2), 90-108, 2001.

[3] E.M. Dede, Multiphysics topology optimization of heat transfer and fluid flow systems. *proceedings of the COMSOL Users Conference* 2009.

[4] M.B. Dühring, J.S. Jensen, O. Sigmund, Acoustic design by topology optimization. *Journal of sound and vibration*, **317**(3), 557-575, 2008.

[5] M.P. Bendsøe, N. Kikuchi, Generating optimal topologies in structural design using a homogenization method. *Computer Methods in Applied Mechanics and Engineering*, **71**(2), 197-224, 1988.

[6] T. Borrvall, J. Petersson, Topology optimization of fluids in Stokes flow. *International Journal for Numerical Methods in Fluids*, **41**(1), 77-107 2003.

[7] O. Sigmund, J. Petersson, Numerical instabilities in topology optimization: a survey on procedures dealing with checkerboards, mesh-dependencies and local minima. *Structural optimization*, **16**(1), 68-75, 1998.

[8] J.A. Sethian, A. Wiegmann, Structural Boundary Design via Level Set and Immersed Interface Methods. *Journal of Computational Physics*, **163**(2), 489-528, 2000.

[9] D. Guirguis, K. Hamza, M. Aly, H. Hegazi, K. Saitou, Multi-objective topology optimization of multi-component continuum structures via a Kriging-interpolated level set approach. *Structural and Multidisciplinary Optimization*, **51**(3), 733-748, 2015.

[10] G. Allaire, F. Jouve, A.M. Toader, Structural optimization using sensitivity analysis and a level-set method. *Journal of Computational Physics*, **194**(1), 363-393, 2004.

[11] S. Yamasaki, T. Nomura, A. Kawamoto, K. Sato, S. Nishiwaki. A level set-based topology optimization method targeting metallic waveguide design problems. *International Journal for Numerical Methods in Engineering*, **87**(9), 844-868, 2011.

[12] M. Yoshimura, T. Misaka, K. Shimoyama, S. Obayashi, Topology optimization of flow channels with heat transfer using a genetic algorithm assisted by the Kriging model. *EUROGEN 2015*, Glasgow, UK, Sept 14-16, 2015.

[13] V.J. Challis, J.K. Guest, Level set topology optimization of fluids in Stokes flow. *International Journal for Numerical Methods in Engineering*, **79**(10), 1284-1308, 2009.

[14] B.S Lazarov, O. Sigmund, Filters in topology optimization based on Helmholtz-type differential equations. *International Journal for Numerical Methods in Engineering*, **86**(6), 765-781, 2011.

[15] K. Nakahashi, L.S. Kim. Building-Cube Method for Large-Scale, High Resolution Flow Computations. *AIAA paper*, 2004-0423, 2004.

[16] T. Kawamura, K. Kuwahara, Computation of High Reynolds number Flow around Circular Cylinder with Surface Roughness. *AIAA paper,* 84-0340, 1984.

[17] J. Kim, P. Moin, Application of a Fractional-Step Method to Incompressible Navier-Stokes Equations. *Journal of Computational Physics*, **59**, 308-323, 1985.

[18] R. Mittal, H. Dong, M. Bozkurttas, F.M. Najjar, A. Vargas, A. von Loebbecke. A versatile sharp interface immersed boundary method for incompressible flows with complex boundaries. *Journal of Computational Physics*, **227**(10), 4825-4852, 2008.

[19] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGAII. *Evolutionary Computation, IEEE Transactions*, **6**(2), 182–197, 2002.

[20] J.N. Richardson, F.C. Rajan, A. Sigrid, Robust topology optimization of truss structures with random loading and material properties: A multiobjective perspective. *Computers & Structures*, **154**, 41-47, 2015.

[21] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**(4), 455–492, 1998.

[22] F.A.C. Viana, T.W. Simpson, V. Balabonov, V. Toropov. Metamodeling in Multidisciplinary Design Optimization: How Far Have We Really Come?. *AIAA Journal*, **52**(4), 670-690, 2014.

[23] S.Y. Wang, K. Tai, M.Y. Wang, An enhanced genetic algorithm for structural topology optimization. *International Journal for Numerical Methods in Engineering*, **65**(1), 18-44, 2006.

[24] J.K. Guest, L.C. Smith Genut, Reducing dimensionality in topology optimization using adaptive design variable fields. *International Journal for Numerical Methods in Engineering*, **81**(8), 1019-1045, 2010.

[25] J.D. Deaton, R.V. Grandhi, A survey of structural and multidisciplinary continuum topology optimization: post 2000. *Structural and Multidisciplinary Optimization*, **49**(1), 1-38, 2014.

[26] T. Matsumori, T. Kondoh, A. Kawamoto, T. Nomura, Flow channel design in heat exchanger for maximum heat transfer by topology optimization. *Proceedings of 20th Design & Systems Conference*, 3103.1-4, 2010. (in Japanese)

[27] T. Matsumori, T. Kondoh, A. Kawamoto, T. Nomura, Topology optimization for fluid-thermal interaction problems under constant input power. *Structural and Multidisciplinary Optimization*, **47**(4), 571-581, 2013.

[28] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review. *ACM computing surveys (CSUR)*, **31**(3), 264-323, 1999.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOPOLOGY OPTIMIZATION OF CONTACT PROBLEMS BASED ON ALLEN-CAHN APPROACH

Andrzej M. Myśliński[1,2]

[1]Systems Research Institute
ul. Newelska 6, 01-447 Warsaw, Poland
e-mail: myslinsk@ibspan.waw.pl

[2] Warsaw University of Technology, Faculty of Manufacturing Engineering
ul. Narbutta 85, 02-524 Warsaw, Poland
e-mail: a.myslinski@wip.pw.edu.pl

Keywords: Topology optimization, Contact problems, Phase field regularization, Gradient flow.

Abstract. *The paper deals with a phase field model for formulation and solution of the topology optimization problems of bodies in unilateral contact consisting in the normal contact stress minimization. The contact problem with Tresca friction is governed by the system of elasticity equations with inequality type boundary conditions. The structural optimization problem consists in finding such material distribution within design domain to minimize the normal contact stress along the boundary of the body. The original structural optimization problem is reformulated in terms of material density function. Moreover the original cost functional is regularized using also surface and bulk energy terms. These terms allow to control global perimeter constraint and the occurence of the intermediate solution values. Using Lagrange multiplier approach the derivative of the regularized cost functional with respect to the control variable is calculated. The neccessary optimality condition is formulated in the form of Allen-Cahn gradient flow equation. The optimal topology is obtained as the steady state of the phase transition governed by this equation. This equation is discretized using finite difference and finite element methods. Numerical examples are provided.*

1 INTRODUCTION

The paper deals with the topology optimization for an elastic body in unilateral contact with a rigid foundation. This optimization problem consists in finding such topology of the domain occupied by the body and/or the shape of its boundary that the normal contact stress along the boundary of the body is minimized. Many successful methods have been proposed to analyze and to solve numerically topology optimization problems, including Simple Isotropic Material Penalization method and Evolutionary Structural Optimization approach, topology derivative method or different level set methods [1, 6, 7, 11, 19, 15].

In this paper phase field approach [4, 5, 6, 8, 17, 18] is proposed to regularize two phase topology optimization problem for unilateral elastic contact system and to solve it numerically. Material density function is a variable subject to optimization. This approach consists in using Ginzburg-Landau free energy term [8, 14, 17, 18, 19] as the regularization term rather than the perimeter constraint term. Although the proposed regularization for topology optimization of contact problems is more complicated than the perimeter one it has advantages comparing to the standard one. The derivative formula of the cost functional with respect to the material density function is calculated and is employed to formulate a necessary optimality condition for the topology optimization problem. This necessary optimality condition takes the form of the generalized Allen–Cahn equation rather than Cahn–Hilliard equation as in authors previous paper [10]. The derivative of the cost functional appears in the right hand side of these equation. Moreover the cost functional derivative is employed to calculate a descent direction in the numerical algorithm. Finite difference and finite element methods are used as the approximation methods. Implementation details are introduced. Numerical examples are provided and discussed.

2 PROBLEM FORMULATION

Consider deformations of an elastic body occupying two – dimensional domain Ω with the smooth boundary Γ (see Fig. 1). Assume $\Omega \subset D$ where D is a bounded smooth hold – all subset of R^2. The body is subject to body forces $f(x) = (f_1(x),\, f_2(x))$, $x \in \Omega$. Moreover, surface tractions $p(x) = (p_1(x), p_2(x))$, $x \in \Gamma$, are applied to a portion Γ_1 of the boundary Γ. We assume, that the body is clamped along the portion Γ_0 of the boundary Γ, and that the contact conditions are prescribed on the portion Γ_2, where $\Gamma_i \cap \Gamma_j = \emptyset$, $i \neq j,\ i, j = 0, 1, 2$, $\Gamma = \bar{\Gamma}_0 \cup \bar{\Gamma}_1 \cup \bar{\Gamma}_2$.

Let $\rho = \rho(x) : \Omega \to R$ denote the material density function at any generic point x in a design domain Ω. It is a phase field variable taking value close to 1 in the presence of material, while $\rho = 0$ corresponds to regions of Ω where the material is absent, i.e. there is a void. In the phase field approach the interface between material and void is described by a diffusive interfacial layer of a thickness proportional to a small lenght scale parameter $\epsilon > 0$ and at the interface the phase field ρ rapidly but smoothly changes its value [6]. We require that $0 \leq \rho \leq 1$. The ρ values outside this range do not seem to correspond to admissible material distributions. The elastic tensor $\mathcal{A}$ of the material body is assumed to be a function depending on density function ρ:

$$\mathcal{A} = g(\rho)\mathcal{A}_0, \quad \mathcal{A}_0 = \{a_{ijkl}\}_{i,j,k,l=1}^{2} \tag{1}$$

and $g(\rho)$ is a suitable chosen function [2, 4, 6, 15]. We denote by $u = (u_1, u_2)$, $u = u(x)$, $x \in \Omega$, the displacement of the body and by $\sigma(x) = \{\sigma_{ij}(u(x))\}$, $i, j = 1, 2$, the stress field in the body. Consider elastic bodies obeying Hooke's law, i.e., for $x \in \Omega$ and $i, j, k, l = 1, 2$

$$\sigma_{ij}(u(x)) = g(\rho)a_{ijkl}(x)e_{kl}(u(x)). \tag{2}$$

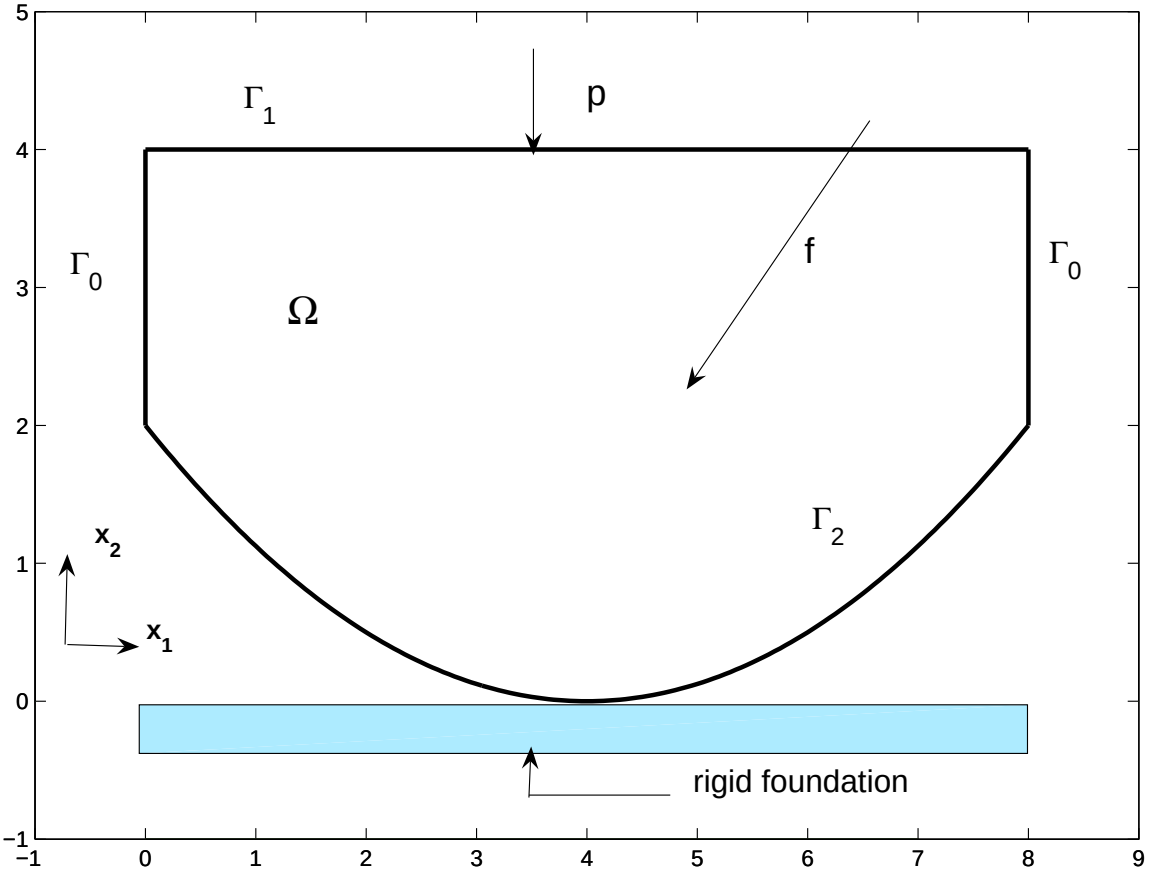


Figure 1: Initial domain Ω.

We use here and throughout the paper the summation convention over repeated indices [9]. The strain $e_{kl}(u(x))$, $k,l = 1,2$, is defined by:

$$e_{kl}(u(x)) = \frac{1}{2}(u_{k,l}(x) + u_{l,k}(x)), \tag{3}$$

where $u_{k,l}(x) = \frac{\partial u_k(x)}{\partial x_l}$. The stress field σ satisfies the system of equations in the domain Ω [9]

$$-\sigma_{ij}(x)_{,j} = f_i(x) \quad x \in \Omega, i,j = 1,2, \tag{4}$$

where $\sigma_{ij}(x)_{,j} = \frac{\partial \sigma_{ij}(x)}{\partial x_j}$, $i,j = 1,2$. The following boundary conditions are imposed on the boundary $\partial\Omega$

$$u_i(x) = 0 \quad \text{on} \quad \Gamma_0, \quad i = 1,2, \tag{5}$$

$$\sigma_{ij}(x)n_j = p_i \quad \text{on} \quad \Gamma_1, \quad i,j = 1,2, \tag{6}$$

$$u_N \leq 0, \quad \sigma_N \leq 0, \quad u_N\sigma_N = 0 \quad \text{on } \Gamma_2, \tag{7}$$

$$\mid \sigma_T \mid \leq 1, \quad u_T\sigma_T + \mid u_T \mid = 0 \quad \text{on } \Gamma_2, \tag{8}$$

where $n = (n_1, n_2)$ is the unit outward versor to the boundary Γ. Here $u_N = u_i n_i$ and $\sigma_N = \sigma_{ij}n_i n_j$, $i,j = 1,2$, represent the normal components of displacement u and stress σ, respectively. The tangential components of displacement u and stress σ are given by $(u_T)_i = u_i - u_N n_i$ and $(\sigma_T)_i = \sigma_{ij}n_j - \sigma_N n_i$, $i,j = 1,2$, respectively. $\mid u_T \mid$ denotes the Euclidean norm in R^2 of the tangent vector u_T.

2.1 Structural Optimization Problem

Before formulating a structural optimization problem for (4)-(8) let us introduce the set U_{ad} of admissible domains. This set has the form

$$U_{ad} = \{\Omega \ : \ E \subset \Omega \subset D \subset R^2 \ : \ \Omega \text{ is Lipschitz continuous,} \\ Vol(\Omega) - Vol^{giv} \leq 0, \ \ Vol(\Omega) = \int_\Omega \rho(x)dx.\}, \tag{9}$$

where $E \subset R^2$ is a given domain such that Ω as well as all perturbations of it satisfy $E \subset \Omega$. The constant $const_1 > 0$ is assumed to exist. The set U_{ad} is assumed to be nonempty. The constant $Vol^{giv} = const_0 > 0$ is given. For the shape optimization prpblem for system (4)-(8) the optimized domain Ω is assumed to satisfy equality volume condition, i.e., (9) is assumed to be satisfied as equality. In a case of topology optimization Vol^{giv} is assumed to be the initial domain volume and (9) is satisfied in the form $Vol(\Omega) = r_{fr}Vol^{giv}$ with $r_{fr} \in (0, 1)$ [15]. Recall from [11, 12] the cost functional approximating the normal contact stress on the contact boundary

$$J_\eta(u(\Omega)) = \int_{\Gamma_2} \sigma_N(u)\eta_N(x)ds, \tag{10}$$

depending on the auxiliary given bounded function $\eta(x) \in M^{st}$. The auxiliary set $M^{st} = \{\eta = (\eta_1, \eta_2) \in [H^1(D)]^2 : \eta_i \leq 0 \ \text{ on } D, \ i = 1, 2, \quad \| \ \eta \ \|_{[H^1(D)]^2} \leq \ 1\}$. Functions σ_N and η_N are the normal components of the stress field σ corresponding to a solution u satisfying system (4) - (8) and the function η, respectively. The cost functional 10) approximates the normal contact stress and is associated with the elastic energy functional [9]. Consider the following structural optimization problem: *for a given function $\eta \in M^{st}$, find a domain $\Omega^\star \in U_{ad}$ such that*

$$J_\eta(u(\Omega^\star)) = \min_{\Omega \in U_{ad}} J_\eta(u(\Omega)). \tag{11}$$

Adding to (9) a perimeter constraint $P_D(\Omega) \leq const_1$, where $P_D(\Omega) = \int_\Gamma dx$ is a perimeter of a domain Ω in D [5, 11, 16] and $const_1 > 0$ is a given constant the existence of an optimal domain $\Omega^\star \in U_{ad}$ to the problem (11) is ensured (see [4, 5, 16]).

3 PHASE FIELD BASED TOPOLOGY OPTIMIZATION PROBLEM

Let us introduce the regularized cost functional $J(\rho, u)$ in the form:

$$J(\rho, u) = J_\eta(u) + E(\rho), \tag{12}$$

where the functional $J_\eta(u)$ is given by (10). The Ginzburg-Landau free energy functional $E(\rho)$ is expressed as

$$E(\rho) = \int_\Omega \psi(\rho)d\Omega, \ \ \psi(\rho) = \frac{\gamma\epsilon}{2} \mid \nabla\rho \mid^2 + \frac{\gamma}{\epsilon}\psi_B(\rho), \tag{13}$$

where $\epsilon > 0$ is a constant, $\gamma > 0$ is a parameter related to the interfacial energy density. Function $\psi_B(\rho) = \rho^2(1 - \rho^2)$ is a double-well potential [10] which characterizes the two phases [2, 6]. The structural optimization problem (11) takes the form: *find $\rho^\star \in U^\rho_{ad}$ such that*

$$J(\rho^\star, u^\star) = \min_{\rho \in U^\rho_{ad}} J(\rho, u), \tag{14}$$

where $u^\star = u(\rho^\star)$ denotes a solution to the state system (4)-(8) depending on $\rho^\star$ and $U^\rho_{ad} = \{\rho : Vol(\Omega) \leq Vol^{giv}\}$ denotes the set of admissible material density functions.

In order to compute the first variation of the cost functional (12) we apply a formal Lagrangian approach combined with Allen–Cahn approach [2]. Let us introduce the Lagrangian $L(\rho) = L(\rho, u, \lambda, p^a, q^a, \mu)$:

$$L(\rho, u, \lambda, p^a, q^a, \mu) = J_\eta(u) + E(\rho) + \int_\Omega g(\rho) a_{ijkl} e_{ij}(u) e_{kl}(p^a) dx - \tag{15}$$
$$\int_\Omega f_i p_i^a dx - \int_{\Gamma_1} p_i p_i^a ds + \int_{\Gamma_2} \lambda p_T^a ds + \int_{\Gamma_2} q^a u_T ds + \mu(\int_\Omega \rho(x) dx - Vol^{giv}),$$

where $(p^a, q^a) \in K_1 \times \Lambda_1$ denotes an adjoint state defined as follows:

$$\int_\Omega g(\rho) a_{ijkl} e_{ij}(\eta + p^a) e_{kl}(\varphi) dx + \int_{\Gamma_2} q^a \varphi_T ds = 0 \quad \forall \varphi \in K_1, \tag{16}$$

and

$$\int_{\Gamma_2} \zeta(p_T^a + \eta_T) ds = 0 \ \ \forall \zeta \in \Lambda_1. \tag{17}$$

The sets K_1 and Λ_1 are given by

$$K_1 = \{\xi \in V_{sp} \ : \ \xi_N = 0 \ \text{ on } A^{st} \}, \tag{18}$$

$$\Lambda_1 = \{\zeta \in \Lambda \ : \ \zeta(x) = 0 \text{ on } B_1 \cup B_2 \cup B_1^+ \cup B_2^+\}, \tag{19}$$

while the coincidence set $A^{st} = \{x \in \Gamma_2 \ : \ u_N + v = 0\}$. Moreover $B_1 = \{x \in \Gamma_2 : \lambda(x) = -1\}$, $B_2 = \{x \in \Gamma_2 : \lambda(x) = +1\}$, $\tilde{B}_i = \{x \in B_i : u_N(x) + v = 0\}$, $i = 1, 2$, $B_i^+ = B_i \setminus \tilde{B}_i$, $i = 1, 2$. The derivative of the Lagrangian L with rescpect to ρ has the form:

$$\int_\Omega \frac{\partial J}{\partial \rho}(\rho, u)\zeta dx = \int_\Omega \frac{\partial L}{\partial \rho}(\rho, u, \lambda, p^a, q^a, \mu)\zeta dx = \int_\Omega [\gamma\epsilon \nabla\rho \cdot \nabla\zeta + \frac{\gamma}{\epsilon}\psi_B'(\rho)\zeta + \mu\zeta] dx +$$
$$\int_\Omega [g'(\rho) a_{ijkl} e_{ij}(u_\epsilon) e_{kl}(p^a + \eta) - f(p^a + \eta)]\zeta dx, \ \ \forall \zeta \in H^1(\Omega) \tag{20}$$

Using (20) we formulate a modified Allen-Cahn equation with constant mobility function as a gradient flow dynamic problem in an artificial time variable. It leads to a pseudo time stepping approach. This problem is as follows: ***find sufficiently regular*** $(\rho, u, \lambda, p^a, q^a, \mu)$ ***satisfying (4)-(8), (16)-(17) as well as***

$$\frac{\partial \rho}{\partial t} = \varphi_E(\rho) \qquad \text{in } \ \Omega, \ \forall t \in [0, T), \tag{21}$$
$$\nabla\rho \cdot n = 0 \qquad \text{on } \ \partial\Omega, \ \forall t \in [0, T), \tag{22}$$
$$\rho(0, x) = \rho_0(x) \qquad \text{in } \ \Omega, \ t = 0. \tag{23}$$

where the potential function φ_E is given by

$$\varphi_E = -\gamma\epsilon \triangle \rho + \frac{\gamma}{\epsilon}\psi_B'(\rho) + \mu -$$
$$g'(\rho) a_{ijkl} e_{ij}(u_\epsilon) e_{kl}(p^a + \eta) - f(p^a + \eta), \quad \text{a.e. in } \ \Omega. \tag{24}$$

The necessary optimality condition to optimization problem (14) has the form: if $(\rho^\star, u^\star, \lambda^\star, p^{a\star}, q^{a\star}, \mu^\star)$ is an optimal solution to structural optimization problem (14) than it satisfies (4)-(8), (16)-(17) and (21)-(24).

4 NUMERICAL RESULTS

The discretized structural optimization problem (14) is solved numerically. Time derivatives are approximated by the forward finite difference. Piecewise constant and piecewise linear finite element method is used as disretization method in space variables. The derivative of the double well potential is linearized with respect to ρ. Primal-dual active set method has been used to solve state and adjoint systems (4)-(8) and (16)-(17). Biconjugate gradient method has been used to solve (21)-(23). The algorithms are programmed in Matlab enviroment. As an example a body occupying 2D domain

$$\Omega = \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ 0 < v(x_1) \leq x_2 \leq 4\}, \tag{25}$$

is considered. The boundary Γ of the domain Ω is divided into three pieces

$$\begin{aligned} \Gamma_0 &= \{(x_1, x_2) \in R^2 : x_1 = 0, 8 \ \wedge \ 0 < v(x_1) \leq x_2 \leq 4\}, \\ \Gamma_1 &= \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ x_2 = 4\}, \\ \Gamma_2 &= \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ v(x_1) = x_2\}. \end{aligned} \tag{26}$$

The domain Ω and the boundary Γ_2 depend on the function v. The initial position of the boundary Γ_2 is given as in Fig. 1. The computations are carried out for the elastic body characterized by the Poisson's ratio $\nu = 0.29$, the Young modulus $E = 2.1 \cdot 10^{11} N/m^2$. The body is loaded by boundary traction $p_1 = 0$, $p_2 = -5.6 \cdot 10^6 N$ along Γ_1, body forces $f_i = 0$, $i = 1, 2$. Auxiliary function η is selected as piecewise constant (or linear) on D and is aproximated by a piecewise constant (or bilinear) functions. The computational domain $D = [0, 8] \times [0, 4]$ is selected. Domain D is discretized with a fixed rectangular mesh of 80 $\times$ 40. Other parameters are: $\epsilon = 0.02$, $\gamma = 1$, $T = 0.200$. Following [17] $g(\rho) = \frac{\rho}{1+\exp(-40\rho)} + g_\epsilon$, $g_\epsilon = 0.02$, $\psi_B(\rho) = \rho^2(1-\rho)^2(\frac{1}{10}\exp(15(1/2-\rho)^2) + 1)$.

Fig. 2 presents the optimal domain obtained by solving structural optimization problem (14) in the computational domain D using the optimality condition (21)-(24). The areas with low values of density function appear in the central part of the body and near the fixed edges. The obtained normal contact stress is almost constant along the optimal shape boundary and has been significantly reduced comparing to the initial one (see Fig. 3).

5 CONCLUSIONS

The structural optimization problem for elastic contact problem with the prescribed friction is solved numerically in the paper. Obtained numerical results indicate that the proposed numerical algorithm allows for significant improvements of the structure from one iteration to the next. Phase field approach based on the Allen-Cahn equation is flexible and can be easily combined with material density field. In this sense this approach follows SIMP method. On the other hand this approach can be also coupled with other physical fields allowing to consider different topology optimization problems.

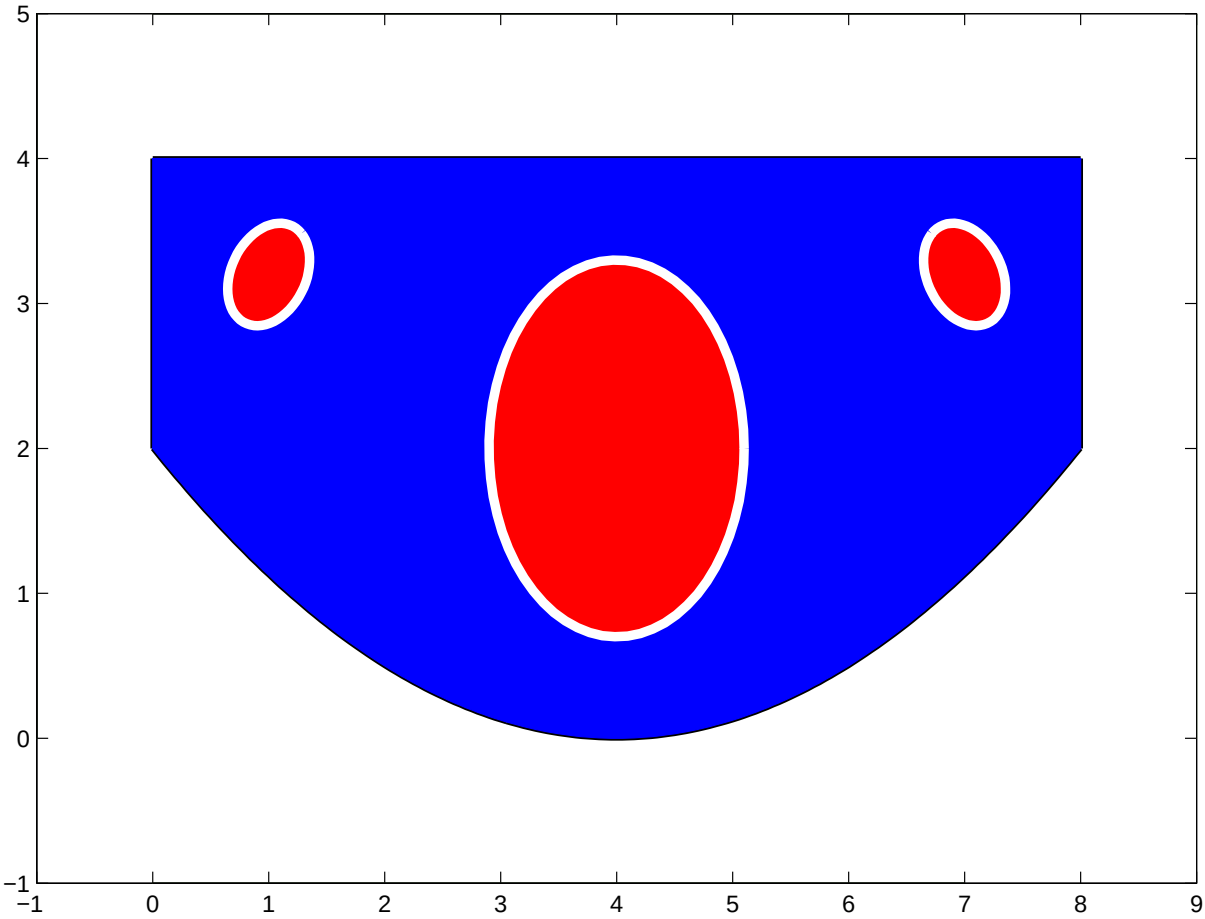


Figure 2: Optimal material density distribution in domain $\Omega^\star$.

REFERENCES

[1] F. Allaire, A. Jouve, A. Toader, Structural Optimization Using Sensitivity Analysis and a Level Set Method, *Journal of Computational Physics*, **194**, 363–393, 2004.

[2] L. Blank, M. Butz, H. Garcke, L. Sarbu, V. Styles, Allen-Cahn and Cahn-Hiliard variational inequalities solved with optimization techniques, *Constrained Optimization and Optimal Control for Partial Differential Equations*, G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, S. Ulbrich (Eds.), International Series of Numerical Mathematics, **160**, Birkhäuser, Basel, 21–35, 2012.

[3] L. Blank, H. Garcke, L. Sarbu, T. Srisupattarawanit, V. Styles, A. Voigt, Phase-field approaches to structural topology optimization, *Constrained Optimization and Opimal Control for Partial Differential Equations*, G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, S. Ulbrich (Eds.), International Series of Numerical Mathematics, **160**, Birkhäuser, Basel, 245–256, 2012.

[4] B. Bourdin, A. Chambolle, The phase-field method in optimal design, *IUTAM Symposium on Topological Design Optimization of Structures, Machines and Material*, M.P.

Bendsoe, N. Olhoff, and O. Sigmund (Eds.), Solid Mechanics and its Applications, Springer, 207–216, 2006.

[5] M. Burger, R. Stainko, Phase-field relaxation of topology optimization with local stress constraints, *SIAM J. Control. Optim.*, **45**, 1447-1466, 2006.

[6] L. Dede, M.J. Boroden, T.J.R. Hughes, Isogeometric analysis for topology optimization with a phase field model, *Archives of Computational Methods in Engineering*, **19**(3), 427–465, 2012.

[7] N.P. van Dijk, K. Maute, M. Langlaar, F. van Keulen, Level-set methods for structural topology optimization: a review, *Structural and Multidisciplinary Optimization*, **48**, 437–472, 2013.

[8] A. L. Gain, G. H Paulino, Phase-field based topology optimization with polygonal elements: a finite volume approach for the evolution equation, *Struct. Multidisc Optim.*, **46**, 327–342, 2012.

[9] J. Haslinger, R. Mäkinen, *Introduction to Shape Optimization. Theory, Approximation, and Computation*, SIAM Publications, Philadelphia, 2003.

[10] A. Myśliński, Phase Field Approach to Topology Optimization of Contact Problems, CD ROM Proceedings of the 10th World Congress on Structural and Multidisciplinary Optimization, May 19 - 24, 2013, Orlando, Florida, USA, Ed. R. Haftka, ISSMO, 2013.

[11] A. Myśliński, Level Set Method for Optimization of Contact Problems, *Engineering Analysis with Boundary Elements*, **32**, 986-994, 2008.

[12] A. Myśliński, Radial Basis Function Level Set Method for Structural Optimization, *Control and Cybernetics*, **39**(3), 627–645, 2010.

[13] A. Myśliński, Piecewise Constant Level Set Method for Topology Optimization of Unilateral Contact Problems, *Advances in Engineering Software*, **80**, 25–32, 2015.

[14] M. Scherzer, R. Denzer, P. Steinmann, A fictitious energy approach for shape optimization, *International Journal for Numerical Methods in Engineering*, **82**(3), 269–302, 2010.

[15] J. Sokołowski, A. Żochowski, On topological derivative in shape optimization, *Optimal Shape Design and Modelling*, T. Lewiński, O. Sigmund, J. Sokołowski, A. Żochowski (Eds.), Academic Printing House EXIT, Warsaw, Poland, 55–143, 2004.

[16] J. Sokołowski, J.P. Zolesio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer, Berlin, 1992.

[17] M. Wallin, M. Ristinmaa, H. Askfelt, Optimal topologies derived from a phase-field method, *Struct. Multidisc Optim.*, **45**, 171–183, 2012.

[18] M. Wallin, M. Ristinmaa, Boundary effects in a phase-field approach to topology optimization, *Computer Methods in Applied Mechanics and Engineering*, **278**, 145–159, 2014.

[19] T. Yamada, K. Izui, S. Nishiwaki, A. Takezawa, A Topology Optimization Method Based on the Level Set Method Incorporating a Fictitious Interface Energy, *Comput. Methods Appl. Mech. Engrg.*, **199**(45-48), 2876–2891, 2010.

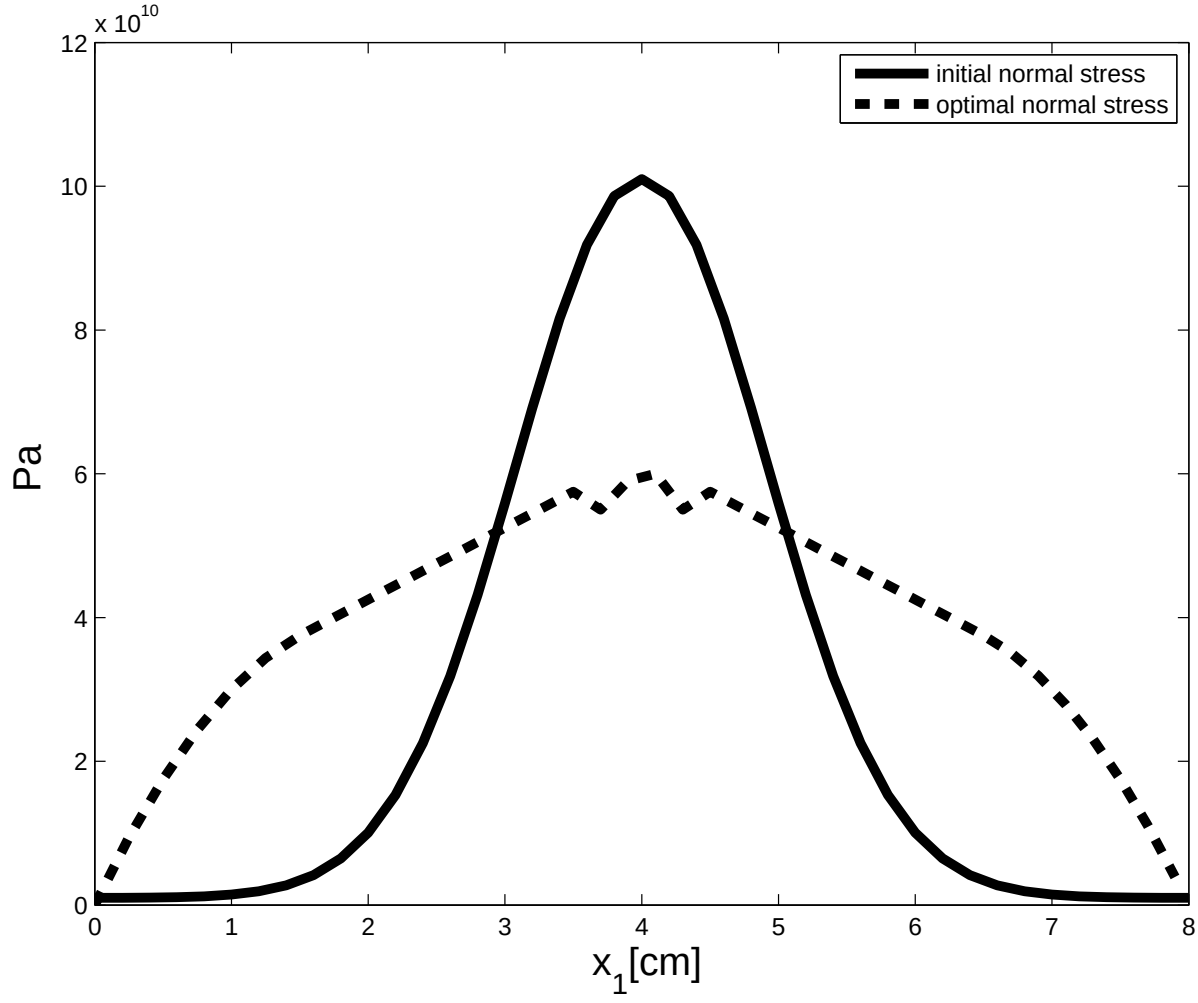


Figure 3: Initial and optimal normal contact stress.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON THE OPTIMAL DESIGN OF CABLE-STAYED BRIDGES

G. Carpentieri[1], M. Modano[2], F. Fabbrocino[3], L. Feo[1], and F. Fraternali[1]

[1] Department of Civil Engineering, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy, {gcarpentieri, l.feo}@unisa.it

[2] Department of Civil Engineering, "Federico II" University of Naples, Via Claudio 21, 80125 Naples, Italy, email: modano@unina.it

[3] Department of Engineering, Pegaso University, Piazza Trieste e Trento 48, 80132 Naples, Italy, email: francesco.fabbrocino@unipegaso.it

Keywords: Cable-Stayed Bridges, Stressing Sequence, Optimizations, Force Equilibrium Methods.

Abstract. *This paper studies cable-stayed bridges (CBSs), with special focus on the initial force distributions in cables during construction phases. An algorithm for the optimal design of the pre-tensioning sequence of cables is presented. A procedure for the optimization of cable forces is developed, according to a given objective function. Particular attention is given to the choice of the parameters to be optimized, and numeric examples are provided. The proposed method is employed to study the At Tannumah bridge in Basrah, Iraq; and is suitable for the optimization of the pre-tensioning sequence of arbitrary cable-stayed bridges.*
We show in the rest of the paper that an initial design (named S_0 configuration) that does not include any pre-tensioning forces in cables can lead to a highly non uniform bending moment distribution over the deck; which is not ideal for an optimal structure. Fort that reason we develop an "optimal design" (named S_d configuration), that corresponds to pre-tensioning forces inducing an "optimal" bending moment distribution over the deck.

1 INTRODUCTION

The construction of cable-stayed bridges is characterized by a series of phases in which geometry, boundaries, and loads vary significantly, causing changes in the state of stress [1-9]. The optimization of the construction process via the regulation of the initial forces in cables is important for the optimal control of the whole structural behavior [10-16]. One of the most common problems dealing with cable-stayed bridges concerns the computing of the initial cable forces and the pre-tensioning sequence, needed to obtain the designed configuration [17-20]. An optimal pre-tensioning sequence is useful for the control of the state of stress and strain during and after the construction phases. This question is faced in literature with several approaches [8,9,14], and is nowadays an open issue. In fact, there aren't closed form analytical solutions that allow the computing of the pre-tensioning sequence given a final design configuration. Only iterative algorithms are available [8,9,14], but such approaches require several cable tightening operations that cause technological, structural, and economic problems.

The absence of closed form solutions to this problem is due to the large number of parameters needed to characterize the cable stress distribution in CSBs. In fact, the structural behavior of CSBs depends on geometry, statics, material properties, construction process, and technology. The development of tools that allow the control of the behavior of these structures is therefore an open issue.

The present work deals with the formulation of procedure for the optimization of cable pre-tensioning forces that is suitable for any kind of CSBs. The proposed approach allows for the optimization of the bending moment distribution in the deck under suitable values of the pre-tensioning forces. We employ the "influence matrix method" [18] to compute the optimal pre-tensioning sequence that guarantee the achievement of the designed bending moment distribution (BMD) which is statically equivalent to another target distribution.

The procedure belongs to the "force equilibrium methods" [11], acting directly on the internal forces and indirectly on the elastic deformations. The proposed procedure is suitable for any construction sequence, and can be generalized to account for time-dependent phenomena [21].

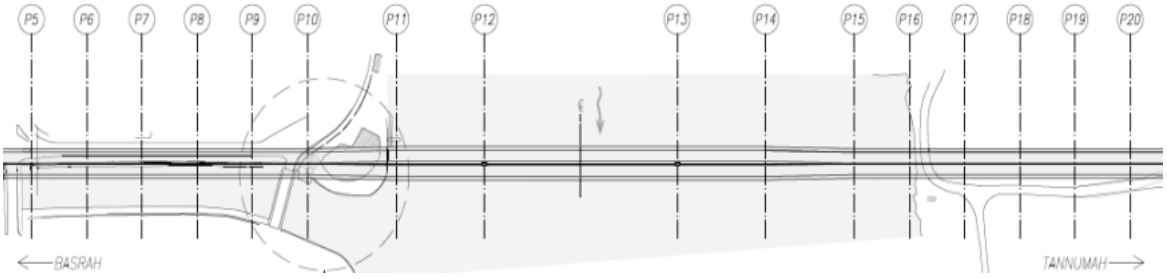
2 DESCRIPTION OF THE CASE STUDY

The At Tannumah Bridge (ATB) has been designed by Studio "De Miranda Associati", Milan, Italy and it is a part of a highway viaduct that connects the city center of Basra to the area of At Tannumah, Iraq, passing over the Shatt al Arab River.

The bridge concept include a semi-fan, central suspended span and self-supported cable-stayed system. In the longitudinal direction, the bridge is symmetric with respect to the mid-span, and it include two towers and two set of cables. The deck is made of three spans, the central one of 150 m and two lateral ones of 75 m each. The deck is obtained assembly semi-precast elements with length of 12.50 m each. The towers are made of pre-built sections of reinforced high-strength concrete (class C45/55 according to the Eurocode 2 [22]). The deck is made of steel welded beams (class S355 [22]) and reinforced concrete sections (a concrete slab of class C25/30 [22], 26 cm thick). The connection between the concrete slab and the steel beams is made through metallic bolts. The cables have a total diameter of 22 cm and are made of a set of 110 braided steel wires with diameter of 16 mm each.

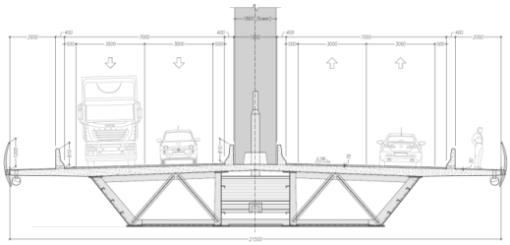
Fig. 1 shows different views of the ATB, while Fig. 2 illustrates the layout of the structural model that we employed to describe such a bridge. The elastic problem of the CSB model in Fig. 2 has been solved on assuming linear elastic response of all the bridge elements, through

the numerical algorithm that is detailed in [16]. We refer the reader to [16] for the mechanical and geometric properties of all the bridge elements.



(a)

(b)



(c)

Figure 1: The At Tannumah Bridge: (a) top view, (b) 3d view, (c) cross sectional [16] (courtesy of Studio "De Miranda Associati", Milan, Italy).

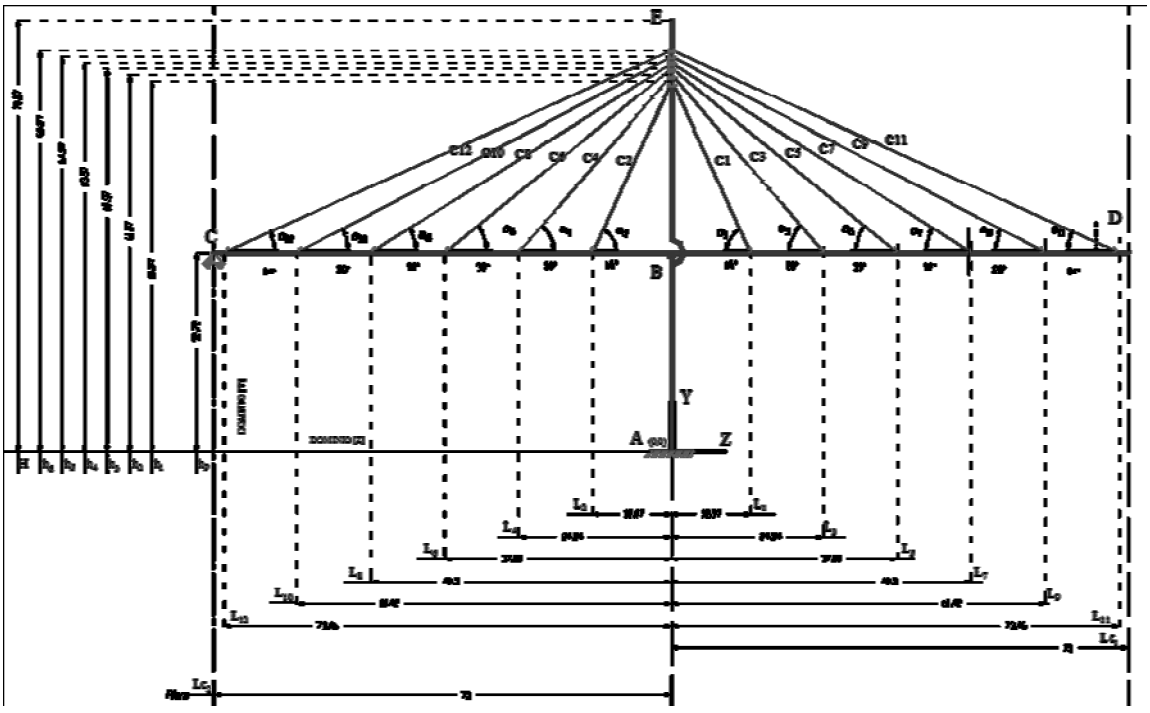


Figure 2: Bridge model.

Hereafter, we name "initial" (Fig. 3a) the bridge model S_0 that corresponds to assuming zero pre-tensioning forces in the cables (bridge unstressed under zero external loads). We shall see in the next section that such a model induces a highly non uniform moment distribution

over the deck, which is not ideal for an optimal use of the material (assuming uniform cross section of the deck along the span).

We instead name "optimal design" (Fig. 3b) the realization S_d of the bridge model in Fig. 2 that corresponds to pre-tensioning forces inducing a bending moment distribution over the deck identical to that of a low-stiffness deck model under zero pre-tensioning ("optimal" bending moment distribution). It is shown in [16] that such a bending moment distribution is nearly uniform over the span, determining an optimized use of the material composing the deck, which is assumed to have uniform cross-section. It is worth noting that the bridge model S_0, accounts for self-weight, external loads and the major dynamic effects (such as, e.g., wind and fluttering). The proposed procedure allows to obtain a target bending moment distribution, through the application of a self-equilibrated state of stress induced by an optimal cable pre-tensioning. As a matter of fact, our final goal is to play with the pre-tensioning forces of the cables of the model in Fig. 2, in order to achieve such the above, optimal bending moment distribution on the ATB.

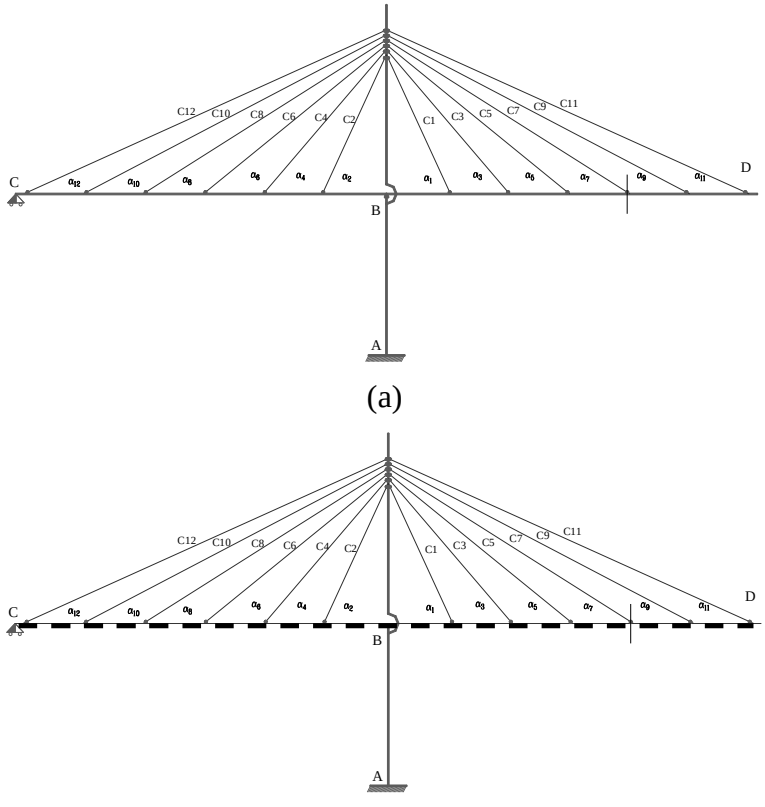


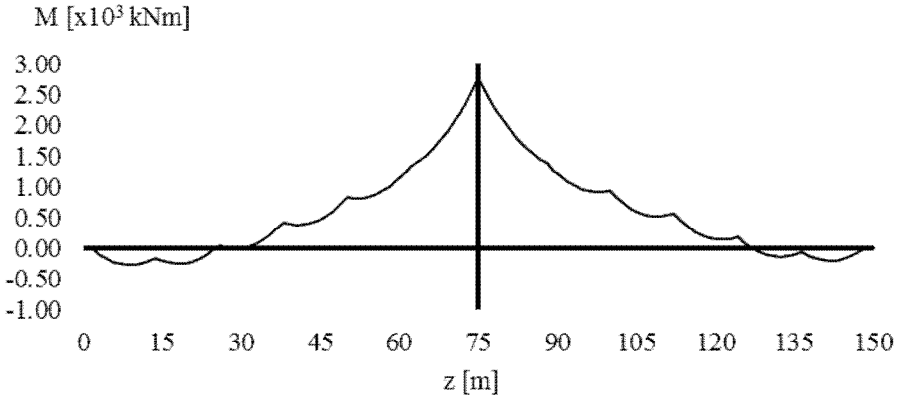
Figure 3: (a) initial bridge model S_0, (b) optimal bridge model S_d. The dashed thick line indicate a low stiffness deck.

3 NUMERICAL RESULTS FOR INITIAL AND OPTIMAL DESIGN MODELS

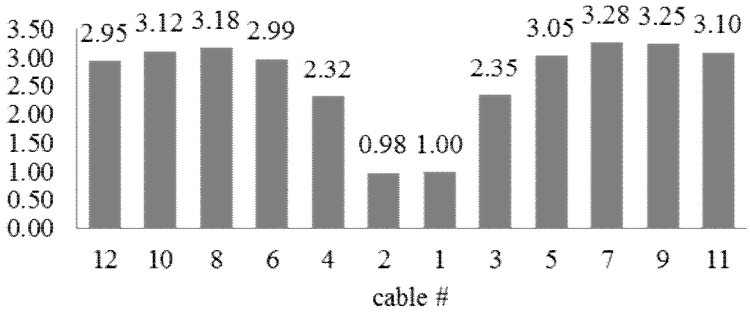
Figs. 4 and 5 show the bending moment and cable force distributions for the initial and optimal bridge models defined in the previous section. It is worth noting that the bending moment distribution of the initial model (M_0) is rather non uniform along the deck, featuring a pick value in correspondence of the deck-pier junction that is more than 5 times larger than the moment at the middle of the span (Fig. 4a). In the same model, the axial force carried by

the most stressed cable (cable # 7) is 3.28 times larger than the axial force carried by cable # 1 (Fig. 4b).

The bending moment distribution (M_d) and the cable force distribution of the optimal design model are shown in Fig. 5a-b. We observe that M_d shows a much more uniform distribution over the span as compared to M_0 (Fig. 4a). The cable force distribution also shows a more uniform profile as compared to that of the initial model, featuring a maximum cable force (in cable # 9) that is 1.96 times larger than the force carried by cable # 1.



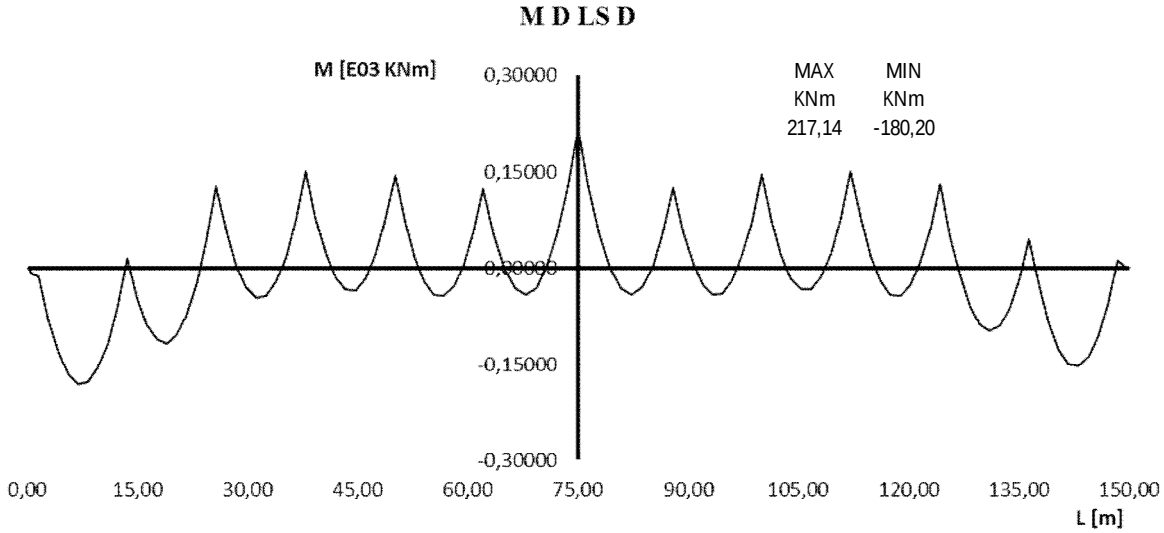
(a)



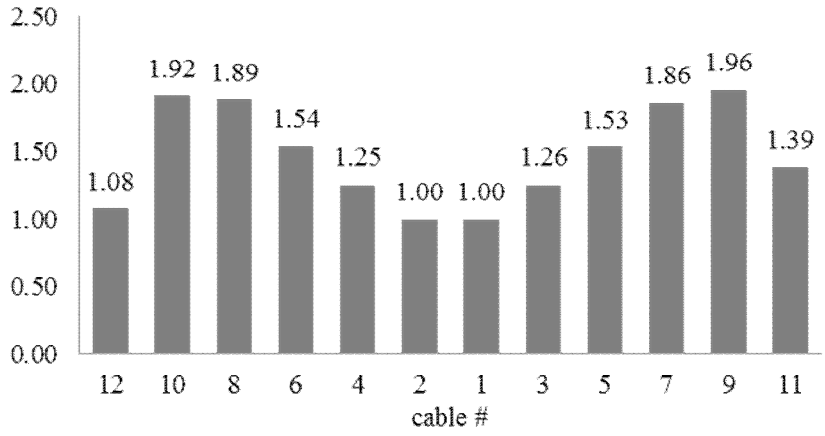
(b)

Figure 4: (a) bending moment and (b) normalized cable forces (divided by the axial force carried by cable # 1) for the initial model.

We recall that M_0 was computed on the real model of the ATB prescribing zero pretensioning forces in all cables. Such moment distribution resembles that of a cantilever beam under uniform transverse loading, on the initial portions of the two deck branches departing from the central pier (Fig. 4a). The M_d bending moment distribution was instead computed on a fictitious bridge model with a low stiffness deck [16]. Such moment distribution resembles that of a multi-support continuous beam under uniform transverse loading, always over the initial portions of the deck departing from the pier (Fig. 5a).



(a)



(b)

Figure 5: (a) bending moment and (b) normalized cable forces (divided by the axial force carried by cable # 1) for the optimal model.

4 OPTIMAL PRE-TENSIONING DESIGN

The present section is devoted to the formulation of a pre-tensioning design of the ATB model, which ensures that the bending moment distribution over the deck corresponds to M_d (Fig. 5a). Let $\boldsymbol{X} = \{x_1, \cdots x_n\}^T$ denote the vector collecting the pre-tensioning forces of all cables (n = 12). By repeatedly solving elastic problems of the bridge model in Fig. 2, we compute the axial force carried by the j-th cable when the i-th cable is subject to a unit axial force (S_i system). Let d_{ij} denote such a force coefficient and let **D** denote the n by n influence matrix collecting all such entries [16]. We are interested in solving the following linear problem [23]:

$$\mathbf{D}^T \mathbf{X} = \mathbf{\Delta T} \tag{1}$$

where $\boldsymbol{\Delta T} = \{\Delta t_1, \cdots \Delta t_n\}^T$ is the vector with current entry $\Delta t_i = t_{di} - t_{0i}$, t_{di} and t_{0i} respectively denoting the forces in the i-th cable in correspondence of the optimal design and in the initial models. The algebraic system of equations (1) is obtained by solving the n+2 elastic systems S_0, S_d, S_1, …, S_n. Its solution allows us to determine the pre-tensioning forces to be

applied to the different cables in order to achieve the target bending moment distribution M_d (Fig. 5a).

5 CONCLUSIONS

We have presented an approach to the optimal pre-tensioning design of cable-stayed bridges that is aimed at achieving a target bending moment distribution over the deck. Such a design approach allows the designers to recover from construction errors that could compromise the structural safety of the bridge.

The proposed methodology is based on the matrix of influence method, and relies on the determination of the cable forces on n+2 elastic models, n denoting the total number of cables. It can be applied in correspondence with any construction phase and can be easily generalized to account time dependent phenomena due, e.g., to material viscosity [21], and/or fracture damage [24]. Such a bridge design technique can also be applied to prestressed concrete structures, arch bridges with suspended decks, and all prestressable structures [25]-[31]. Another field of application of the influence matrix approach proposed in the present work regards the study of the optimal state of prestress of tensile reinforcements of existing masonry structures[32]-[35].

ACKNOLEDGEMENTS

The authors wish to acknowledge the precious support received by Maeg Costruzioni Spa – Treviso, Italy and Studio "De Miranda Associati", Milan - Italy during the course of present work.

REFERENCES

[1] M.P. Petrangeli, *Progettazione e costruzione di ponti con cenni di patologia e diagnostica delle opere esistenti*, Casa Editrice Ambrosiana, 1996.

[2] A. Raithel, *Costruzione di ponti*, Liguori Editore, 1963.

[3] Chen Wai-Fah, L. Duan, *Bridge Engineering Handbook*, CRC Press, Washington. 1999.

[4] V. Franciosi, *Scienza delle Costruzioni* Vol.III. Liguori Editore, 1965.

[5] N.J. Gimsing, *Cable supported bridges. Concepts and design*. John Wiley. 1997.

[6] M.B. Schlaich, Erection of cable-stayed bridges having composite decks with precast concrete slabs. *Journal of Bridge Engineering ASCE*. 6(5), 333-339. 2001.

[7] R. Walther, B. Houriet, W. Isler, P. Moia, J.F. Klein. *Cable Stayed Bridges*. Second edition. Thomas Teldford Publishing. 1999.

[8] P.H. Wang, T.C. Tseng, C.G. Yang, Initial shape of cable-stayed bridges. *Computers and Structures*. 46(6), 1095-1106. 1993.

[9] P.H. Wang, T.Y. Tang, H.N. Zheng, Analysis of cable-stayed bridges during construction by cantilever methods. *Computers and Structures*. 82, 329-346. 2004.

[10] P. Cassity, E.Y. He, Sequential stressing of cable-stayed bridges erected on falsework. *IABSE Conference on cable-supported bridges*. Seoul. 2001.

[11] D.W. Chen, F.T.K. Au, L.G. Tham, and P.K.K. Lee. Determination of initial cable force in prestressed concrete cable-stayed bridges for given design deck profiles using the force equilibrium method. *Computers & Structures*, 2000.

[12] F. Dischinger, Hangebrucken fur schwerste Verkehrslasten. Bauingenieur, 2(3,4):65–67, 107–113, 1949.

[13] H.J. Ernst, Der E-modul von Seilen unter Beruecksichtigung des Durchhanges.Der Bauingenieur, 40(2):52–55, 1965.

[14] D. Janjic, M. Pircher, H. Pircher, Optimization of Cable Tensioning in Cable-Stayed Bridges. *Journal of Bridge Engineering ASCE*. 8(3), 131-137, 2003.

[15] A. Recupero, Initial Cable-Forces Adjustments in Cable–Stayed Bridges: Effects of Parameters Uncertainties. *Advances in Structural Engineering and Mechanics ASEM'04*. Seoul, 2004.

[16] M.R. Maresca, *Studio delle sequenze di accordatura dei cavi nei ponti strallati. Il caso dell'At Tannumah Bridge di Bassora*, BsC thesis, University of Naples "Federico II", 2013.

[17] A.M.S. Freire, J.H.O. Negrão, and A.V. Lopes. Geometrical nonlinearities on the static analysis of highly flexible steel cable-stayed bridges. *Computers & Structures*, 2006.

[18] R. Xiao, L. Jia, X. Song, H. Xiang, Influence matrix method of cable tension optimization for long span cable-stayed bridges. *IABSE Conference on cable-supported bridges*. Seoul, 2001.

[19] Y.C. Loo, and G. Iseppi. Nonlinear Effects Of Cable Sag - A Case Study. *Cable-Stayed Bridges (Experience and Practice)*, 289–300, 1987.

[20] M.P. Petrangeli, The bridge over the Wadi Dib in Algeria. *L'Industria Italiana del Cemento*. LXIX (743), 362-379, 1999.

[21] M. Arici, M.F. Granata, A. Recupero, The influence of time-dependent phenomena in-segmental construction of concrete cable-stayed bridges. *Bridge Structures*. 7(4), 2011.

[22] European Committee for Standardization. *Eurocode 2: Design of concrete structures - Part 1-2: General rules*, EN 1992-1-1:2004, Brussels, Belgium, 2014.

[23] A. Luciano, Lomonaco. Un'introduzione all'algebra lineare. *Aracne editrice*, 2005.

[24] F. Fraternali, Free discontinuity finite element models in two-dimensions for in-plane crack problems, *Theoretical and Applied Fracture Mechanics*, 47 (3), 274-282, 2007.

[25] G. Carpentieri, R.E. Skelton, F. Fraternali, Minimum mass and optimal complexity of planar tensegrity bridges. *International Journal of Space Structures*, 30(3-4), 221-244, ISSN: 0266-3511, DOI: http://dx.doi.org/10.1260/0266-3511.30.3-4.221, 2015.

[26] F. Fraternali, G. Carpentieri, R.E. Skelton, A. Micheletti, Progettazione parametrica di ponti tensegrity (Parametric design of tensegrity bridges, in Italian). *STRUCTURAL*, 201, paper 02 (11pp). ISSN 2282-3794. DOI: 10.12917/Stru201.02, 2016.

[27] G. Carpentieri, M. Modano, F. Fabbrocino, F. Fraternali, Optimal design and dynamics of truss bridges, *COMPDYN 2015 - 5th ECCOMAS* Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering, 1731-1740, 2015.

[28] F. Fraternali, G. Carpentieri, A. Amendola, R.E. Skelton, V.F. Nesterenko, Multiscale tunability of solitary wave dynamics in tensegrity metamaterials, *Applied Physics Letters*, 105 (20), art. no. 201903, 2014.

[29] A. Bossio, F. Fabbrocino, G.P. Lignola, A. Prota, G. Manfredi, Simplified model for strengthening design of beam-column internal joints in Reinforced Concrete Frames, *Polymers*, 7 (9), 1732-1754, 2015.

[30] A. Amendola, C.J. Smith, R. Goodall, F. Auricchio, L. Feo, G. Benzoni, F. Fraternali, Experimental response of additively manufactured metallic pentamode materials confined between stiffening plates. *Composite Structures*, 142, 254–262, 2016.

[31] R.E. Skelton, R. Montuori, V. Pecoraro, Globally stable minimal mass compressive tensegrity structures. *Composite Structures*, 141, 346–354, 2016

[32] F. Fraternali, I. Farina, C. Polzone, E. Pagliuca, L. Feo, On the use of R-PET strips for the reinforcement of cement mortars. *Composites. Part B, Engineering*, 46, 207-210, 2013.

[33] F. Fraternali, G. Carpentieri, M. Modano, F. Fabbrocino, R.E. Skelton, A tensegrity approach to the optimal reinforcement of masonry domes and vaults through fiber-reinforced composite materials. *Composite Structures*, 134, 247-254, 2015.

[34] F. Fabbrocino, I. Farina, V.P. Berardi, A.J.M. Ferreira, F. Fraternali, On the thrust surface of unreinforced and FRP-/FRCM-reinforced masonry domes. *Composites. Part B, Engineering*, 83, 297-305, 2015.

[35] I. Farina, F. Fabbrocino, G. Carpentieri, M. Modano, A. Amendola, R. Goodall, L. Feo, F. Fraternali, On the reinforcement of cement mortars through 3D printed polymeric and metallic fibers. *Composites. Part B, Engineering*, 90, 76-85, 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A MIXED-INTEGER LINEAR PROGRAMMING APPROACH FOR GLOBAL DISCRETE SIZE OPTIMIZATION OF FRAME STRUCTURES

R. Van Mellaert[1], K. Mela[2], T. Tiainen[2], M. Heinisuo[2], G. Lombaert[3], and M. Schevenels[1]

[1]Department of Architecture, Faculty of Engineering Science, KU Leuven, Belgium
Kasteelpark Arenberg 1 box 2431, 3001 Leuven
e-mail: roxane.vanmellaert@kuleuven.be

[2] Department of Civil Engineering, Tampere University of Technology, Finland

[3] Department of Civil Engineering, Faculty of Engineering Science, KU Leuven, Belgium

Keywords: Global optimization, discrete optimization, size optimization, frame structures, mixed-integer linear programs.

Abstract. *This paper proposes a method to solve discrete size optimization problems of frame structures to global optimality. Global optimality is guaranteed by reformulating the optimization problem as a mixed-integer linear program (MILP) and solving it with the branch-and-bound method. The presented mixed variable formulation extends the existing mixed variable formulation for size and topology optimization of truss structures. The MILP is obtained by adopting the simultaneous analysis and design approach. The variables consist of binary decision variables to select a profile section from the catalog, and state variables representing the member end forces. The equilibrium equations and member stiffness relations are included as constraints. The displacement and stress constraints are formulated such that for each member limit values are imposed at predefined locations along the member. The proposed method is applied to a three-bay three-story frame.*

1 Introduction

In structural optimization, the problem formulation has a critical role as it imposes restrictions on the methods that can be used to solve the problem and it determines the applicability of the solution for the structural designer. If the principles of design, fabrication and economy [1] are followed, the problem formulation should contain as many requirements of these aspects as possible such that the need for post-processing the solution is minimized. On the other hand, incorporating all the requirements of the design codes, manufacturing practices and the economy of the structure into the problem as well-behaving (e.g. smooth) functions is very difficult in general, and enforcing them tend to make the problem intractable for most solution methods. Thus, it is fair to say that formulating an optimization problem is a compromise between meeting the needs of the structural designer and the capabilities of contemporary solution procedures.

For practical optimization of frame structures, the member profiles have to be chosen from a catalog of commercially available alternatives. When this feature is coupled with conventional formulations based on elastic structural analysis, the problem is not only nonlinear [2], but it also contains discrete design variables. The resulting mixed-integer nonlinear program (MINLP) can be treated by several optimization methods that have been proposed in the literature on discrete structural design of frames (for reviews, see [3, 4, 5, 6]). However, these methods have in common that they cannot guarantee that the global optimum is found.

In a detailed review, Arora [6] discusses various methods for structural optimization with discrete variables. These include branch-and-bound for nonlinear problems, sequential linearization, dynamic rounding-off, penalty approach and various stochastic methods, among others. Some of the more recent approaches include the discrete Lagrangian-based algorithm [7], and a scheme based on the optimality criteria method [8].

Currently, evolutionary algorithms are widely used for solving discrete frame optimization problems. These methods include simulated annealing [9], genetic algorithms [10, 11], ant colony optimization [12], firefly algorithm [13], artificial bee colony algorithm [14], and particle swarm optimization [15]. The general idea is to explore the design space in a random fashion, thereby using information collected from previous analyses to gradually move towards a better performing design. Evolutionary algorithms owe their popularity to the fact that they are easy to understand and to implement. They can cope with discrete parameters and are able to take into account complex constraints. However, evolutionary algorithms converge slowly, involve algorithmic parameters that require careful tuning, and global optimality cannot be guaranteed since no conclusive convergence checks can be made.

This paper presents a method for global discrete size optimization of frame structures. The adopted approach is to reformulate the optimization problem into a mixed-integer linear program (MILP). In the classical approach for structural optimization the nested analysis and design (NAND) approach is employed [16]: in every iteration a finite element analysis is performed in order to obtain the state variables (the structural nodal displacements and the member end forces). In order to facilitate the reformulation of the optimization problem as an MILP, the simultaneous analysis and design (SAND) approach is adopted [16] in this study: the state variables are considered as additional design variables and the state equations (the equilibrium equations and member stiffness relations) are enforced by means of additional constraints. In addition, a set of binary decision variables is introduced for each member of the structure to select a profile from the profile section catalog. The obtained MILP can be solved to global optimality with well-established algorithms such as branch-and-bound methods.

The MILP formulation approach has originally been proposed by Grossmann et al. [17] for discrete size optimization problems and is extended by Rasmussen et al. [18] for truss topology design problems. Mela [19] included member strength and buckling constraints specified by the Eurocode in the truss topology design problem. Van Mellaert et al. [20] included both the member and the joint constraints for size optimization of statically determinate trusses.

The formulation approach proposed in this paper differs from the formulation for truss optimization problems. In the formulation for truss structures, the variables consist of the binary decision variables, the displacement, and the normal forces. The internal resistance constraints are imposed by limiting the normal force of the member and displacement constraints are imposed by limiting the nodal displacements. For frame structures, the normal forces are replaced with the internal member end forces, including bending moments and shear forces. The proposed approach to take into account the member resistance and displacement constraints makes it possible to limit the relevant displacement components and stresses at several predefined locations of the members. These displacement and stress constraints are evaluated in a number of points using shape functions that interpolate the nodal displacements in order to calculate the stresses and displacements at the predefined locations of the members.

The paper is organized as follows. In section 2, the mixed-integer linear program of frame optimization problems is introduced: the design variables as well as the constraints are described. In section 3, the method is applied to a three-bay three-story frame.

2 Mixed variable formulation

In this section, the mixed integer linear program reformulation for frame structures is described. The design variables as well as the constraints of the optimization problem are introduced.

The proposed formulation is written for plane frames with prismatic members analysed by linear elastic theory. For simplicity, it is assumed that all members are made of the same material. Moreover, only a single loading condition is considered, but the formulation is easily extended to multiple loading conditions. The joints are assumed to be rigid, although hinged connections can be incorporated in the formulation as well.

Consider a frame structure defined by n_m members and n_n nodes with n_{dof} degrees of freedom. The number of profile alternatives is n_s. Denote by $\mathcal{M} = \{1, 2, \ldots, n_m\}$ and $\mathcal{C} = \{1, 2, \ldots, n_s\}$ the index sets for members and profiles. Each member may have its own set of profile alternatives. The index set of profiles of member i is denoted by $\mathcal{C}_i \subseteq \mathcal{C}$.

2.1 Design variables

The design variables include a vector with binary decision variables $\mathbf{y}$, a vector with continuous nodal displacement variables $\mathbf{u}$, and a vector with continuous force variables $\mathbf{q}$. The binary variables select a profile from the set of available alternatives. For each member i, profile j is selected when the corresponding variable $y_{ij} = 1$. Profile j is not selected for member i when the corresponding variable $y_{ij} = 0$. The force variables represent the member end forces. This means that for each member i and for each section j the force variables are represented by the normal end forces $N_{1,ij}$ and $N_{2,ij}$, the shear end forces $V_{1,ij}$ and $V_{2,ij}$, and the bending moment end forces $M_{1,ij}$ and $M_{2,ij}$ as shown in figure 1. The member end forces for each member i and for each section j are collected in the vector $\mathbf{q}_{ij}$:

$$\mathbf{q}_{ij} = \begin{bmatrix} N_{1,ij} & V_{1,ij} & M_{1,ij} & N_{2,ij} & V_{2,ij} & M_{2,ij} \end{bmatrix}^{\mathrm{T}} \tag{1}$$

Figure 1: Member end forces $\mathbf{q}_{ij}$.

The vector with the design variables $\mathbf{x}$ is given by:

$$\mathbf{x} = \begin{bmatrix} \mathbf{y}^{\mathrm{T}} & \mathbf{u}^{\mathrm{T}} & \mathbf{q}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \quad \mathbf{y} \in \mathbb{B}^{n_{\mathrm{b}}}, \mathbf{u} \in \mathbb{R}^{n_{\mathrm{dof}}}, \mathbf{q} \in \mathbb{R}^{6n_{\mathrm{b}}} \tag{2}$$

The total number of binary decision variables is denoted by $n_{\mathrm{b}} = \sum_{i=1}^{n_{\mathrm{m}}} n_{\mathrm{s}i}$, where n_{m} is the total number of members in the structure, and $n_{\mathrm{s}i}$ is the total number of available sections for member i. The total number of degrees of freedom is denoted by n_{dof}. The total number of force variables is $6n_{\mathrm{b}}$. The total number of design variables is calculated as $n_{\mathrm{dv}} = n_{\mathrm{dof}} + 7n_{\mathrm{b}}$.

2.2 MILP

The mixed-integer linear program for frame structures is given by equations (3) to (9):

$$\min_{\mathbf{x}} \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{C}_i} c_{ij} y_{ij} \tag{3}$$

such that

$$\sum_{j \in \mathcal{C}_i} y_{ij} = 1 \qquad \forall\, i \in \mathcal{M} \tag{4}$$

$$\sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{C}_i} \mathbf{L}_i^{\mathrm{T}} \mathbf{T}_i^{\mathrm{T}} \mathbf{q}_{ij} = \mathbf{f} \tag{5}$$

$$(1 - y_{ij}) \underline{\mathbf{q}}'_{ij} \leqslant \mathbf{K}_{ij} \mathbf{T}_i \mathbf{L}_i \mathbf{u} - \mathbf{q}_{ij} \leqslant (1 - y_{ij}) \bar{\mathbf{q}}'_{ij} \qquad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{6}$$

$$\underline{\mathbf{q}}'_{ij} y_{ij} \leqslant \mathbf{q}_{ij} \leqslant \bar{\mathbf{q}}'_{ij} y_{ij} \qquad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{7}$$

$$\underline{\mathbf{d}}'_{ij} + (\underline{\mathbf{d}}_{ij} - \underline{\mathbf{d}}'_{ij}) y_{ij} \leqslant \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \leqslant \bar{\mathbf{d}}'_{ij} + (\bar{\mathbf{d}}_{ij} - \bar{\mathbf{d}}'_{ij}) y_{ij} \qquad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{8}$$

$$\underline{\mathbf{s}}'_{ij} + (\underline{\mathbf{s}}_{ij} - \underline{\mathbf{s}}'_{ij}) y_{ij} \leqslant \mathbf{B}_{ij} \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{s}}_{ij} \leqslant \bar{\mathbf{s}}'_{ij} + (\bar{\mathbf{s}}_{ij} - \bar{\mathbf{s}}'_{ij}) y_{ij} \qquad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{9}$$

The objective function is given by equation (3). The subsequent equations are the constraints of the optimization problem. Equation (4) represents the profile selection. The equilibrium equations and the member stiffness relations are given by equation (5) and (6) respectively. Equation (7) ensures that the member forces are zero when profile j is not selected for member i. The displacements at predefined locations of member i are limited by the constraints given by equation (8). The stresses at predefined locations of member i are limited by the constraints given by equation (9). In the following subsections the details of deriving these equations are described.

It is possible to account for multiple load cases by extending the MILP with additional nodal displacement and force variables for each load case and additional constraints: the constraints given by equation (5) to (9) are repeated for each load case considering the appropriate nodal displacement and forces variables [19]. In this paper, only one load case is assumed.

2.3 Objective function

The objective function is given by equation (3):

$$\min_{\mathbf{x}} \quad \sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{C}_i} c_{ij}y_{ij}$$

where c_{ij} is the cost of profile j for member i. When the total weight of the structure is minimized, the cost of profile j for member i is defined as $c_{ij} = \rho L_i A_{ij}$, where ρ is the density of the material, L_i is the length of member i, and A_{ij} is the section area of section j for member i. In this paper a single material is used, but multiple materials can be adopted with the formulation.

2.4 Selection constraints

The constraint given by equation (4) ensures that a single section j is chosen from the catalog $\mathcal{C}_i$ for member i.

$$\sum_{j\in\mathcal{C}_i} y_{ij} = 1 \quad \forall\, i \in \mathcal{M}$$

2.5 Equilibrium equations

The nodal equilibrium is imposed by the equality constraints given by equation (5):

$$\sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{C}_i} \mathbf{L}_i^{\mathrm{T}}\mathbf{T}_i^{\mathrm{T}}\mathbf{q}_{ij} = \mathbf{f}$$

where $\mathbf{L}_i$ is a $6 \times n_{\text{dof}}$ binary location matrix that maps the element degrees of freedom to the system degrees of freedom, $\mathbf{T}_i$ is a 6×6 transformation matrix that accounts for the orientation of the element [21], and $\mathbf{f}$ is the $n_{\text{dof}} \times 1$ nodal load vector. Element loads are taken into account as equivalent nodal loads in the nodal load vector $\mathbf{f}$.

2.6 Member stiffness relations

In addition to nodal equilibrium, the material law and compatibility conditions are needed in structural analysis. For trusses, Hooke's law and compatibility conditions can be written as a single equation, because the normal force is the only stress resultant appearing in the members [18]. As frame members have three (6 in 3D) stress resultants in each node, altogether six (12 in 3D) force-displacement relations are needed. Thus, the relation between the member end forces and the nodal displacements can be written as:

$$\mathbf{q}_{ij} = \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u}y_{ij} \quad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{10}$$

where $\mathbf{K}_{ij}$ is the element stiffness matrix:

$$\mathbf{K}_{ij} = \begin{bmatrix} \frac{EA_{ij}}{L_i} & 0 & 0 & -\frac{EA_{ij}}{L_i} & 0 & 0 \\ 0 & \frac{12EI_{ij}}{L_i^3} & \frac{6EI_{ij}}{L_i^2} & 0 & -\frac{12EI_{ij}}{L_i^3} & \frac{6EI_{ij}}{L_i^2} \\ 0 & \frac{6EI_{ij}}{L_i^2} & \frac{4EI_{ij}}{L_i} & 0 & -\frac{6EI_{ij}}{L_i^2} & \frac{2EI_{ij}}{L_i} \\ -\frac{EA_{ij}}{L_i} & 0 & 0 & \frac{EA_{ij}}{L_i} & 0 & 0 \\ 0 & -\frac{12EI_{ij}}{L_i^3} & -\frac{6EI_{ij}}{L_i^2} & 0 & \frac{12EI_{ij}}{L_i^3} & -\frac{6EI_{ij}}{L_i^2} \\ 0 & \frac{6EI_{ij}}{L_i^2} & \frac{2EI_{ij}}{L_i} & 0 & -\frac{6EI_{ij}}{L_i^2} & \frac{4EI_{ij}}{L_i} \end{bmatrix}$$

where E is the Young's modulus of the material, L_i is the length of member i, and A_{ij} and I_{ij} are the section area and second moment of area of profile j for member i, respectively.

Equation (10) ensures that the force variables $\mathbf{q}_{ij}$ become zero when profile j is not selected for member i ($y_{ij} = 0$) and $\mathbf{q}_{ij} = \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u}$ when profile j is selected for member i ($y_{ij} = 1$).

In a regular finite element analysis, the global stiffness matrix $\mathbf{K}$ is assembled by replacing $\mathbf{q}_{ij}$ in equation (5) with the expression given by equation (10). The resulting equilibrium equation can not be reformulated as a linear system of equations in terms of the design variables since the global stiffness matrix depends on the binary decision variables. Therefore, the linear nodal equilibrium equation (5) and the member stiffness relation equation (10) are adopted as separate constraints.

The member stiffness relation in equation (10) is nonlinear in terms of the design variables but can be equivalently reformulated as a set of linear inequality constraints by introducing artificial upper and lower bounds (big-M) [18] as equation (6):

$$(1 - y_{ij})\underline{\mathbf{q}}'_{ij} \leqslant \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} - \mathbf{q}_{ij} \leqslant (1 - y_{ij})\bar{\mathbf{q}}'_{ij} \quad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i$$

In this equation, the force variables become equal to $\mathbf{q}_{ij} = \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u}$ when profile j is selected for member i ($y_{ij} = 1$). When profile j is not selected for member i ($y_{ij} = 0$), the force variables do not become zero but are bounded by $\mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} - \bar{\mathbf{q}}'_{ij} \leqslant \mathbf{q}_{ij} \leqslant \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} - \underline{\mathbf{q}}'_{ij}$. In order to ensure that the force variables become zero when profile j is not selected for member i, an additional constraint given by equation (7) is introduced:

$$\underline{\mathbf{q}}'_{ij} y_{ij} \leqslant \mathbf{q}_{ij} \leqslant \bar{\mathbf{q}}'_{ij} y_{ij} \quad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i$$

The artificial upper and lower bounds $\bar{\mathbf{q}}'_{ij}$ and $\underline{\mathbf{q}}'_{ij}$ ensure feasibility when profile j is not selected for member i and are calculated as follows [22]:

$$\begin{aligned} \underline{\mathbf{q}}'_{ij} = \min_{\mathbf{u}} \quad & \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} \\ \text{s.t.} \quad & \underline{\mathbf{u}} \leqslant \mathbf{u} \leqslant \overline{\mathbf{u}} \end{aligned} \tag{11}$$

$$\begin{aligned} \bar{\mathbf{q}}'_{ij} = \max_{\mathbf{u}} \quad & \mathbf{K}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} \\ \text{s.t.} \quad & \underline{\mathbf{u}} \leqslant \mathbf{u} \leqslant \overline{\mathbf{u}} \end{aligned} \tag{12}$$

where $\underline{\mathbf{u}}$ and $\overline{\mathbf{u}}$ are the prescribed minimum and maximum allowed nodal displacements, respectively. Note that equations (12) and (13) are linear optimization problems with bound constraints, that can be solved without effort [22].

2.7 Displacements along elements

For truss structures, the loads are applied to the nodes and the normal forces remain constant along the members. For each member, the internal resistance constraints can therefore be imposed by limiting the normal force of the member and displacement constraints can be imposed by limiting the nodal displacements. For frame structures, the internal forces may vary along the members and element loads (e.g. distributed loads) or bending moments at the end of the member can cause large deflections along the member. Consequently, the internal resistance and the displacements have to be checked at multiple locations along the member. In a standard finite element analysis, a possible strategy is to refine the finite element mesh by dividing the

members in multiple elements in order to obtain the forces and displacements at a certain location of the members. In the MILP reformulation the introduction of multiple elements not only leads to additional continuous force variables, but also to additional equilibrium equations and member stiffness relation constraints which have to be taken into account. The optimization problem thus becomes large when this approach is adopted.

In this subsection, an approach is proposed to limit the relevant displacement components at predefined locations of the members without the introduction of extra elements by refining the finite element mesh. These displacement constraints are imposed using shape functions that interpolate the nodal displacements in order to calculate the displacements at predefined locations of the members. In the following section, a similar approach is proposed in order to limit the relevant internal stresses at predefined locations of the members.

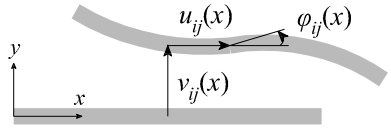


Figure 2: Local displacements

The displacements in the local coordinate system as defined in figure 2 at location x of member i are calculated as:

$$u_{ij}(x) = \mathbf{N}_i^u(x)\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{u}_{ij}(x) \tag{13}$$

$$v_{ij}(x) = \mathbf{N}_i^v(x)\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{v}_{ij}(x) \tag{14}$$

$$\varphi_{ij}(x) = \mathbf{N}_i^\varphi(x)\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{\varphi}_{ij}(x) \tag{15}$$

where the location x is defined along the local x-axis of the member and can have a minimum value of 0 representing the first end of member i, and a maximum value of L_i representing the second end of member i. $u_{ij}(x)$, $v_{ij}(x)$ and $\varphi_{ij}(x)$ are respectively the displacement along the local x-axis, the displacement along the local y-axis and the rotation at location x of member i for profile j as shown in figure 2, and $\tilde{u}_{ij}(x)$, $\tilde{v}_{ij}(x)$ and $\tilde{\varphi}_{ij}(x)$ are the displacements at location x of member i for profile j for the case of the beam with clamped-clamped boundary conditions subjected to the element loads . The shape function vectors $\mathbf{N}_i^u(x)$, $\mathbf{N}_i^v(x)$ and $\mathbf{N}_i^\varphi(x)$ are given by:

$$\mathbf{N}_i^u(x) = \left[\begin{array}{cccccc} 1-\frac{x}{L_i} & 0 & 0 & \frac{x}{L_i} & 0 & 0 \end{array}\right] \tag{16}$$

$$\mathbf{N}_i^v(x) = \left[\begin{array}{cccccc} 0 & \left[1-\frac{3x^2}{L_i^2}+\frac{2x^3}{L_i^3}\right] & \left[x\left(1-\frac{x}{L_i}\right)^2\right] & 0 & \left[\frac{3x^2}{L_i^2}-\frac{2x^3}{L_i^3}\right] & \left[x\left(-\frac{x}{L_i}+\frac{x^2}{L_i^2}\right)\right] \end{array}\right] \tag{17}$$

$$\mathbf{N}_i^\varphi(x) = \left[\begin{array}{cccccc} 0 & \left[\frac{-6x}{L_i^2}+\frac{6x^2}{L_i^3}\right] & \left[1-\frac{4x}{L_i}+\frac{3x^2}{L_i^2}\right] & 0 & \left[\frac{6x}{L_i^2}-\frac{6x^2}{L_i^3}\right] & \left[\frac{-2x}{L_i}+\frac{3x^2}{L_i^2}\right] \end{array}\right] \tag{18}$$

The relevant displacement components u_{ij}, v_{ij} and/or φ_{ij} of element i with section j at the output locations x_k of interest are collected in a vector $\mathbf{d}_{ij}$. For example, if the transverse displacement v_i of element i must be checked at the locations x_1, x_2, and x_3, the vector $\mathbf{d}_{ij}$ is given by:

$$\mathbf{d}_{ij} = \left[\begin{array}{c} v_{ij}(x_1) \\ v_{ij}(x_2) \\ v_{ij}(x_3) \end{array}\right] \tag{19}$$

This vector is obtained as follows:

$$\mathbf{d}_{ij} = \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \tag{20}$$

where $\mathbf{N}_i$ is in this case composed as:

$$\mathbf{N}_i = \begin{bmatrix} \mathbf{N}_i^v(x_1) \\ \mathbf{N}_i^v(x_2) \\ \mathbf{N}_i^v(x_3) \end{bmatrix} \tag{21}$$

and $\tilde{\mathbf{d}}_{ij}$ as:

$$\tilde{\mathbf{d}}_{ij} = \begin{bmatrix} \tilde{v}_{ij}(x_1) \\ \tilde{v}_{ij}(x_2) \\ \tilde{v}_{ij}(x_3) \end{bmatrix} \tag{22}$$

In order to limit the relevant displacement components at all predefined locations of member i the following constraints are introduced:

$$\underline{\mathbf{d}}'_{ij} + (\underline{\mathbf{d}}_{ij} - \underline{\mathbf{d}}'_{ij}) y_{ij} \leqslant \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \leqslant \bar{\mathbf{d}}'_{ij} + (\bar{\mathbf{d}}_{ij} - \bar{\mathbf{d}}'_{ij}) y_{ij} \quad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{23}$$

where $\underline{\mathbf{d}}_{ij}$ and $\bar{\mathbf{d}}_{ij}$ are the prescribed minimum and maximum allowed value of displacements, respectively, and $\underline{\mathbf{d}}'_{ij}$ and $\bar{\mathbf{d}}'_{ij}$ are artificial lower and upper bounds, respectively, in order to ensure feasibility when profile j is not selected for member i. When profile j is selected for member i ($y_{ij} = 1$) equation (23) becomes $\underline{\mathbf{d}}_{ij} \leqslant \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \leqslant \bar{\mathbf{d}}_{ij}$ and the relevant displacement components are limited. When profile j is not selected for member i ($y_{ij} = 0$) equation (23) reduces to $\underline{\mathbf{d}}'_{ij} \leqslant \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \leqslant \bar{\mathbf{d}}'_{ij}$. The lower bounds $\underline{\mathbf{d}}'_{ij}$ and upper bounds $\bar{\mathbf{d}}'_{ij}$ are calculated as:

$$\begin{aligned} \underline{\mathbf{d}}'_{ij} = \min_{\mathbf{u}} \quad & \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \\ \text{s.t.} \quad & \underline{\mathbf{u}} \leqslant \mathbf{u} \leqslant \overline{\mathbf{u}} \end{aligned} \tag{24}$$

$$\begin{aligned} \bar{\mathbf{d}}'_{ij} = \max_{\mathbf{u}} \quad & \mathbf{N}_i \mathbf{T}_i \mathbf{L}_i \mathbf{u} + \tilde{\mathbf{d}}_{ij} \\ \text{s.t.} \quad & \underline{\mathbf{u}} \leqslant \mathbf{u} \leqslant \overline{\mathbf{u}} \end{aligned} \tag{25}$$

where $\underline{\mathbf{u}}$ and $\overline{\mathbf{u}}$ are the minimum and maximum allowed nodal displacements, respectively.

When only nodal displacements are limited, the constraints given by equation (23) can be substituted by the following constraints:

$$\underline{\mathbf{u}} \leqslant \mathbf{u} \leqslant \overline{\mathbf{u}} \tag{26}$$

2.8 Stresses

The resistance of cross-sections subjected to shear forces, normal forces, and bending moments must be checked at predefined locations of the members. For elastic design, the resistance constraints can be written in terms of stresses as:

$$\sigma_{\text{min}} \leqslant \sigma_{\text{t},ij}(x) \leqslant \sigma_{\text{max}} \tag{27}$$

$$\sigma_{\text{min}} \leqslant \sigma_{\text{b},ij}(x) \leqslant \sigma_{\text{max}} \tag{28}$$

$$\tau_{\text{min}} \leqslant \tau_{ij}(x) \leqslant \tau_{\text{max}} \tag{29}$$

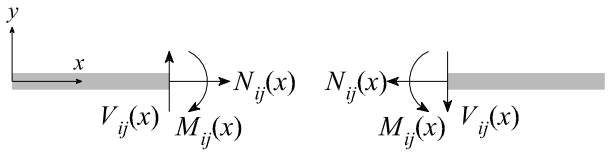


Figure 3: Internal forces of member i for profile j at location x

where $\sigma_{\mathrm{t},ij}(x)$, $\sigma_{\mathrm{b},ij}(x)$, and $\tau_{ij}(x)$ are the normal stress at the top, the normal stress at the bottom and the maximum shear stress of member i for profile j at location x. These stresses are calculated as

$$\sigma_{\mathrm{t},ij}(x) = \frac{N_{ij}(x)}{A_{ij}} + \frac{M_{ij}(x)}{W_{\mathrm{t},ij}} \tag{30}$$

$$\sigma_{\mathrm{b},ij}(x) = \frac{N_{ij}(x)}{A_{ij}} - \frac{M_{ij}(x)}{W_{\mathrm{b},ij}} \tag{31}$$

$$\tau_{ij}(x) = \frac{V_{ij}(x)S_{ij}}{I_{ij}b_{ij}} \tag{32}$$

where $W_{\mathrm{t},ij}$ is the section modulus at the top, $W_{\mathrm{b},ij}$ is the section modulus at the bottom, S_{ij} is the first moment of area, I_{ij} is the second moment of area, and b_{ij} is the width of the profile at the point where the maximum shear stress occurs. $N_{ij}(x)$, $V_{ij}(x)$, and $M_{ij}(x)$ represent, respectively, the normal force, shear force, and bending moment at location x as given by figure 3. For plastic design, similar constraints related to the stress resultants can be formulated.

The internal stresses at location x of member i for profile j are calculated as:

$$\sigma_{\mathrm{t},ij}(x) = \mathbf{B}_{ij}^{\sigma_\mathrm{t}}(x)\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{\sigma}_{\mathrm{t},ij}(x) \tag{33}$$

$$\sigma_{\mathrm{b},ij}(x) = \mathbf{B}_{ij}^{\sigma_\mathrm{b}}(x)\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{\sigma}_{\mathrm{b},ij}(x) \tag{34}$$

$$\tau_{ij}(x) = \mathbf{B}_{ij}^{\tau}(x)\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{\tau}_{ij}(x) \tag{35}$$

where $\tilde{\sigma}_{\mathrm{t},ij}(x)$, $\tilde{\sigma}_{\mathrm{b},ij}(x)$ and $\tilde{\tau}_{ij}(x)$ are the stresses of the beam with clamped-clamped boundary conditions subjected to the element loads at location x of member i for profile j. They are calculated as:

$$\tilde{\sigma}_{\mathrm{t},ij}(x) = \frac{\tilde{N}_i(x)}{A_{ij}} + \frac{\tilde{M}_i(x)}{W_{\mathrm{t},ij}} \tag{36}$$

$$\tilde{\sigma}_{\mathrm{b},ij}(x) = \frac{\tilde{N}_i(x)}{A_{ij}} - \frac{\tilde{M}_i(x)}{W_{\mathrm{b},ij}} \tag{37}$$

$$\tilde{\tau}_{ij}(x) = \frac{\tilde{V}_i(x)S_{ij}}{I_{ij}b_{ij}} \tag{38}$$

where $\tilde{N}_i(x)$, $\tilde{V}_i(x)$ and $\tilde{M}_i(x)$ are respectively the normal forces, shear forces and bending moments of the beam with clamped-clamped boundary conditions subjected to the element loads at location x of member i. Starting from the given displacements $\mathbf{u}$ the stresses can be

calculated using the vectors $\mathbf{B}_i^{\sigma_\mathrm{t}}(x)$, $\mathbf{B}_i^{\sigma_\mathrm{b}}(x)$ and $\mathbf{B}_i^{\tau}(x)$:

$$\mathbf{B}_{ij}^{\sigma_\mathrm{t}}(x) = \begin{bmatrix} -\frac{E}{L_i} & \frac{6EI_{ij}}{L_i^2 W_{\mathrm{t},ij}} - \frac{12xEI_{ij}}{L_i^3 W_{\mathrm{t},ij}} & \frac{4EI_{ij}}{L_i W_{\mathrm{t},ij}} - \frac{6xEI_{ij}}{L_i^2 W_{\mathrm{t},ij}} & \frac{E}{L_i} & -\frac{6EI_{ij}}{L_i^2 W_{\mathrm{t},ij}} + \frac{12xEI_{ij}}{L_i^3 W_{\mathrm{t},ij}} & \frac{2EI_{ij}}{L_i W_{\mathrm{t},ij}} - \frac{6xEI_{ij}}{L_i^2 W_{\mathrm{t},ij}} \end{bmatrix} \tag{39}$$

$$\mathbf{B}_{ij}^{\sigma_\mathrm{b}}(x) = \begin{bmatrix} -\frac{E}{L_i} & -\frac{6EI_{ij}}{L_i^2 W_{\mathrm{b},ij}} + \frac{12xEI_{ij}}{L_i^3 W_{\mathrm{b},ij}} & -\frac{4EI_{ij}}{L_i W_{\mathrm{b},ij}} + \frac{6xEI_{ij}}{L_i^2 W_{\mathrm{b},ij}} & \frac{E}{L_i} & \frac{6EI_{ij}}{L_i^2 W_{\mathrm{b},ij}} - \frac{12xEI_{ij}}{L_i^3 W_{\mathrm{b},ij}} & -\frac{2EI_{ij}}{L_i W_{\mathrm{b},ij}} + \frac{6xEI_{ij}}{L_i^2 W_{\mathrm{b},ij}} \end{bmatrix} \tag{40}$$

$$\mathbf{B}_{ij}^{\tau}(x) = \begin{bmatrix} 0 & -\frac{12ES_{ij}}{L_i^3 b_{ij}} & -\frac{6ES_{ij}}{L_i^2 b_{ij}} & 0 & \frac{12ES_{ij}}{L_i^3 b_{ij}} & -\frac{6ES_{ij}}{L_i^2 b_{ij}} \end{bmatrix} \tag{41}$$

The relevant stress components $\sigma_{\mathrm{t},ij}$, $\sigma_{\mathrm{b},ij}$ and/or τ_{ij} of element i with section j at the selected output locations x_k are collected in a vector $\mathbf{s}_{ij}$. For example, if the normal stress at the top of element i must be checked at the locations x_1, x_2, and x_3, the vector $\mathbf{s}_{ij}$ is given by:

$$\mathbf{s}_{ij} = \begin{bmatrix} \sigma_{\mathrm{t},ij}(x_1) \\ \sigma_{\mathrm{t},ij}(x_2) \\ \sigma_{\mathrm{t},ij}(x_3) \end{bmatrix} \tag{42}$$

This vector is obtained as follows:

$$\mathbf{s}_{ij} = \mathbf{B}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{\mathbf{s}}_{ij} \tag{43}$$

where $\mathbf{B}_{ij}$ is in this case composed as:

$$\mathbf{B}_{ij} = \begin{bmatrix} \mathbf{B}_{ij}^{\sigma_\mathrm{t}}(x_1) \\ \mathbf{B}_{ij}^{\sigma_\mathrm{t}}(x_2) \\ \mathbf{B}_{ij}^{\sigma_\mathrm{t}}(x_3) \end{bmatrix} \tag{44}$$

and $\tilde{\mathbf{s}}_{ij}$ as:

$$\tilde{\mathbf{s}}_{ij} = \begin{bmatrix} \tilde{\sigma}_{\mathrm{t},ij}(x_1) \\ \tilde{\sigma}_{\mathrm{t},ij}(x_2) \\ \tilde{\sigma}_{\mathrm{t},ij}(x_3) \end{bmatrix} \tag{45}$$

The internal resistance constraints read as:

$$\underline{\mathbf{s}}'_{ij} + (\underline{\mathbf{s}}_{ij} - \underline{\mathbf{s}}'_{ij})y_{ij} \leqslant \mathbf{B}_{ij}\mathbf{T}_i\mathbf{L}_i\mathbf{u} + \tilde{\mathbf{s}}_{ij} \leqslant \bar{\mathbf{s}}'_{ij} + (\bar{\mathbf{s}}_{ij} - \bar{\mathbf{s}}'_{ij})y_{ij} \quad \forall\, i \in \mathcal{M},\, \forall\, j \in \mathcal{C}_i \tag{46}$$

where $\tilde{\mathbf{s}}_{ij}$ is a vector containing the selected stress components at the predefined locations of the beam with clamped-clamped boundary conditions subjected to the element loads of member i for profile j, $\underline{\mathbf{s}}_{ij}$ and $\bar{\mathbf{s}}_{ij}$ are the prescribed minimum and maximum allowed stresses, and $\underline{\mathbf{s}}'_{ij}$ and $\bar{\mathbf{s}}'_{ij}$ are artificial lower and upper bounds in order to ensure feasibility when profile j is not selected for member i, calculated similarly to equations (24) and (25).

3 Three-bay three-story frame example problem

3.1 Problem description

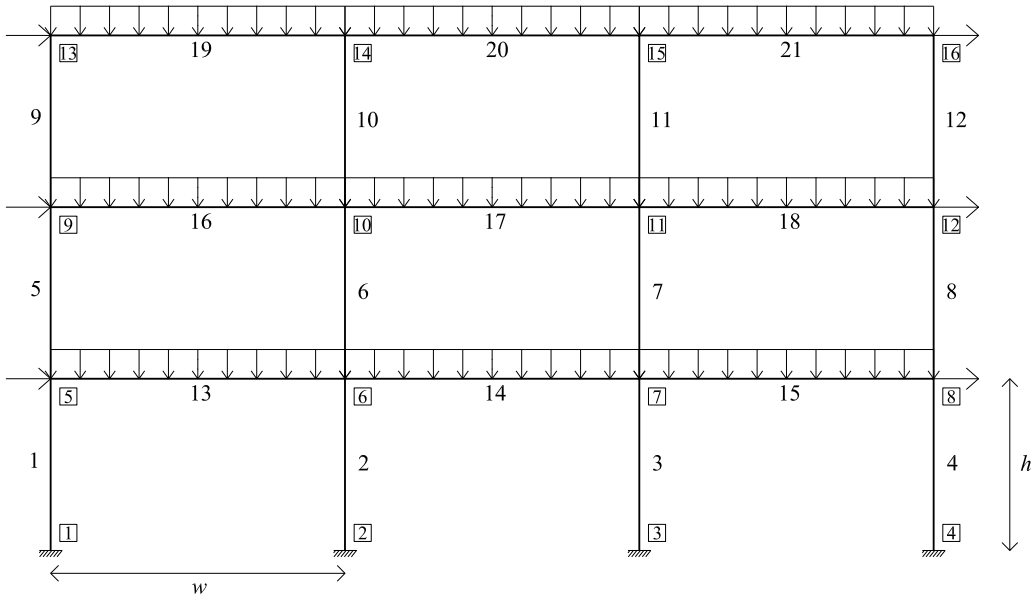


Figure 4: Three-bay three-story frame

Figure 4 shows a three-bay three-story frame structure with twenty-one members. The height of each story is $h = 3.5$ m, the width of each bay is $w = 6$ m, the value of the horizontal load is $F = 22.05$ kN, and the value of the distributed vertical load is $p = 50.1$ kN/m. The members are subdivided in seven groups: the profiles of all beams are the same, and for each story the profiles of the outer columns, and the profiles of the inner columns are the same.

The objective of the optimization problem is to minimize the weight of the structure. In order to reduce the computation time, a limited set of available profiles is used: the profile sections have to be chosen from a HEA catalog, but the range is limited from HEA 100 to HEA 400, meaning 15 alternatives. The Young's modulus of the material is 210 GPa, the material density is 7850 kg/m^3, and the yield strength of the material is $f_\mathrm{y} = 235$ MPa. The maximum allowed normal stress is $f_\mathrm{y} = 235$MPa, and the maximum allowed shear stress is $f_\mathrm{y}/\sqrt{3} = 136$MPa. For each member, all stress components are limited at three equidistant locations $x_1 = 0$, $x_2 = L_i/2$, and $x_3 = L_i$. For each column, the interstory drift $\triangle u$ is limited by $h/300 = 0.0117$ m. For each beam, the vertical deflection is limited at location $x_1 = L_i/2$ by $w/200 = 0.03$ m.

The minimum weight problem is given by equations (3) to (9). The total number of members is $n_\mathrm{m} = 21$, and for each member the number of available profiles is $n_\mathrm{s} = 15$, resulting in $n_\mathrm{b} = 315$ binary decision variables. The number of force variables is $6n_\mathrm{b} = 1890$, and the number of degrees of freedom is $n_\mathrm{dof} = 36$.

3.2 Results

The MILP consists of 2241 design variables, 13791 constraints, and 61105 nonzeros in the constraint coefficient matrix. The number of design combinations is $15^7 = 171 \times 10^6$. There are 21 equality constraints to ensure that only one profile is selected for each member (given by equation (4)), 36 nodal equilibrium equality constraints (5), 3780 member stiffness relation inequality constraints (6), 3780 force inequality constraints (7), 270 deflection constraints (8),

5670 stress inequality constraints (9), and 210 grouping constraints. The interstory drift constraints are imposed by limiting the difference of the horizontal nodal displacements at the top and bottom of each column. Consequently, there are 24 interstory drift constraints. The MILP is solved by the software Gurobi 6.0.2 [23] on a computer with an Intel Core i7-5600U processor (2.6 GHz clock frequency) and 8 GB RAM. The feasibility tolerance is set to 10^{-9}, the integer feasibility tolerance is set to 10^{-9}, and the optimality gap is set to 5×10^{-3}. The optimization problem is solved in 74781 seconds or 21 hours, and 1764305 nodes of the branch-and-bound tree are explored. The optimum design is given in figure 5. The total weight of the structure is 6131.87 kg.

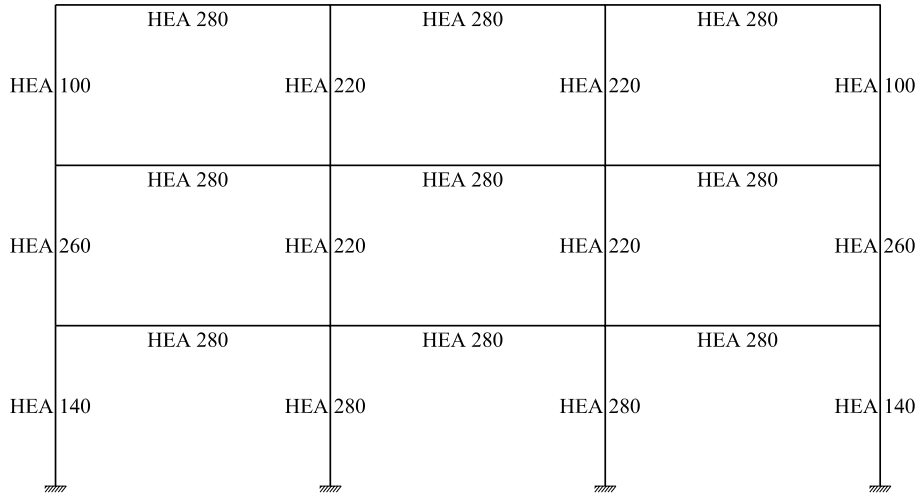


Figure 5: Optimal design

4 Conclusion

A method is proposed for the discrete size optimization of frames to global optimality. The optimization problem is reformulated as a mixed-integer linear program and solved with the branch-and-bound method in order to guarantee global optimality. The design variables consist of binary decision variables, continuous displacements, and continuous member end force variables. The equilibrium of the structure is imposed by nodal equilibrium constraints. The relation between the forces and the displacements is given by the member stiffness relation constraints. The relevant displacement components as well as the internal stresses are limited at several predefined locations by interpolating the nodal displacements using shape functions.

In this paper, only member resistance constraints and displacement constraints are considered. It is hard to consider all relevant code constraints, since not all constraints can be reformulated as linear constraints. Member stability constraints are therefore not taken into account. More research is needed to develop a fast optimization method that can take all relevant code constraints into account and that can solve discrete optimization problems to global optimality.

REFERENCES

[1] J. Farkas. Structural optimization as a harmony of design, fabrication and economy. *Structural and Multidisciplinary Optimization*, 30:66–75, 2005.

[2] Q. Wang and J. S. Arora. Alternative formulations for structural optimization: An evaluation using frames. *Journal of Structural Engineering*, 132(12):1880–1189, 2006.

[3] P.B. Thanedar and G.N. Vanderplaats. Survey of discrete variable optimization for structural design. *Journal of Structural Engineering*, 121:301–306, 1995.

[4] J.S. Arora and M.-W. Huang. Discrete structural optimization with commercially available sections. *Structural Engineering/Earthquake Engineering*, 13:93–110, 1996.

[5] M.-W. Huang and J.S. Arora. Optimal design of steel structures using standard sections. *Structural Optimization*, 14:24–35, 1997.

[6] J. Arora. Methods for discrete variable structural optimization. In Scott A. Burns, editor, *Recent Advances in Optimal Structural Design*, pages 1–40. ASCE, 2002.

[7] D.S. Juang and W.T. Chang. A revised discrete lagrangian-based search algorithm for the optimal design of skeletal structures using available sections. *Structural and Multidisciplinary Optimization*, 31:201–210, 2006.

[8] M. Schevenels, S. McGinn, A. Rolvink, and J. Coenders. An optimality criteria based method for discrete design optimization taking into account buildability constraints. *Structural and Multidisciplinary Optimization*, 50(5):755–774, 2014.

[9] R.J. Balling. Optimal steel frame design by simulated annealing. *Journal of structural engineering*, 117(6):1780–1795, 1991.

[10] C. Camp, S. Pezeshk, and C. Guozhong. Optimized design of two-dimensional structures using a genetic algorithm. *Journal of structural engineering*, 124(5):551–559, 1998.

[11] S. Rajeev and C.S. Krishnamoorthy. Discrete optimization of structures using genetic algorithms. *Journal of Structural Engineering*, 118(5):1233–1250, 1992.

[12] C.V. Camp, B.J. Bichon, and S.P. Stovall. Design of steel frames using ant colony optimization. *Journal of structural engineering*, 131(7):369–379, 2005.

[13] A.H. Gandomi, X.S. Yang, and A.H. Alavi. Mixed variable structural optimization using firefly algorithm. *Computers & Structures*, 89(23):2325–2336, 2011.

[14] M. Sonmez. Discrete optimum design of truss structures using artificial bee colony algorithm. *Structural and Multidisciplinary Optimization*, 43(1):85–97, 2011.

[15] G. Venter and J. Sobieszczanski-Sobieski. Particle swarm optimization. *AIAA Journal*, 41(8):1583–1589, 2003.

[16] J.S. Arora and Q. Wang. Review of formulations for structural and mechanical system optimization. *Structural and Multidisciplinary Optimization*, 30(4):251–272, 2005.

[17] I.E. Grossmann, V.T. Voudouris, and O. Ghattas. Mixed-integer linear programming reformulations for some nonlinear discrete design optimization problems. In C.A. Floudas and P.M. Pardalos, editors, *Recent advances in global optimization*, pages 478–512. Princeton University Press, 1992.

[18] M.H. Rasmussen and M. Stolpe. Global optimization of discrete truss topology design problems using a parallel cut-and-branch method. *Computers & Structures*, 86(13):1527–1538, 2008.

[19] K. Mela. Resolving issues with member buckling in truss topology optimization using a mixed variable approach. *Structural and Multidisciplinary Optimization*, 50(6):1037–1049, 2014.

[20] R. Van Mellaert, G. Lombaert, and M. Schevenels. Global size optimization of statically determinate trusses considering displacement, member, and joint constraints. *Journal of Structural Engineering*, 2015. Published online.

[21] A. Kassimali. *Matrix Analysis of Structures*. Brooks/Cole Publishing Company, London, UK, 1999.

[22] M. Stolpe and K. Svanberg. Modelling topology optimization problems as linear mixed 0-1 programs. *International Journal for Numerical Methods in Engineering*, 57:723–739, 2003.

[23] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A SUPER-STRUCTURE BASED OPTIMISATION APPROACH FOR BUILDING SPATIAL DESIGNS

Koen van der Blom[1], **Sjonnie Boonstra**[2], **Hèrm Hofmeyer**[2] **and Michael T. M. Emmerich**[1]

[1] Leiden Institute of Advanced Computer Science,
Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
e-mail: {k.van.der.blom, m.t.m.emmerich}@liacs.leidenuniv.nl

[2] Eindhoven University of Technology, Department of the Built Environment,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: h.hofmeyer@tue.nl

Keywords: Evolutionary Algorithms, Super-structure, Mixed-Integer Optimisation, Building Spatial Design, Building Structural Design.

Abstract. *Building design can be supported effectively by computer-aided design exploration. This paper investigates optimisation based on a mixed-integer super-structure representation of the search space of building spatial designs. It can take into account parametric as well as topological variations. In the suggested super-structure – the so-called supercube representation – discrete and continuous variables determine the existence, respectively, dimensioning of spaces of the building spatial design. Constraints are formulated as closed form equations and can be used to numerically assess the feasibility of designs. A population-based constraint-handling evolutionary strategy is developed. In the constraint handling repair and penalty methods are combined in a domain specific way. The method is tested on different search space sizes and first promising results are reported.*

1 INTRODUCTION

The use of automated search for finding optimal buildings with respect to various criteria, such as, e.g. structural strength, and stiffness; and energy performance will be an important research topic in the 21st century. Many physical phenomena can be modelled and studied by computer simulations, e.g., mechanical stress, heat transport, and light radiation. However, via these means, search for improvement is still widely limited to a trial and error approach. New research is required for enabling the search in larger design spaces.

This paper is about the foundations of such research. Namely, it will address the question how to represent a search space for optimisation and how search methods can navigate this design space in search for better designs. For this, global optimisation methods for large search spaces, such as evolutionary algorithms, will be regarded here, for across various engineering domains they have shown their potential to discover new, often unexpected design solutions. A method that can take into account structural as well as continuous variables will be proposed. Moreover, it is discussed how to handle constraints on discrete and continuous variables. Both, repair and penalty methods are used and it is investigated how often constraints are violated in stochastic search. The results on building spatial designs of small and moderate size are promising and provide interesting indications on how to plan future research.

The paper is structured as follows: In Section 2 the problem of building design is introduced as well as a review of related work. The basic outline of the optimisation algorithm is provided in Section 3. In Section 4 the search space, objective functions and constraints are discussed in more detail. Then in Section 5 the coupling of optimisation and constraint handling is discussed and the setup of experiments used for validating the approach. An in-depth discussion of results is presented in Section 6. In Section 7 a summary of the main results is provided and directions for future work are indicated.

2 BUILDING DESIGN

Building design is traditionally performed by architects and engineers who create solutions for discipline specific design problems. These solutions are nowadays usually assessed by and modified in accordance with design analysis tools. Such tools are for example finite element methods (FEM), to simulate structural performance or computational fluid dynamics (CFD) to simulate simulate heat, air, and moisture problems. The division created by the different disciplines within the field also calls for tools that allow engineers within different disciplines to cooperate. Examples are computer aided design (CAD) that is used to create and share designs. However, currently, building information modelling (BIM) is on the rise. BIM is a method that uses data management in order to dynamically share information with other disciplines. This allows engineers to – among other things – take other disciplines into account in the early design phase. The early design phase is important for optimal building designs, because decisions in the conceptual design stage often affect performances across all disciplines. A single disciplinary design may therefore lead to a sub-optimal multi-disciplinary design. Optimisation in the built environment is mostly performed by parametrising building components, e.g. installation type, construction type, material type, dimensions or shapes. Software tools for building optimisation purposes have been developed: Palonen *et al.* [1] give an overview and present their own tool. In the tool of Palonen, design variables of a building design can be selected for optimisation, an optimisation strategy can be selected thereafter. Although these tools can improve and alter a designs appearance greatly, they cannot lead to new designs (e.g. a new window cannot appear). Very recently, advances in early design optimisation are made: Hofmeyer

& Davila Delgado [2] for example use a design process inspired optimisation approach to optimise a building spatial design for the structural performance of its related structural design. Attia *et al.* [3] present software that gives information about building physics performances during early design stages. Hopfe *et al.* [4] use statistical sensitivity analysis to predict the impact of design variables on the optimality of a building design. The analysis by Hopfe *et al.* is interesting for early design optimisation as the impact of each design variable in distinct design stages can be investigated.

Here it is tried to perform building optimisation for early stage building spatial design using a new super-structure design space representation. A building spatial design is here merely a layout of building spaces that can be rearranged and resized for optimal performance. Optimisation methods are investigated for two different objective functions on the design space: Structural performance, measured by minimal compliance, and building physics, measured by minimal outside surface area, are selected for the optimisation objectives. These disciplines are selected because they are known to be dependent on the building spatial design. It is intended to use a RC-network to analyse building heat-balance in the future, but for the sake of developing the optimisation strategy presented here only the minimal surface area is taken as objective function. Details on the setup of the analysis are deferred to Section 5.

3 OPTIMISATION

Evolutionary algorithms subsume different algorithms that mimic natural evolution, in order to find improved or optimised technological designs [5]. Population-based evolutionary algorithms generally work according to a basic loop structure – the so-called generational loop. It starts after an initialisation phase where an initial population of individuals (solution candidates) is generated and evaluated. In the evolutionary loop first a ranking among individuals according to their fitness (evaluation results) is established. The following step is to select the parents to generate an offspring population. In this step the ranking of the population might be taken into account, although in Evolution Strategies – an important EA variant – parent individuals are chosen randomly. From the selected parent individuals λ offspring individuals are created. Recombination is applied to allow parts of the genome from multiple parents to together from a new genome. In order to introduce new, possibly not previously considered, information into the genome, random perturbations are applied through mutation of the newly produced genome. When applicable, this is followed by constraint evaluation, where invalid individuals may either be repaired, penalised or discarded. Finally, the offspring population is evaluated on the objective function and a new parent population is produced.

The specific evolutionary algorithm type chosen in this paper is the $(\mu + \lambda)$-Evolution Strategy. Evolution Strategies (ESs) were developed by Ingo Rechenberg and Hans-Paul Schwefel at the TU Berlin in the 60s and are especially well suited for solving engineering design problems [6]. They are interesting for this work, as they can deal with discrete as well as with continuous design variables, as outlined in Li *et al.* [7]. In Figure 1 the main loop of a $(\mu + \lambda)$-ES is summarised. Basically, the population (multiset) of individuals is used as a template to generate the offspring population M of size λ and from the combination of parents and offspring the best individuals are selected as the parents of the next round. The initialisation, mutation and recombination operators are chosen in a domain specific way, as will be discussed in more detail in Section 5. For a more detailed discussion on evolution strategies and their properties the reader is referred to [5] and [8].

1. Initialise Population $P_0 \in S$
2. Evaluate P_0
3. $t \to 0$
4. While (Termination criterion is not satisfied)
5. $\quad t \to t + 1$
6. $\quad C_t =$ Randomly select λ pairs of individuals from P_{t-1}
7. $\quad R_t = \{recombine(c, c') | (c, c') \in C_t\}$
8. $\quad M_t = \{mutate(r) | \, r \in R_t\}$
9. $\quad$ Evaluate M_t
10. $\quad P_t =$ Select μ best individuals from $M_t \cup P_{t-1}$
11. End While
12. Return best individual in P_t

Figure 1: $(\mu + \lambda)$-Evolution Strategy.

4 PROBLEM

One important task before optimisation can be used is to formulate a search space. In this work, it is also intended to use parameter optimisation to optimise the building design. This makes it necessary to encode the existence or non-existence of spaces by discrete variables. Moreover the connectivity of spaces – the topology of the building spatial design – is encoded by discrete variables. In contrast, the sizing of the spaces is encoded by continuous variables.

In the following, the supercube representation of building spatial designs will be used. It can encode a large search space of building spatial designs by a fixed size vector of discrete and continuous variables. This makes it easy to apply mixed-integer optimisation algorithms directly for optimisation, as will be exemplified by using a mixed-integer evolution strategy later in this work.

The supercube representation has recently been introduced in [9]. It is here assumed a building spatial design consists of maximally N_{spaces} spaces that can be seen as mapped to a $N_w \times N_d \times N_h$ 3D rectangular (cuboid) grid, consisting of $N_w \times N_d \times N_h$ cells. Here N_w, N_d and N_h are the number of subdivisions in width, depth and respectively height. An example of a supercube grid is shown in Figure 2. The size of the cells is determined by the dimensioning variables w_i, d_j and h_k for the width, depth and respectively height of the cells. Here, $i \in \{1, \ldots, N_w\}$, $j \in \{1, \ldots, N_d\}$ and $k \in \{1, \ldots, N_k\}$. The building spatial design is represented by binary variables $b^{\ell}_{i,j,k}$ where i, j, k are again the indices to the width, depth and respectively height and ℓ refers to the different spaces. If $b^{\ell}_{i,j,k} = 1$ this means that the cell with indices i, j, k is part of space ℓ, otherwise this is not the case. Not all configurations of cells make sense from a building spatial design perspective. Constraints will therefore be formulated to restrict the design space to reasonable solutions.

In summary the following variables will be subject to optimisation:

Discrete variables: (all binary)

$$b^{\ell}_{i,j,k}, i \in \{1, \ldots, N_w\}, j \in \{1, \ldots, N_d\}, k \in \{1, \ldots, N_h\}, \text{ and } \ell \in \{1, \ldots, N_{spaces}\}$$

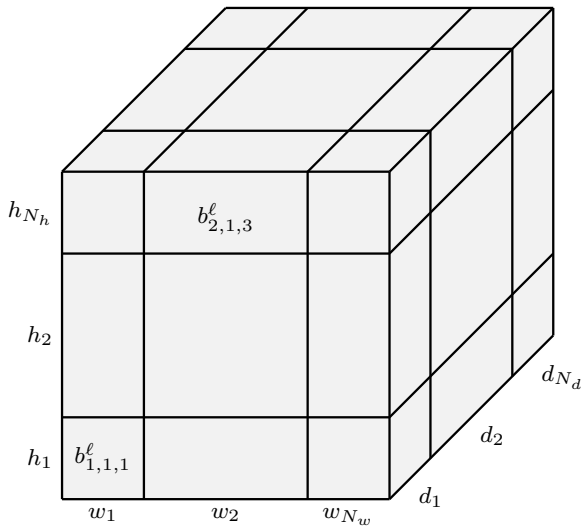


Figure 2: Grid used in the supercube representation.

Continuous variables:

$$w_i, i \in \{1, \ldots, N_w\}, d_j, j \in \{1, \ldots, N_d\}, \text{ and } h_k, k \in \{1, \ldots, N_h\}$$

4.1 Constraints on the discrete variables

In the following four types of topology constraints are introduced, see also [9]. To avoid overlap of spaces every cell can belong to at most one space. This will be enforced by the following constraint in Equation 1.

$$\forall_{i,j,k} \sum_{\ell=1}^{N_{spaces}} b^{\ell}_{i,j,k} \leq 1 \tag{1}$$

Spaces should have the form of a cuboid (3D rectangle). To enforce this constraint one can first extend the supercube grid by adding a layer of cells with binary variables equal to zero around it (Equation 2).

$$\begin{gathered} \forall_{\ell} : \forall_{i,j,k} \in \{0, \ldots, N_w + 1\} \times \{0, \ldots, N_d + 1\} \times \{0, \ldots, N_h + 1\} : \\ i = 0 \vee j = 0 \vee k = 0 \vee i = N_w + 1 \vee j = N_d + 1 \vee k = N_h + 1 \Rightarrow b^{\ell}_{i,j,k} = 0 \end{gathered} \tag{2}$$

Moreover a i, j beam is defined as a set of cells that share the same i, j index. Accordingly j, k and i, k beams are defined. For all i, j beams and all spaces ℓ transitions from zero to one (Equation 3) should always occur at the same position and transitions from one to zero (Equation 4) for space ℓ should also always occur at the same position. The same holds for j, k and i, k beams (not in equations here).

$$\begin{gathered} \forall_{\ell} : \forall_{i_1,j_1,i_2,j_2} : \left(\left(\sum_{k=1}^{N_h} k \left(1 - b^{\ell}_{i_1,j_1,k-1}\right) b^{\ell}_{i_1,j_1,k} \right) - \left(\sum_{k=1}^{N_h} k \left(1 - b^{\ell}_{i_2,j_2,k-1}\right) b^{\ell}_{i_2,j_2,k} \right) \right) \\ \left(\sum_{k=1}^{N_h} b^{\ell}_{i_1,j_1,k} \right) \left(\sum_{k=1}^{N_h} b^{\ell}_{i_2,j_2,k} \right) = 0 \end{gathered} \tag{3}$$

$$\forall_\ell : \forall_{i_1,j_1,i_2,j_2} : \left(\left(\sum_{k=1}^{N_h} k b^\ell_{i_1,j_1,k} \left(1 - b^\ell_{i_1,j_1,k+1}\right) \right) - \left(\sum_{k=1}^{N_h} k b^\ell_{i_2,j_2,k} \left(1 - b^\ell_{i_2,j_2,k+1}\right) \right) \right) \left(\sum_{k=1}^{N_h} b^\ell_{i_1,j_1,k} \right) \left(\sum_{k=1}^{N_h} b^\ell_{i_2,j_2,k} \right) = 0 \tag{4}$$

Moreover, to ensure connectedness (if the previous two equations also hold) of the cells of a space ℓ all beams (in all directions) should have at most one transition from zero to one (Equation 5).

$$\begin{gathered} \forall_\ell : \\ \forall_{i,j} : \sum_{k=0}^{N_h} \left(1 - b^\ell_{i,j,k}\right) b^\ell_{i,j,k+1} \leq 1 \qquad \forall_{i,k} : \sum_{j=0}^{N_d} \left(1 - b^\ell_{i,j,k}\right) b^\ell_{i,j+1,k} \leq 1 \\ \forall_{j,k} : \sum_{i=0}^{N_w} \left(1 - b^\ell_{i,j,k}\right) b^\ell_{i+1,j,k} \leq 1 \end{gathered} \tag{5}$$

Building spatial designs normally stand on the ground. This is difficult to check by a simple equation if vertical gaps are allowed in the spatial design. Therefore it is suggested to enforce a no vertical gaps constraint as well. As a result of this constraint it is not possible to describe structures with cantilevers, overhangs or archways. The no vertical gaps constraint could be abandoned if one is willing to use more complex procedures to check constraints. These constraints are simultaneously enforced by disallowing transitions from zero to one for i, j beams. Let $b_{i,j,k}$ be the outcome of a logical OR of all ℓ bits belonging to cell i, j, k. In equations: $\forall_{i,j,k} : b_{i,j,k} = sgn(\sum_{\ell=1}^{N_{spaces}} b^\ell_{i,j,k})$, where the $sgn()$ may be omitted if the no-overlap constraint (Equation 1) is satisfied. If Equation 6 holds, the building spatial design has no vertical gaps and stands on the ground.

$$\forall_{i,j} : \left(\sum_{k=1}^{N_h - 1} \left(1 - b_{i,j,k}\right) b_{i,j,k+1} \right) = 0 \tag{6}$$

The number of described spaces is kept constant. This is achieved by ensuring every space is described by at least one cell (Equation 7).

$$\forall_\ell : \sum_{i=1}^{N_w} \sum_{j=1}^{N_d} \sum_{k=1}^{N_h} b^\ell_{i,j,k} \geq 1 \tag{7}$$

4.2 Constraints on the continuous variables

In the design of buildings the total volume V_0 of the building is could be provided as a constraint. To exclude inactive cells (not part of any building space) from the volume computation here $b_{i,j,k}$ is again taken to be the result of a logical OR over all ℓ bits of a cell i, j, k. This yields the equality constraint in Equation 8 below:

$$\sum_{i=1}^{N_w} \sum_{j=1}^{N_d} \sum_{k=1}^{N_h} b_{i,j,k} w_i d_j h_k = V_0 \tag{8}$$

In addition, all continuous variables should be positive or taken from a range of positive values.

4.3 Objective function

Optimisation is performed with respect to two objective functions from two different disciplines: Minimum compliance, which relates to structural design performance, and total outer surface area, which relates to energy performance.

The *structural design performance* is assessed by first providing the building spatial design with a building structural design. This is carried out by applying a so-called structural grammar on each building space. The grammar selected here adds four walls ($t = 150mm$), with a slab on top ($t = 150mm$), all made of concrete (elasticity modulus: $E = 30000N/mm^2$ and Poisson's ratio: $v = 0.3$). The building spatial design is then loaded with a live load on each floor surface ($1.8kN/m^2$) and wind loads on each outside surface ($1.0kN/m^2$ for pressure, $0.5kN/m^2$ for suction and $0.4kN/m^2$ for shear) from eight general directions (N, N/W, W, etc.). After the transfer of the loads to the building structural design and meshing of the structure, finite element analyses are carried out to find the total compliance for all loads together. Detailed information can be found in the paper by Hofmeyer & Davila Delgado [2]. In summary the first optimisation task is to minimise the total compliance S_C subject to the given constraints.

The *total surface area* objective can be computed for the supercube representation as follows. Note that for this computation it is required that the building spatial design contains no cantilevers or archways, this is ensured by the no vertical gaps constraint previously described with Equation 6. Additionally the computation requires a layer cells with their binary variables equal to zero around the supercube (Equation 2).

Every ray through the supercube in width and depth measures the number of changes from zero to one and then multiplies this number of changes by the area of the fixed indices and then by two to take into account both the entry and exit points. In case of the height direction the multiplication by two is omitted, because for the height direction the connection with the ground layer is not counted as surface area. Since there are no vertical gaps, the height direction is essentially the sum of areas of rays where space exists. The total sum $S_A := S_h + S_d + S_w$ of Equations 9, 10 and 11 below is then the total surface area.

$$S_h = \sum_{i=1}^{N_w} \sum_{j=1}^{N_d} \left(\left(\sum_{k=1}^{N_h+1} \left(1 - b_{i,j,k-1}\right) b_{i,j,k} \right) w_i d_j \right) \tag{9}$$

$$S_d = \sum_{i=1}^{N_w} \sum_{k=1}^{N_h} \left(2 \left(\sum_{j=1}^{N_d+1} \left(1 - b_{i,j-1,k}\right) b_{i,j,k} \right) w_i h_k \right) \tag{10}$$

$$S_w = \sum_{j=1}^{N_d} \sum_{k=1}^{N_h} \left(2 \left(\sum_{i=1}^{N_w+1} \left(1 - b_{i-1,j,k}\right) b_{i,j,k} \right) d_j h_k \right) \tag{11}$$

In summary the second optimisation task is to minimise the surface area S_A subject to the given constraints.

5 OPTIMISATION METHOD SETUP

Next, a description follows of how the earlier introduced $(\mu + \lambda)$-ES is customised for the optimisation based on the supercube representation, including the handling of constraints.

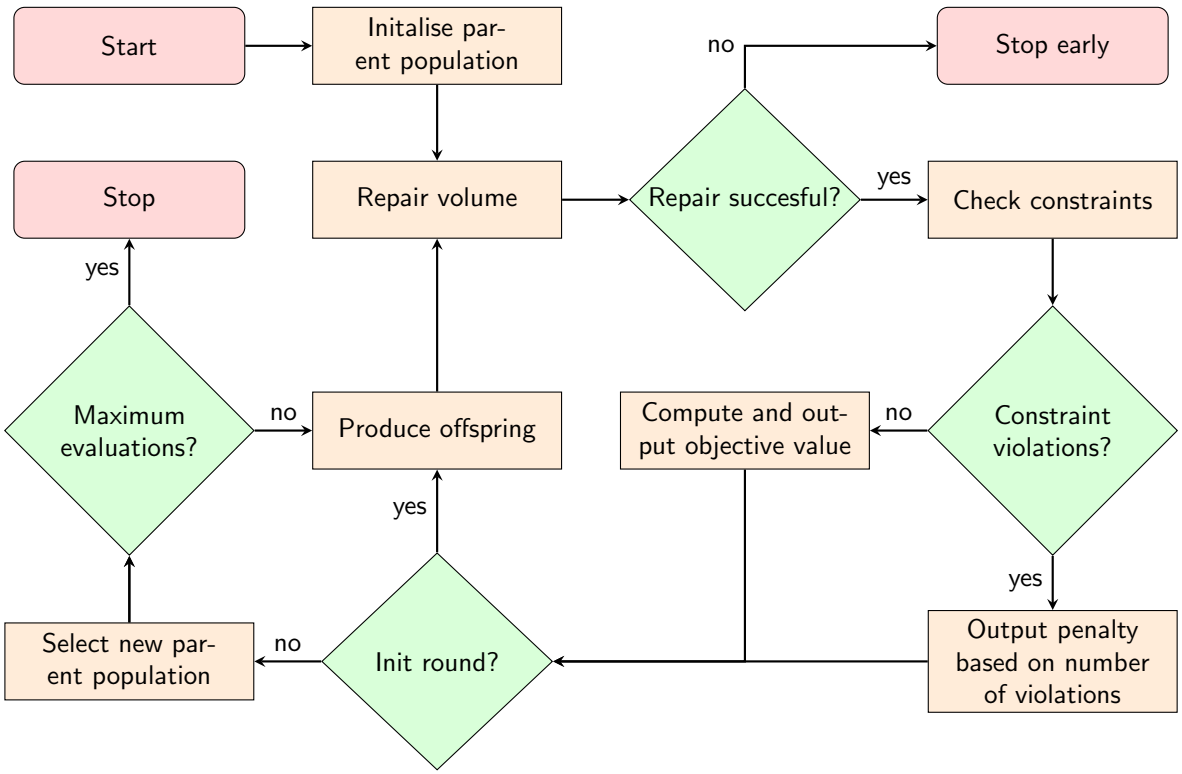


Figure 3: Optimisation outline.

The optimisation procedure is outlined in Figure 3. For convenience three definitions are introduced, the number of cells: $N_{cells} := N_w \times N_d \times N_h$, the number of continuous parameters: $N_{cont} := N_w + N_d + N_h$, and the total number of dimensions: $N_{dims} := N_{cont} + N_{cells} \times N_{spaces}$. The process starts by initialising the parent population of size $\mu = 20$. Here continuous parameters $(x_1, \ldots, x_{N_{cont}})$ are initialised to a uniformly random value in $[lb = 1.5, \ldots, ub = 9.9]$, and the binary parameters $(x_{N_{cont}+1}, \ldots, x_{N_{dims}})$ to one with probability $1/N_{cells}$ or zero otherwise. Step sizes of the continuous parameters are initialised to 0.1. Following this the volume of all new individuals is repaired to be within 1% of the desired volume $2^3 \times N_{cells}$ (a detailed description follows after this outline). In order to check constraints the continuous parameters are then converted to millimetres (multiply by 2000). In case any constraints are violated either a single penalty value of $pen = 999,999,999$ is output or a penalty value equal to $pen + CV - 1$ based on the number of constraint violations $CV \in 1, \ldots, 5$ is output as objective value. Five constraints are considered for the value of CV: All spaces exist, no-overlap, cuboid shape, connected cuboid and no vertical gaps. When no constraints are violated the objective value is computed and output, either surface area in square metres or compliance in Newton metres. For the initialisation round $\lambda = 100$ offspring are produced immediately, all later iterations first make a selection of the μ best (lowest objective value) individuals from the $\mu+\lambda$ individuals (parents+offspring). For every to be produced offspring individual two parents are selected (possibly the same twice) uniformly at random. Using these two parents, crossover is applied to produce a new individual, which is then mutated, and finally repaired if it exceeds the bounds. This is done by Modified Interval Bounds Treatment, where for parameters the previously defined lb and ub parameters are used and for the step sizes bounds $lb_s = 0.01$ and $ub_s = ub \times 0.1$ are used. Intermediate crossover, taking the mean value of the parents, is applied to the continuous parameters as well as their corresponding step sizes. Gaussian mutation with

individual step sizes is applied to the continuous parameters. Their step sizes are mutated by the local learning rate $\tau_1 = 1/\sqrt{2\sqrt{N_{dims}}}$ and global learning rate $\tau_2 = 1/\sqrt{2 \times N_{dims}}$ as in Li *et al.* [7]. Implementation wise a global value g_1 is drawn from a Gaussian distribution $g_1 = G(0, 1)$ for every individual. For every parameter separately a second value $g_2 = G(0, 1)$ and third value $g_3 = G(0, 1)$ are drawn. The step size s of a parameter is then mutated as: $s' = s \times exp(g_1 \times \tau_2 + g_2 \times \tau_1)$. The newly mutated step size is then used to mutate its corresponding parameter x as: $x' = x + g_3 \times s'$. Binary parameters apply crossover by copying the value from one of the parents, chosen uniformly at random for each bit. Binary parameters are mutated by flipping each bit with a probability of $1/(N_{cells} \times N_{spaces})$. Following this the volume of the new offspring individuals is repaired and the loop repeats until the desired number of evaluations is reached. Note that here evaluations are counted based on the number of valid (non-constraint violating) solutions.

A detailed description of volume repair follows. When a newly created individual does not satisfy the volume constraint in Equation 8 its volume is repaired, which is done by scaling the continuous parameters. The desired total volume V_0 is defined for each experiment. The current total volume V_c is simply the outcome of the left-hand side of Equation 8. These two values allow for the computation of a factor $\alpha = V_0/V_c$. The desired volume is then reached by multiplying the continuous parameters by the cubic root of α as shown in Equation 12. Note that this should only be done for active cells in the supercube (cells occupied by at least one building space) since as defined before inactive cells do not contribute to the volume. As a result of the creation process (recombination, mutation) of new individuals and the described volume repair, continuous parameters of individuals may end up exceeding their lower bound lb or upper bound ub. To correct for this, parameters exceeding the lower bound are set to the lower bound, parameters exceeding the upper bound are multiplied by 0.95 until they are within the bound. Clearly the corrections for the bounds affect the volume. Therefore the volume repair and bound corrections are iteratively solved up to 26 times until both requirements are satisfied. When this number of corrections is insufficient for any of the individuals the optimisation process is stopped and considered as unsuccessful. This did not occur in the later presented experiments, the likelihood of this occurring depends on the chosen bounds and the desired volume.

$$\forall_i : w_i = \sqrt[3]{\alpha w_i} \qquad \forall_j : d_j = \sqrt[3]{\alpha d_j} \qquad \forall_k : h_k = \sqrt[3]{\alpha h_k} \tag{12}$$

Based on the described optimisation outline a number of experiments have been carried out. Namely, individual optimisation of both the surface area and the compliance using a single penalty value is used to test how well the proposed supercube description works in practice. Moreover, this will set a baseline for the objective value of both functions. In addition a comparison is made to individual optimisation of the same objectives when using a penalty based on the number of constraint violations. This comparison aims to provide some first insights into constraint handling for this heavily constrained problem.

All experiments are conducted with an evaluation budget of 1000, for six different configurations. A supercube of size two ($2 \times 2 \times 2$ cells) with one, three and five spaces, and a supercube of size three ($3 \times 3 \times 3$ cells), also with one, three and five spaces. After generating one million candidate solutions the optimisation is stopped, even if 1000 valid candidates are not yet found.

6 RESULTS

First experiments with a simple, single penalty, constraint handling technique are reported as shown in Figures 4 and 5. Here infeasible solutions are penalised by a constant penalty that

is higher than all objective function values of feasible solutions. For each plot mean values for different numbers of spaces are shown. Different problem configurations are denoted by four numbers. For instance '1234' would refer to a configuration with a width of one, depth of two and height of three cells, and describing four spaces. Five runs of the optimisation process are performed for every configuration.

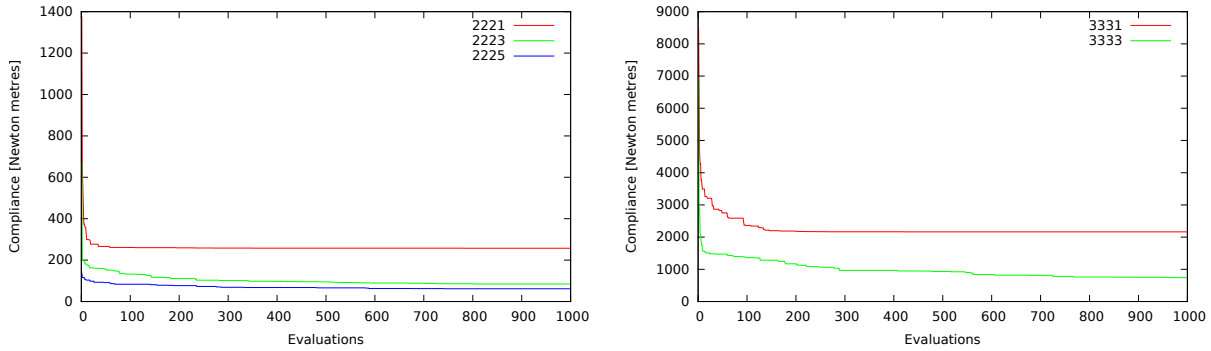


Figure 4: Mean convergence of the compliance for all completed runs (maximum of five) of single penalty 222x and 333x configurations, for one, three and five spaces. 2225 and 3333 completed 3 of 5 runs and 3335 completed 0 of 5 runs. All incomplete runs did not find any valid spatial designs. All not specifically mentioned configurations completed all five runs.

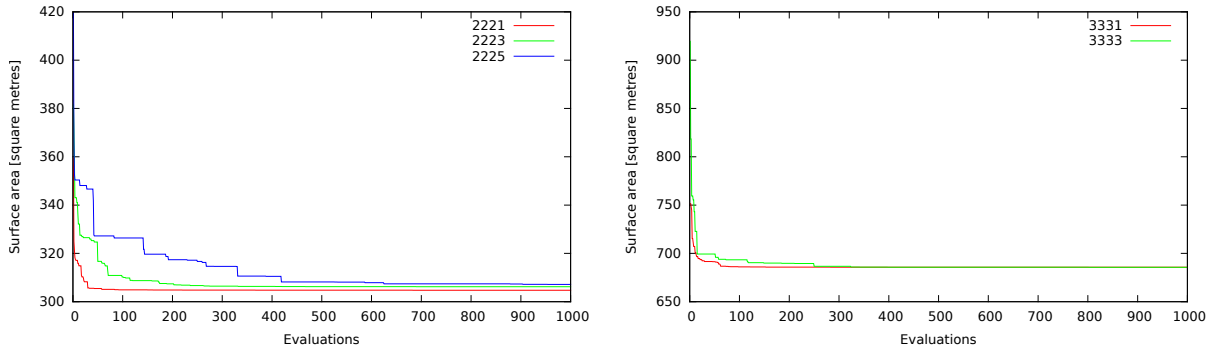


Figure 5: Mean convergence of the surface area for completed runs (maximum of five) of single penalty 222x and 333x configurations, for one, three and five spaces. 2225 completed 2 of 5 runs, 3333 completed 3 of 5 runs and 3335 completed 0 of 5 runs. All incomplete runs did not find any valid spatial designs. All not specifically mentioned configurations completed all five runs.

The results in Figure 4 show convergence plots for the compliance objective. After a rapid decrease in the function value during the first few hundred evaluations the optimisation process tends to stagnate. Configurations 2225, 3333 and 3335 did not find any valid solutions in all of their runs. Namely configurations 2225 and 3333 both completed three of the five runs and configuration 3335 completed none of the five runs. Mean convergence values in the plots are computed over the completed runs, that is, runs that found 1000 feasible solutions. For surface area similar results were found as shown in Figure 5. Here too, some configurations did not complete all their runs, respectively configurations 2225, 3333 and 3335 completed two, three and zero runs.

For the ease of discussion the constraints are summarised here with a reference to the relevant equations. *Existence*, all spaces exist, i.e. they are described by at least one cell (Equation 7). *No-overlap*, each cell belongs to no more than one space (Equation 1). *Cuboid shape*, the cells describing a space together form a cuboid shape, possibly with voids (Equations 3 and 4). *Connected cuboid*, given the cuboid shape constraint holds, all cells forming a space are connected and form a cuboid without voids (Equation 5). *No vertical gaps*, all cells are either on the ground level, or have an active cell on the level below them (Equation 6). Note that for some problem configurations certain constraints are always satisfied, these are indicated as not applicable (N/A).

Table 1 shows the ratios of constraint violations for every configuration and constraint type for the minimal compliance objective with a single penalty value. For small supercube sizes the existence constraint poses no problem. The no-overlap constraint depends on the ratio between the number of spaces and the number of cells. The probability of constraint violation for the 3333 and 3335 configurations is extremely high (0.999218016 and 0.999294014), making it impractical to search. The remaining three constraints show a similar pattern to the no-overlap constraint. The constraint violations of the surface area in Table 2 show a largely similar behaviour, with a large portion of the constraints occurring with high probabilities. Constraint violation appears to be a major problem for this approach.

Configuration	Existence	No-overlap	Cuboid shape	Connected cuboid	No vertical gaps
2221	0.072030329	N/A	0.430918281	N/A	0.350463353
2223	0.148770246	0.373125375	0.528494301	N/A	0.393521296
2225	0.197997775	0.610956619	0.565072303	N/A	0.482480534
3331	0.014877790	N/A	0.590860786	0.160821821	0.501239816
3333	0.000360993	0.999218016	0.999928001	0.999147017	0.999986000
3335	0.816557669	0.999294014	0.999885002	0.999008020	0.999954001

Table 1: Mean constraint violation probability over five runs for minimal compliance optimisation with a single penalty value for various problem configurations.

Configuration	Existence	No-overlap	Cuboid shape	Connected cuboid	No vertical gaps
2221	0.054300608	N/A	0.477410947	N/A	0.288010426
2223	0.176321781	0.433341766	0.589931697	N/A	0.348343031
2225	0.021001580	0.999803004	0.999840003	N/A	0.998622028
3331	0.040989160	N/A	0.525406504	0.168021680	0.483739837
3333	0.080888636	0.195670749	0.656508117	0.158074623	0.558245514
3335	0.816465671	0.999535009	0.999898002	0.999337013	0.999976001

Table 2: Mean constraint violation probability over five runs for surface area optimisation with a single penalty value for various problem configurations.

To remedy this problem a graduated penalty function was created. It penalises based on the number of constraints that are violated. This allows evolutionary search to gradually correct bits and find feasible solutions faster. Clearly Figures 6 and 7 show the success of this strategy where all runs converge and it is now possible to deal with bigger grid sizes (e.g. 3335).

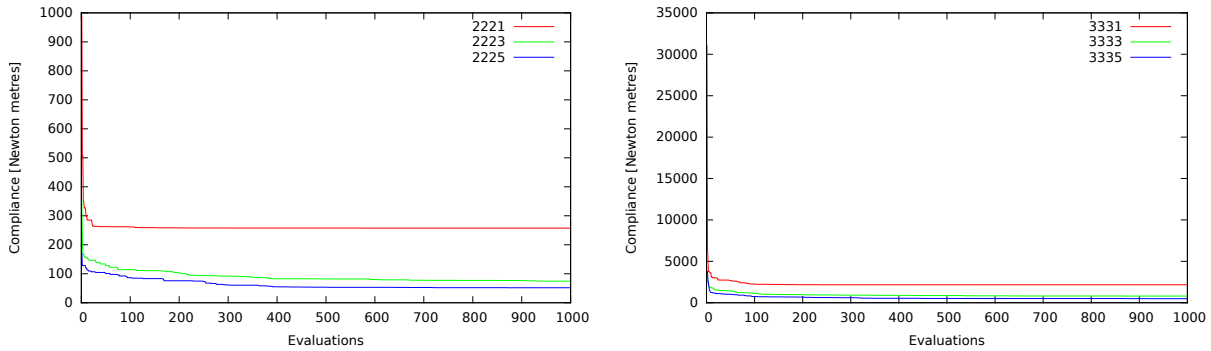


Figure 6: Mean convergence of the minimal compliance optimisation with constraint penalties based on the number of violations for five runs of 222x and 333x configurations, for one, three and five spaces.

In terms of constraint violations it is observed that penalising based on the number of violated constraints vastly improves the chance of finding valid solutions, as shown in Tables 3 and 4.

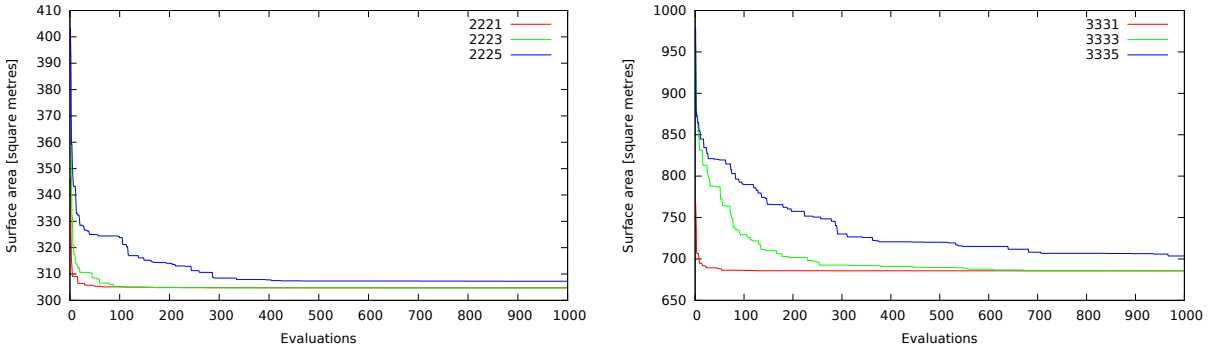


Figure 7: Mean convergence of the surface area optimisation with constraint penalties based on the number of violations for five runs of 222x and 333x configurations, for one, three and five spaces.

Where with a single penalty a significant number of constraints were violated with probabilities above 80% this is now reduced to a single case. Only the existence constraint remains a major problem, and even this is only true for the largest configuration (3335). Even so, many other constraint violations still occur around 50% of the time, and clearly these would benefit from further improvement as well.

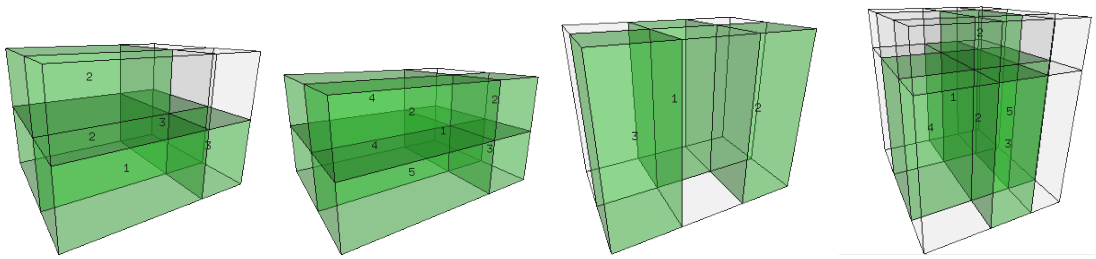
Configuration	Existence	No-overlap	Cuboid shape	Connected cuboid	No vertical gaps
2221	0.066305819	N/A	0.423545332	N/A	0.328371673
2223	0.156955204	0.322667565	0.467497474	N/A	0.313910408
2225	0.305234899	0.484563758	0.455570470	N/A	0.334228188
3331	0.013302295	N/A	0.605919521	0.119388094	0.518789491
3333	0.143656716	0.112873134	0.559701493	0.078358209	0.456778607
3335	0.840781999	0.154225102	0.462875022	0.082638026	0.422732114

Table 3: Mean constraint violation probability with penalties based on the number of violations over five runs for minimal compliance optimisation for various problem configurations.

Configuration	Existence	No-overlap	Cuboid shape	Connected cuboid	No vertical gaps
2221	0.063636365	N/A	0.501581028	N/A	0.269960474
2223	0.144760533	0.320849838	0.453366943	N/A	0.142599928
2225	0.197621226	0.532174443	0.471790170	N/A	0.365965233
3331	0.063481457	N/A	0.521550284	0.105913799	0.467758102
3333	0.142694064	0.146689498	0.582191781	0.099029680	0.469463470
3335	0.859725404	0.193019717	0.495911166	0.101861297	0.420656555

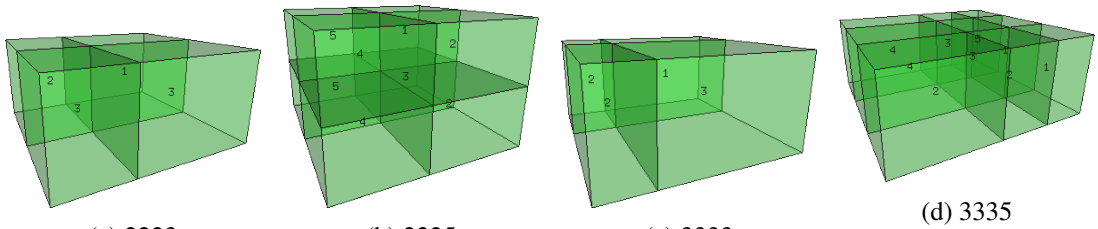
Table 4: Mean constraint violation probability with penalties based on the number of violations over five runs for surface area optimisation for various problem configurations.

The example results of some building spatial designs are given here, it was observed that different configurations resulted in different spatial designs. There is a clear distinction between the results from minimal compliance optimisation (Figure 8) and those of surface area optimisation (Figure 9). Surface area optimisation leads to compact cuboid, or near cuboid, shapes, as might be expected. Minimal compliance optimisation on the other hand produces a variety of shapes. The similarity between both 2225 configurations is striking, but probably the result of the limited variety of space arrangements within an order two supercupe. On the other hand, little use is made of the extra space in the 333x configurations. For the 3333 configurations this may be explained by the availability of only three spaces; there is a limited number of valid building spatial designs that can make use of the larger number of cells with only three spaces. Note that while it is possible to produce spaces consisting of a large number of cells, reaching such a situation becomes increasingly difficult with the number of cells while also satisfying the constraints. For example the largest space in the 3335 configuration for surface area optimisation (Figure 9d) consists of just two cells. This is likely to play a role in the limited use of space for both the 3333 and 3335 configurations. These frequent issues with constraints complicate exploring all feasible solutions in the search space, in particular transitions between different feasible parts of the search space are challenging when many moves end up in infeasible parts of the search space.



(a) 2223 (b) 2225 (c) 3333 (d) 3335

Figure 8: Examples of output spatial designs of compliance optimisation.



(a) 2223 (b) 2225 (c) 3333 (d) 3335

Figure 9: Examples of output spatial designs of surface area optimisation.

7 DISCUSSION

A newly described optimisation outline for building spatial designs in early stages of building design has been shown to make effective use of a mixed-integer super-structure representation. The optimisation process was shown to converge for both small and medium sized building spatial designs for two individual objectives, compliance and surface area. Moreover, different constraint handling techniques were applied to significantly improve the ability to traverse a search space in search of valid designs. Both repair procedures and graded penalties were used. Altogether these results form a promising first step in multi-disciplinary design optimisation.

A problem that was encountered is that different runs found different topologies. A possible explanation is that transitions from one feasible subspace to another are very unlikely. In future work this should be addressed by global optimisation strategies such as niching [10]. Another issue is that even the largest problem configurations considered here are not very large compared to most practical building spatial designs. Given a significant portion of the constraints are violated with frequencies in the 40 to 50% range and in the worst cases over 80%, even when using the multi-penalty approach, it is clear that for larger designs more effective constraint handling methods have to be introduced.

Acknowledgements: The authors gratefully acknowledge the financing of this project by the Dutch STW via project 13596 (*Excellent Buildings via Forefront MDO, Lowest Energy Use, Optimal Spatial and Structural Performance*).

REFERENCES

[1] Matti Palonen, Mohamed Hamdy, and Ala Hasan. Mobo a new software for multi-objective building performance optimization. In *Proceedings of the 13th Internationcal Conference of the IBPSA*, pages 2567–2574, 2013.

[2] Hèrm Hofmeyer and Juan Manuel Davila Delgado. Coevolutionary and genetic algorithm based building spatial and structural design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 29(04):351–370, 2015.

[3] Shady Attia, Elisabeth Gratia, André De Herde, and Jan L.M. Hensen. Simulation-based decision support tool for early stages of zero-energy building design. *Energy and Buildings*, 49:2–15, 2012.

[4] Christina J. Hopfe, Michael T.M. Emmerich, Robert Marijt, and Jan L.M. Hensen. Robust multi-criteria design optimisation in building design. *Proceedings of Building Simulation and Optimization, Loughborough, UK*, pages 118–125, 2012.

[5] Thomas Bäck, David B. Fogel, and Zbigniew Michalewicz. *Handbook of Evolutionary Computation*. IOP Publishing Ltd., 1997.

[6] Ingo Rechenberg. *Evolutionsstrategie '94, volume 1 of Werkstatt Bionik und Evolutionstechnik*. Frommann Holzboog, Stuttgart, 1994.

[7] Rui Li, Michael T.M. Emmerich, Jeroen Eggermont, Thomas Bäck, Martin Schütz, Jouke Dijkstra, and Johan H.C. Reiber. Mixed integer evolution strategies for parameter optimization. *Evolutionary computation*, 21(1):29–64, 2013.

[8] Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies–a comprehensive introduction. *Natural computing*, 1(1):3–52, 2002.

[9] Sjonnie Boonstra, Koen van der Blom, Hèrm Hofmeyer, Robert Amor, and Michael T.M. Emmerich. Super-structure and super-structure free design search space representations for a building spatial design in multi-disciplinary building optimisation. Submitted for EG-ICE 2016, Poland June 29 – July 1, 2016.

[10] Ofer M. Shir, Michael T.M. Emmerich, and Thomas Bäck. Adaptive niche radii and niche shapes approaches for niching with the CMA-ES. *Evolutionary Computation*, 18(1):97–126, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

BENDING STIFFNESS OF A MULTILAYERED PLATE

Petr E. Tovstik[1], **Tatiana M. Tovstik**[2]

[1,2]Sankt-Petersburg State University, Universitetskaya nab. 7/9, Sankt-Petersburg, Russian Federation
e-mail: peter.tovstik@mail.ru

Keywords: Multilayered Plate, Bending Stiffness, Asymptotic Methods, Monte Carlo Method.

Abstract. *A thin elastic multilayered plate consisting of alternating hard and soft isotropic layers is studied. One of the face planes is subject to a normal pressure and the other face plane is free. A formula for the deflection of an infinite plate under a doubly periodic external force is delivered using asymptotic expansions. This formula is also applied for a rectangular plate with Navier boundary conditions on its edges. The maximal deflection is accepted as a measure of the plate stiffness. The purpose of the present paper is to obtain an expression for the plate deflection and find an optimal distribution of hard and soft layers assuming that their total thicknesses are given. The Monte Carlo method is used for finding the optimal distribution of layers.*

1 INTRODUCTION

The 2D classical equation for a plate was obtained by Sophie Germain in 1808 with the purpose of explaining Chladni figures. The equation of a plate bending can be obtained on the basis of the Kirchhoff–Love (KL) hypotheses [1, 2]. The more involved and sometimes more exact equation, which takes into account the transversal shear, follows from the Timoshenko–Reissner (TR) hypotheses [3, 4].

The 2D models of plates and shells are chiefly based on the 3D equations of the theory of elasticity. The methods of unknown functions expansions in series of Legendre polynomials in the thickness direction were used in [5, 6]. Numerous investigations [7–9] were devoted to the derivation of 2D equations by using asymptotic expansions in power series in the small parameter $\mu = h/L$, which is equal to the dimensionless plate thickness (h and L are, respectively, the thickness and the typical wave length in the tangential directions). A different approach [10,11] rests on the direct derivation of 2D equations of plates and shells without referring to a 3D media. An account of general problems of the plates theory may be found in the books [12–14].

The present paper is concerned with a thin plate of constant thickness made of a linearly elastic material that is transversally isotropic and heterogeneous in the thickness direction. For a transversally isotropic material the accepted simplification [15], for which a 3D system of sixth order of the theory of elasticity splits into systems of second and fourth orders, is possible. Asymptotic expansions in powers of the small thickness parameter μ are constructed, the bending equation of second-order accuracy (the SA model) is obtained. The origin of this paper is the paper [16], in which an isotropic homogeneous plate is studied, and the paper [8], in which a heterogeneous plate is briefly examined. For a multi-layer plate with alternating hard and soft layers, an explicit formula based on the SA model for the maximum deflection is delivered. The results of the SA model are compared with the KL classical model and also with the exact numerical solution. The ratio $\eta = E_2/E_1$ of the soft and hard Young moduli changes in a very wide range ($0.0001 \le \eta \le 1$).

The aforementioned formula is obtained for an infinite in the tangential directions plate subject to a doubly periodic external load. After employing the Fourier expansions this formula is applied to a rectangular plate with the Navier boundary conditions.

For a multi-layer plate with alternating hard and soft layers, the distribution of layer thicknesses, which gives the maximum bending stiffness, is found. Here it is assumed that the summary thicknesses of the hard and of the soft layers are given. This problem is attacked by the Monte Carlo method. For a very small ratio η of elastic moduli, the three-layer plate with the soft layer lying between two hard layers was found not to be optimal. In this case, the plate with the maximum bending stiffness is multi-layered one with the number of layers $n > 3$.

2 EQUILIBRIUM EQUATIONS AND THEIR SIMPLIFICATION

Consider the linear bending problem of a thin plate made of a transversally isotropic heterogeneous material. The 3D equilibrium equations are

$$\frac{\partial \sigma_{ij}}{\partial x_j} + f_i = 0, \qquad i, j = 1, 2, 3, \qquad 0 \le x_3 = z \le h, \tag{1}$$

where x_j are the Cartesian coordinates, f_i are the projections of the external load intensity, the summation being carried out over repeating subscripts.

The stresses σ_{ij} are related to the strains ε_{ij} as follows:

$$\begin{aligned}
&\sigma_{11} = E_{11}\varepsilon_{11} + E_{12}\varepsilon_{22} + E_{13}\varepsilon_{33}, && \sigma_{12} = G_{12}\varepsilon_{12},\\
&\sigma_{22} = E_{12}\varepsilon_{11} + E_{11}\varepsilon_{22} + E_{13}\varepsilon_{33}, && \sigma_{13} = G_{13}\varepsilon_{13},\\
&\sigma_{33} = E_{13}\varepsilon_{11} + E_{13}\varepsilon_{22} + E_{33}\varepsilon_{33}, && \sigma_{23} = G_{13}\varepsilon_{23},\\
&\varepsilon_{11} = \frac{\partial u_1}{\partial x_1}, \qquad \varepsilon_{12} = \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}, \quad etc.
\end{aligned} \tag{2}$$

Here, $E_{11} = E_{12} + 2G_{12}$, u_i are deflections. The elastic moduli E_{ij}, G_{ij} are independent of the tangential coordinates x_1, x_2, but they may depend on the transversal coordinate $x_3 = z$. For functionally gradient materials the moduli are continuous functions in z, and for multi-layered plates they are piecewise continuous functions.

For an isotropic material

$$E_{11}=E_{33}=\frac{E(1-\nu)}{(1+\nu)(1-2\nu)}, \quad E_{12}=E_{13}=\frac{E\nu}{(1+\nu)(1-2\nu)}, \quad G_{12}=G_{13}=G=\frac{E}{2(1+\nu)}, \tag{3}$$

where E and ν are the Young's modulus and the Poisson ratio, respectively.

We set the homogeneous boundary conditions on the face planes $z = 0$ and $z = h$

$$\sigma_{i3} = 0, \quad i = 1, 2, 3. \tag{4}$$

If the surface forces are given, then they are included in the body forces by using the Dirac delta-function.

Introduce new unknown functions u, v, σ, τ as

$$\begin{aligned}
&u = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}, && v = \frac{\partial u_1}{\partial x_2} - \frac{\partial u_2}{\partial x_1},\\
&\sigma = \frac{\partial \sigma_{13}}{\partial x_1} + \frac{\partial \sigma_{23}}{\partial x_2}, && \tau = \frac{\partial \sigma_{13}}{\partial x_2} - \frac{\partial \sigma_{23}}{\partial x_1}.
\end{aligned} \tag{5}$$

For a transversally isotropic material the system (1), (2) is split into two subsystems [15]:

$$\frac{\partial \tau}{\partial z} + G_{12}\Delta v + m_1 = 0, \quad \tau = G_{13}\frac{\partial v}{\partial z}, \qquad \Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}, \quad m_1 = \frac{\partial f_1}{\partial x_2} - \frac{\partial f_2}{\partial x_1}; \tag{6}$$

$$\begin{aligned}
&\sigma_{33} = E_{13}u + E_{33}\frac{\partial w}{\partial z}, && \sigma = G_{13}\left(\frac{\partial u}{\partial z} + \Delta w\right), && w = u_3,\\
&\frac{\partial \sigma}{\partial z} + E_0\Delta u + \frac{E_{13}}{E_{33}}\Delta\sigma_{33} + m = 0, && \frac{\partial \sigma_{33}}{\partial z} + \sigma + f_3 = 0, && m = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2},
\end{aligned} \tag{7}$$

with

$$E_0 = E_{11} - \frac{E_{13}^2}{E_{33}} = \frac{E}{1-\nu^2}.$$

System (6) of the second differential order in z describes the boundary layer and will not be studied here.

The fourth order system (7) describes the plate bending. The 2D plate model is obtained here by using asymptotic expansions in powers of the small parameter $\mu = h/L$ [8,17].

We introduce the dimensionless variables (denoted with hats)

$$\{u_1, u_2, w, z\} = h\{\hat{u}_1, \hat{u}_2, \hat{w}, \hat{z}\}, \quad \{x_1, x_2\} = L\{\hat{x}_1, \hat{x}_2\}, \quad u = \mu\hat{u},$$
$$\{\sigma_{ij}, \sigma, E_{ij}, G_{ij}, E_0\} = E_*\{\hat{\sigma}_{ij}, \hat{\sigma}, \hat{E}_{ij}, \hat{G}_{ij}, c_0\}, \quad f_i = \frac{E_*}{h}\hat{f}_i, \quad E_* = \frac{1}{h}\int_0^h E_0(z)dz, \tag{8}$$

and rewrite system (7) in the dimensionless form

$$\frac{\partial\hat{w}}{\partial\hat{z}} = -\mu c_\nu\hat{u} + c_3\hat{\sigma}_{33}, \qquad \frac{\partial\hat{u}}{\partial\hat{z}} = -\mu\hat{\Delta}\hat{w} + c_g\hat{\sigma},$$
$$\frac{\partial\hat{\sigma}}{\partial\hat{z}} = Y_3(\hat{z}) = -\mu^2 c_0\hat{\Delta}\hat{u} - \mu c_\nu\hat{\Delta}\hat{\sigma}_{33} - \hat{m}, \qquad \frac{\partial\hat{\sigma}_{33}}{\partial\hat{z}} = Y_4(\hat{z}) = -\mu\hat{\sigma} - \hat{f}_3, \quad 0 \le \hat{z} \le 1, \tag{9}$$

with

$$c_\nu = \frac{E_{13}}{E_{33}} = \frac{\nu}{1-\nu}, \quad c_3 = \frac{E_*}{E_{33}}, \quad c_g = \frac{E_*}{G_{13}},$$
$$\hat{\sigma} = \frac{\partial\hat{\sigma}_{13}}{\partial\hat{x}_1} + \frac{\partial\hat{\sigma}_{23}}{\partial\hat{x}_2}, \quad \hat{m} = \frac{\partial\hat{f}_1}{\partial\hat{x}_1} + \frac{\partial\hat{f}_2}{\partial\hat{x}_2}, \quad \hat{\Delta} = \frac{\partial^2}{\partial\hat{x}_1^2} + \frac{\partial^2}{\partial\hat{x}_2^2}. \tag{10}$$

Here, L is the typical value of the wave length in the tangential directions, and E_* is the average value of the modulus E_0. The dimensionless coefficients c_0, c_ν, c_g, c_3 are the given functions of $\hat{z}$. From the boundary conditions (4) we get

$$\hat{\sigma} = \hat{\sigma}_{33} = 0 \quad at \quad \hat{z} = 0 \quad and \quad \hat{z} = 1. \tag{11}$$

In what follows, the hat sign will be omitted.

3 ASYMPTOTIC SOLUTION OF THE BOUNDARY VALUE PROBLEM (9), (11)

Assume that the dimensionless external forces f_3, m are of the order of unity. Then the orders of the unknown functions are

$$\sigma_{33} = O(1), \quad \sigma = O(\mu^{-1}), \quad u = O(\mu^{-3}), \quad w = O(\mu^{-4}). \tag{12}$$

The right-hand sides of Eqs. (9) are small, and the method of iterations [10,18] is used. To construct the solution of second-order accuracy we seek it as

$$w = \mu^{-4}w_0 + \mu^{-2}w_2, \quad u = \mu^{-3}u_0 + \mu^{-1}u_2, \quad \sigma = \mu^{-1}\sigma_0 + \mu\sigma_2, \quad \sigma_{33} = \sigma_{33,0} + \mu^2\sigma_{33,2}. \tag{13}$$

The arbitrary functions $w^0(x_1, x_2)$ and $u^0(x_1, x_2)$ appear after the integration in z of the first two equations (9). These functions are found from the compatibility conditions of the remaining two equations (9) and the boundary conditions (11) [17]

$$\langle Y_3(z)\rangle = 0, \quad \langle Y_4(z)\rangle = 0, \qquad \langle Z(z)\rangle \equiv \int_0^1 Z(z)dz. \tag{14}$$

In the zero approximation we get

$$w_0 = w_0(x_1, x_2), \quad u_0 = (a - z)\Delta^2 w_0, \quad a = \langle zc_0(z)\rangle,$$
$$\sigma_0 = \varphi_1(z)\Delta^2 w_0^0, \quad \varphi_1(z) = \int_0^z c_0(z)(z - a)dz,$$
$$D\Delta^2 w_0 = F_3, \qquad D = \langle (z - a)^2 c_0(z)\rangle, \quad F_3 = \langle f_3(z)\rangle,$$
$$\sigma_{33,0} = -\frac{F_3}{D}\varphi_2 - \varphi_3, \quad \varphi_2(z) = \int_0^z \varphi_1(z)dz, \quad \varphi_3(x_1, x_2, z) = \int_0^z f_3(x_1, x_2, z)dz, \tag{15}$$

where $z = a$ is the position of the plate neutral layer, D is the bending stiffness of a plate with the variable elastic moduli, F_3 is the full transversal force. The equation $D\Delta^2 w^0 = F_3$ corresponds to the classical KL model.

In the second approximation the solution is more unwieldy. Here, we give only the function w_2, which depends on z. At $z = 0$ it satisfies the equation

$$D\Delta^2 w_2(0) = A\Delta F_3 + N(\Delta f_3) - M, \qquad A = A_g - A_\nu, \tag{16}$$

where

$$\begin{aligned} N(\Delta f_3) &= \int_0^1 c_\nu(z)(a-z)\left(\int_0^z \Delta f_3 dz_1\right) dz, \quad M = \int_0^1 (a-z)m(z)dz, \\ A_g &= \frac{1}{D}\int_0^1 c_0(z)(z-a)\int_0^z c_g(z_1)\int_0^{z_1} c_0(z_2)(z_2-a)dz_2dz_1dz, \\ A_\nu &= \frac{1}{D}\int_0^1 (z-a)\int_0^z \int_0^{z_1} \left(c_\nu(z)c_0(z_2) + c_0(z)c_\nu(z_2)\right)(z_2-a)dz_2dz_1dz. \end{aligned} \tag{17}$$

The full deflection of the reference plane $z = 0$ satisfies the equation

$$D\mu^4\Delta^2 w(0) = F_3 + \mu^2(A\Delta F_3 + L(\Delta f_3) - M) + O(\mu^4), \tag{18}$$

in which the coefficients D and A depend of the elastic moduli distribution in the plate thickness, the summands $L(\Delta f_3)$ and M depend on the distribution of external transversal and tangential loads, respectively. The summands A_g and A_ν take into account the transversal shear and the Poisson's strains, respectively.

The full deflection $w(z)$ of the arbitrary plane z is expressed through $w(0)$ as

$$w(z) = w(0) + \mu^2\Delta w(0)\int_0^z c_\nu(z)(z-a)dz. \tag{19}$$

4 MULTILAYERED PLATE UNDER A NORMAL PRESSURE

Consider a multi-layered plate with hard and soft isotropic homogeneous layers under a normal pressure F_3 acting on the plane $z = 0$. In this case, in Eq. (18) $L(\Delta f_3) = \langle c_\nu(a-z)\rangle\,\Delta F_3$ and $M = 0$, and so Eq. (18) assumes the form

$$D\mu^4\Delta^2 w(0) = F_3 + \mu^2 A_1\Delta F_3, \quad A_1 = A + \int_0^1 c_\nu(z)(z-a)dz. \tag{20}$$

If the plate is homogeneous one-layered, then the elastic moduli c_0, c_g, c_ν are constant, and so the integrals in (17) can be taken, and so in Eq. (20) we have

$$D\mu^4\Delta^2 w(0) = F_3 + \mu^2 A_1\Delta F_3, \quad A_1 = \frac{2c_\nu - c_g}{10} = -\frac{1}{5}, \quad D = \frac{1}{12}. \tag{21}$$

Using the neutral plane $z = 1/2$ as a reference one due to Eq. (19) we get

$$D\mu^4\Delta^2 w(1/2) = F_3 + \mu^2 A_2\Delta F_3, \quad A_2 = \frac{3c_\nu - 4c_g}{40} = \frac{3\nu - 8}{40(1-\nu)}. \tag{22}$$

This value A_2 was obtained in [16] and was repeated in [8].

Consider a plate of thickness h consisting of $n = 2n_0 + 1$ homogeneous isotropic layers of thickness $h_k,\ k = 1, 2, ..., n$ ($h = \sum h_k$). Let E_1, ν_1 and E_2, ν_2 ($E_2/E_1 \le 1$) be the Young's moduli and the Poisson ratios of hard layers with odd numbers and of soft layers with even numbers, respectively. We set

$$z_0 = 0, \quad z_k = \sum_{i=1}^{k} h_i, \quad k = 1, \dots, n. \tag{23}$$

The elastic moduli are known to be piecewise functions in z, and so we put

$$\begin{aligned} e_k &= \frac{E_1}{1-\nu_1^2}, \quad c_k = \frac{\nu_1}{1-\nu_1}, \quad g_k = \frac{E_1}{2(1+\nu_1)}, && k\ is\ odd, \\ e_k &= \frac{E_2}{1-\nu_2^2}, \quad c_k = \frac{\nu_2}{1-\nu_2}, \quad g_k = \frac{E_2}{2(1+\nu_2)}, && k\ is\ even. \end{aligned} \tag{24}$$

Calculating the integrals in Eqs. (15) and (17) we find the coordinate $z = a$ of the neutral layer, the bending stiffness D according to the KL model, and the coefficients A_g and A_ν in Eqs. (17) as follows:

$$\begin{aligned} a &= \frac{1}{2}\sum_{k=1}^{n} e_k(z_k^2 - z_{k-1}^2)\left(\sum_{k=1}^{n} e_k h_k\right)^{-1}, \quad D = \frac{1}{3}\sum_{k=1}^{n} e_k(\hat z_k^3 - \hat z_{k-1}^3), \quad \hat z_k = z_k - a, \\ A_g &= \frac{1}{D}\sum_{k=1}^{n}\left(\frac{e_k f_{1k}}{2}(\hat z_k^2 - \hat z_{k-1}^2) + \frac{f_{2k}}{3g_k}(\hat z_k^3 - \hat z_{k-1}^3) + \frac{e_k}{30g_k}(\hat z_k^5 - \hat z_{k-1}^5)\right), \\ A_\nu &= \frac{1}{D}\sum_{k=1}^{n}\left(\frac{e_k f_{3k}^c + c_k f_{3k}^e}{2}(\hat z_k^2 - \hat z_{k-1}^2) + (c_k f_{4k}^e + e_k f_{4k}^c)(\hat z_k^3 - \hat z_{k-1}^3) + \frac{e_k c_k}{15}(\hat z_k^5 - \hat z_{k-1}^5)\right), \end{aligned} \tag{25}$$

where

$$\begin{aligned} f_{1k} &= \sum_{i=1}^{k-1}\left(\frac{f_{2i}h_i}{g_i} + \frac{e_i}{6g_i}(\hat z_i^3 - \hat z_{i-1}^3)\right) - \frac{f_{2k}h_k}{g_k} - \frac{e_k}{6g_k}\hat z_k^3, \\ f_{2k} &= \frac{1}{2}\sum_{i=1}^{k-1} e_i(\hat z_i^2 - \hat z_{i-1}^2) - \frac{1}{2}e_k\hat z_k^2, \\ f_{3k}^e &= \sum_{i=1}^{k-1}\left(f_{4i}^e h_i + \frac{e_i}{6}(\hat z_i^3 - \hat z_{i-1}^3)\right) - f_{4k}^e h_k - \frac{e_k}{6}\hat z_k^3, \\ f_{4k}^e &= \frac{1}{2}\sum_{i=1}^{k-1} e_i(\hat z_i^2 - \hat z_{i-1}^2) - \frac{1}{2}e_k\hat z_k^2, \quad (e \to c). \end{aligned} \tag{26}$$

Here, $(e \to c)$ means that the similar formulae also hold for the moduli c_k.

Assume, at first, that the plate is infinite ($-\infty < x_1, x_2 < \infty$), and consider a doubly periodic normal pressure $F_3 = F_3^0 \sin r_1 x_1 \sin r_2 x_2$. Then the deflection $W(x_1, x_2, z) = W(z) \sin r_1 x_1 \sin r_2 x_2$ is also doubly periodic. We rewrite Eq. (20) as

$$Dr^4 W(0) = (1 - r^2 A_1) F_3^0, \tag{27}$$

where

$$A_1 = A_g - A_\nu - \frac{1}{2}\sum_{k=1}^{n} c_k(\hat z_k^2 - \hat z_{k-1}^2), \qquad r^2 = r_1^2 + r_2^2, \tag{28}$$

and the coefficients $D,\ A_\nu,\ A_g$ are given in Eqs. (25).

In the dimensionless form (with hats)

$$W(0) = hw(0), \quad D = h^3\hat{D}, \quad A_1 = h^2\hat{A}_1, \quad \mu = rh \tag{29}$$

Equation (27) gives the deflection of the second-order accuracy (SA)

$$w(0)^{SA} = w^{KL}(1 - \mu^2\hat{A}_1), \quad w^{KL} = \frac{F_3^0 h}{\mu^4\hat{D}}, \tag{30}$$

where w^{KL} is the deflection in the KL model.

A comparison of the expressions $\mu = rh$ and $\mu = h/L$ leads to the conclusion that in this problem the typical wave length L should be taken as

$$L = \frac{1}{r} = \frac{1}{\sqrt{r_1^2 + r_2^2}}. \tag{31}$$

To discuss the errors of the approximate expressions (30) for w^{KL} and $w(0)^{SA}$ we require some numerical examples. We consider a multi-layered plate with $n = 11$ layers and $h = 1$, $\nu_1 = \nu_2 = 0.3$, $F_3^0 = 1$, and take $h_k = 1/18,\ E_k = 1,\ k = 2i - 1$, i=1,...,6, for the hard layers, and $h_k = 2/15,\ E_k = \eta,\ k = 2i,\ i = 1, \ldots, 5$, for the soft layers. Such a plate will be referred to as a plate with uniform thickness distribution. We take two values of the relative thickness parameter: $\mu = 0.3776$ and $\mu = 0.2221$. The ratio $\eta = E_2/E_1$ of the elastic moduli will vary in the range $0.0001 \le \eta \le 1$. We compare the approximate values (30) with the exact value w^e found by numerical solution of Eqs. (9). The results are given in Table 1. Columns 2–4 correspond to $\mu = 0.3776$, and columns 5–7, to $\mu = 0.2221$.

1	2	3	4	5	6	7
η	w^e	w^{KL}/w^e	w^{SA}/w^e	w^e	w^{KL}/w^e	w^{SA}/w^e
1.0	553	0.9717	0.9994	4529	0.9902	0.9999
0.31	936	0.9539	0.9992	7578	0.9837	0.9999
0.1	1270	0.8899	0.9985	9830	0.9532	0.9998
0.031	1684	0.7322	0.9975	11587	0.8882	0.9996
0.01	2700	0.4703	0.9987	14723	0.7198	0.9998
0.0031	5746	0.2231	1.0110	23692	0.4510	1.0036
0.001	14258	0.0800	1.0441	51150	0.2098	1.0192
0.00031	41793	0.0308	1.1156	134232	0.0800	1.0557
0.0001	110822	0.0116	1.3054	374835	0.0287	1.1337

Table 1: Error of the approximate models for some values of η with $\mu = 0.3776$ and $\mu = 0.2221$.

The relative error of the KL and SA models is found by comparing the results in columns 3,4,6,7 with one. The exactness of the KL model is of first asymptotical order with respect to μ, and it is acceptable for applications only for $0.1 \le \eta \le 1$. The SA model of second asymptotical order may be used in the very wide range $0.001 \le \eta \le 1$. For $0.1 \le \eta \le 1$ the SA model is essentially more exact than the TR model. For $0.0001 \le \eta \le 0.001$ the 2D models are unacceptable.

The errors of the 2D models decrease simultaneously with the thickness parameter μ. This follows from the comparison of columns 3 and 6, and also of columns 4 and 7.

5 DEFLECTION OF A RECTANGULAR MULTILAYERED PLATE

Consider a rectangular multi-layer plate with $0 \le x_1 \le a_1,\ 0 \le x_2 \le a_2,\ 0 \le z \le h$ (Fig. 1). In the previous sections the boundary conditions at the plate edges were either not imposed or a plate was assumed to be infinite in the tangential directions. Now the following variant of boundary conditions will be accepted:

$$\begin{aligned} &u_2 = w = \sigma_{11} = 0 \qquad at \quad x_1 = 0, \quad x_1 = a_1, \\ &u_1 = w = \sigma_{22} = 0 \qquad at \quad x_2 = 0, \quad x_2 = a_2, \end{aligned} \tag{32}$$

(the so-called Navier conditions) and

$$\sigma_{13} = \sigma_{23} = 0,\ \sigma_{33} = -F_3^0 \quad at \quad z = 0; \qquad \sigma_{13} = \sigma_{23} = \sigma_{33} = 0 \quad at \quad z = h. \tag{33}$$

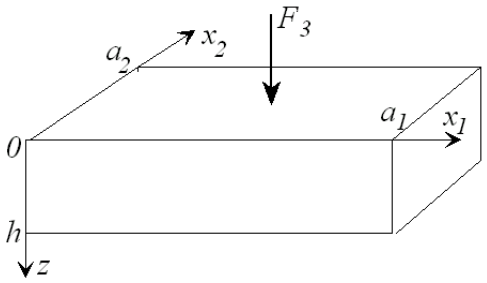


Fig. 1. A plate.

The functions

$$\begin{aligned} &u_1(x_1, x_2, z) = r_{1m}\, u(z) \cos(r_{1m}x_1) \sin(r_{2n}x_2), \\ &u_2(x_1, x_2, z) = r_{2n}\, u(z) \sin(r_{1m}x_1) \cos(r_{2n}x_2), \quad r_{1m} = \frac{m\pi}{a_1},\ r_{2n} = \frac{n\pi}{a_2}, \quad m, n = 1, 2, \ldots, \\ &w(x_1, x_2, z) = w(z) \sin(r_{1m}x_1) \sin(r_{2n}x_2), \end{aligned} \tag{34}$$

satisfy the boundary conditions (32). The boundary layer does not appear, because $v = \partial u_2/\partial x_1 - \partial u_1/\partial x_2 \equiv 0$ (see Eqs. (5)).

Eq. (30) for the deflection is delivered for a doubly periodic external force $F_3 = F_3^0 \sin(r_1 x_1) \sin(r_2 x_2)$ with $\mu = h\sqrt{r_1^2 + r_2^2}$. To use Eq. (30) for a constant force F_3 in the rectangle $0 \le x_1 \le a_1,\ 0 \le x_2 \le a_2$ we expand it in a double Fourier series

$$F_3 = F_3 \sum_{m,n=1,3,\ldots} \frac{16}{mn\pi^2} \sin(r_{1m}x_1) \sin(r_{2n}x_2), \tag{35}$$

and apply Eq. (30) to the each summand of (35). In this case

$$\mu = \mu_{mn} = h\sqrt{\left(\frac{m\pi}{a_1}\right)^2 + \left(\frac{n\pi}{a_2}\right)^2}, \quad m, n = 1, 2, \ldots \tag{36}$$

and so the full deflection is

$$w(x_1, x_2, 0) = \frac{F_3}{D} \sum_{m,n=1,3,\ldots} C_{mn} \sin(r_{1m}x_1) \sin(r_{2n}x_2), \qquad C_{mn} = \frac{16\,(1 - \mu_{mn}^2 A_1)}{mn\pi^2 \mu_{mn}^4}, \tag{37}$$

where D and A_1 are given in (25) and (28), respectively. The similar expression is valid for the deflection of the opposite face plane $w(x_1, x_2, h)$. In this case, the constant A_1 in Eq. (37) needs to be replaced by $A = A_g - A_\nu$.

The deflection of the line $x_2 = a_2/2,\ z = 0$ and the maximum deflection at $z = 0$ are, respectively,

$$\begin{aligned} w(x_1, a_2/2, 0) &= \frac{F_3}{D}\sum_{k=0}^{\infty} B_{2k+1}\sin(r_{1,2k+1}x_1), \quad B_{2k+1} = \sum_{j=0}^{\infty}(-1)^j C_{2k+1,2j+1}, \\ w(a_1/2, a_2/2, 0) &= \frac{F_3}{D}A_0, \quad A_0 = \sum_{k,j=0}^{\infty}(-1)^{k+j}C_{2k+1,2j+1} = \sum_{k=0}^{\infty}(-1)^k B_{2k+1}. \end{aligned} \tag{38}$$

All the series in (37) and (38) converge rapidly, the summands with C_{11} are much larger than the remaining summands (see Table 4).

As an example, we consider a multi-layer square plate with $n = 11$ and with the same data as in Section 4. We assume in addition that $a_1 = a_2 = 20$, and so $\mu = \mu_{11} = 0.2221$. For various values of $\eta = E_2/E_1$, the first coefficients $C_{2k+1,2j+1}$, B_{2k+1}, and A_0 are given in Table 2.

η	C_{11}	C_{13}	C_{33}	C_{15}	C_{35}	C_{55}	B_1	B_3	B_5	A_0
1	4528	191	62	31	19	10	4368	160	22	4230
0.1	9828	463	164	89	58	32	9454	388	63	9129
0.01	14720	1255	593	383	282	183	13848	1045	284	13087
0.001	52134	8717	4737	3251	2475	1675	46668	6455	2451	42669

Table 2: Coefficients $C_{2k+1,2j+1}$, B_{2k+1}, and A_0 for various values η.

The amplitude of deflection is proportional to A_0; it essentially depends on the ratio η (see Table 2), but the deflection mode depends slightly on η and is close to the function $\sin(\pi x_1/a_1)$ (see Fig. 2).

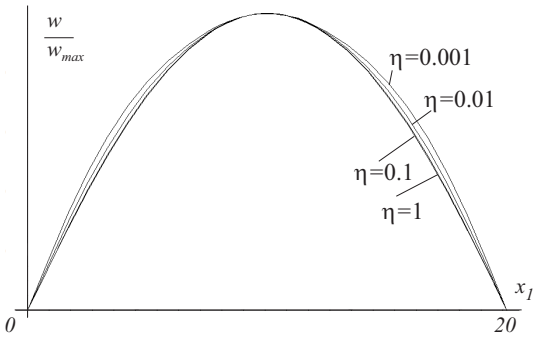


Figure 1: Deflection modes for various values of the ratio η of the Young moduli.

6 OPTIMAL DISTRIBUTION OF THE LAYER THICKNESSES

Consider a multilayer plate with the alternating hard and soft layers. Let the entire number of layers be $n = 2n_0 + 1$ with $n_0 + 1$ layers and n_0 soft layers. We set

$$h_h = \sum_{k=0}^{n_0} h_{2k+1}, \quad h_s = \sum_{k=1}^{n_0} h_{2k}, \quad h = h_h + h_s, \tag{39}$$

where h_h and h_s are the summary thicknesses of the hard and of the soft layers, respectively. Assume that the values h_h, h_s, also the elastic moduli of hard and of soft layers, and the thickness parameter μ are given. We seek the distribution of the layer thicknesses h_k satisfying Eqs. (39), for which the plate bending stiffness is maximum. As the bending stiffness we accept the inverse value the maximum deflection $f(h_1, h_2, \ldots, h_n) = w(0)^{SA}$ (see Eq. (30)).

In most cases a three-layer plate with

$$n = 3, \quad h_1 = h_3 = h_h/2, \quad h_2 = h_s \tag{40}$$

has the maximum stiffness. But if the ratio η of elastic moduli is very small, then the other distribution of thicknesses may lead to the maximum stiffness. Here we investigate this problem in details for the fixed values $\mu = 0.3776,\ \nu_1 = \nu_2 = 0.3$, for two changing parameters η and $s = h_h/h$, and for $n \le 11$ (without loss of generality we may take $E_1 = 1$ and $h = 1$). The small parameter $\mu = 0.3776$ corresponds, in partial, to a rectangular plate with $a_1 = 10,\ a_2 = 15$.

To find the thickness distribution corresponding to the maximum stiffness for the fixed values η and s, we seek the minimum of function $f(h_k) = f(h_1, h_2, \ldots, h_n)$ in h_k satisfying Eqs. (39). To find h_k^0 that delivers the minimum to the function $f(h_k)$ we use the Monte Carlo method in combination with the method of iterations [18]. Thanks to Eqs. (39) the point h_k moves in the space R^{n-2}. Let $h_k^{(i)}$ be some point. We next seek the next point $h_k^{(i+1)}$ for which

$$f(h_k^{(i+1)}) < f(h_k^{(i)}) \tag{41}$$

as follows. We put $h_k^* = \max\{0, h_k^{(i)} + \xi_k\}$, where ξ_k are the uniformly distributed numbers in $[-\varepsilon, \varepsilon]$ (for example, $\varepsilon = 0.01$). Then we normalize the values h_k^* according to Eqs. (39)

$$h_{2i+1}^{**} = \frac{s\, h_{2i+1}^*}{\sum_{i=0}^{5} h_{2i+1}^*}, \qquad h_{2i}^{**} = \frac{(1-s) h_{2i}^*}{\sum_{i=1}^{5} h_{2i}^*}. \tag{42}$$

If $f(h_k^{**}) < f(h_k^{(i)})$, then we take $h_k^{(i+1)} = h_k^{**}$. Otherwise, we repeat the previous step with different random numbers ξ_k.

As the initial distribution we take $h_1^{(1)} = h_{11}^{(1)} = s/2,\ h_6^{(1)} = (1-s)$, assuming that the rest of $h_k^{(1)}$ is 0. Calculations show that it is enough to make 10^6 steps.

The curve tn Fig. 3 divides the plane (s, η) of the problem parameters into two parts; in one of which the three-layer plate has maximum stiffness.

Consider in more detail the case $s = 1/3,\ \mu = 0.3776$ with various values of the parameter η. For $\eta > \eta_* = 0.0122$ the three-layered plate (40) has maximal bending stiffness. In Table 3 for various η the minimal deflection, W_o, as found by the algorithm described above, is compared with the deflection, W_u, of a plate with the uniform thickness distribution (see Section 4) and with the deflection, W_3, of the three-layered plate (40). For $\eta > \eta_*$ the three-layered plate is optimal, $W_3 = W_o$. For $\eta < 0.006$ the stiffness of a plate with uniform thickness distribution is larger than the stiffness of a three-layer plate, $W_u < W_3$.

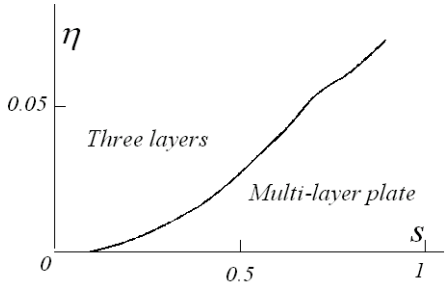


Figure 2: The plane (s, η) for $\mu = 0.3776$.

η	uniform, W_u	three-layered, W_3	optimal, W_o
0.001	15617	17812	11539
0.002	8446	9286	7815
0.005	4139	4170	4007
0.01	2697	2464	2458
0.0122	2434	2155	2155
0.02	1962	1607	1607
0.1	1268	901	901
0.5	785	663	663

Table 3: Comparison of the stiffness parameters for various multilayer plates with $\mu = 0.3776,\ s = 1/3$.

In Table 4 the optimal distribution of thicknesses for $s = 1/3,\ \mu = 0.3776$, and for four values of η is presented. The thicknesses of soft layers are shown in italic. For $\eta = 0.01$ the distribution is symmetric with respect to the mid-plane. With smaller η the distribution is not symmetric. The resulting asymmetry can be accounted for by the fact that the value W_o is the deflection of the plane $z = 0$ to which the external force is applied. At $\eta = 0.001$ the optimal plate degenerates to a three-layer plate with very dissimilar distribution of layers.

η	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
0.01	0.149	*0.192*	0.007	*0.099*	0.010	*0.081*	0.010	*0.099*	0.007	*0.192*	0.149
0.005	0.094	*0.212*	0.025	*0.090*	0.048	*0.063*	0.047	*0.090*	0.025	*0.211*	0.095
0.002	0.030	*0.255*	0.024	*0.075*	0.213	*0.039*	0.027	*0.083*	0.011	*0.215*	0.029
0.001	0.331	*0.667*	0.002	–	–	–	–	–	–	–	–

Table 4: Distribution of thicknesses of the optimal plate with $s = 1/3,\ \mu = 0.3776$.

The similar results also hold for a rectangular plate with the Navier boundary conditions, because the first summand in the expression for $w(x_1, a_2/2, 0)$ in Eq. (38) is much larger than the rest summands.

The results of Section 6 are approximate, since they are based on the approximate equation (30). The error in Eq. (30) is estimated in Section 4 (see Table 1). The exact results may be obtained by numerical solution of Eqs. (9). But for the large number of iterations this approach takes a long time, though the result is almost the same.

7 CONCLUSIONS

- A two-dimensional linear model of second-order accuracy (SA) is used for analysis of a thin multi-layer plate bending.
- The formula for an infinite plate deflection subject to a doubly periodic external force is obtained. This formula is applied to a rectangular multi-layer plate with alternating hard and soft layers and with the Navier boundary conditions.
- The plate deflection depends on the distribution of thicknesses of hard and soft layers.
- The problem of the optimal distribution, which gives the maximum bending stiffness, is attacked using the Monte Carlo method.
- For a very small ratio η of the elastic moduli the three-layer plate with the soft layer lying between two hard layers was found not to be optimal. In this case, the plate with the maximum bending stiffness is the multi-layered one with the number of layers $n > 3$.
- The plate with $n = 11$ layers is examined.

ACKNOWLEDGEMENTS

This research was carried out with the financial support of the Russian Foundation for Basic Research (grants no. 16-01-00580a, 14-01-00271a, 16-51-52025 MHTa).

REFERENCES

[1] G. Kirchhoff, *Vorlesungen uber Matematische Physik. Mechanik*, Leipzig, 1876 [in German].

[2] A.E.H. Love, *A treatise on the mathematical theory of elasticity*, Cambridge Univ. Press, 1927.

[3] S.P. Timoshenko, On the correction for shear of the differential equation for transverse vibrations of prismatic bars, *Philos. Mag.*, **4**, Ser. 6, no. 242, 1921.

[4] E. Reissner, The effect of transverse shear deformation on the bending of elastic plates, *Trans. ASME, J. Appl. Mech.*, **12**, 69–77, 1945.

[5] K.F. Chernykh, V.A. Rodionova, B.F. Titaev, *Applied theory of anisotropic plates and shells*, St.Petersburg Univ. Press, 1996 [in Russian].

[6] I.N. Vekua, On one method of calculating prismatic shells, *Trudy Tbilis. Mat. Inst.*, **21**, 191–259, 1955 [in Russian].

[7] A.L. Goldenweizer, *Theory of Elastic Thin Shells*, Pergamon Press, London, 1961.

[8] P.E. Tovstik, T.P. Tovstik, A thin-plate bending equation of second-order accuracy, *Doklady Physics*, **59**, no. 8, 389–392, 2014.

[9] Y. Vetyukov, A. Kuzin, M. Krommer, Asymptotic splitting in the three-dimensional problem of elasticity for non-homogeneous piezoelectric plates, *Int. J. Solids and Structures*, **48**, 12–23, 2011.

[10] V.A. Eremeev, L.M. Zubov, *Mechanics of elastic shells*, Moscow, Nauka, 2008, [in Russian].

[11] H. Altenbach, G.I. Mikhasev (eds.), *Shell and membrane theories in mechanics and biology*, Springer (2014).

[12] J.N. Reddy, A refined nonlinear theory of plates with transverse shear deformation, *Int. J. Solids and Structures*, **20**, 881–896, 1994.

[13] S.P. Timoshenko, *Strength of materials*, D. Van Nostrand, Toronto, New York, London, 1956.

[14] V. Birman, *Plate Structures*, Springer, Dordrecht, 2011.

[15] N.F. Morozov, P.E. Tovstik, Bulk and surface stability loss of materials. In: *Multi-Scaling of Synthetic and Natural Systems with Self-Adaptive Capacity*, Taiwan, 27–30, 2010.

[16] R. Kienzler, P. Schneider, Comparison of various linear plate theories in the light of a consistent second order approximation, *Shell Structures: Theory and Applications. Proc. 10th SSTA 2013 Conf.*, **3**, 109–112, 2014.

[17] P.E. Tovstik, T.P. Tovstik, On the 2D models of plates and shells including shear, *ZAMM*, **87**, no. 2, 160–171, 2007.

[18] G. Winkler, *Image analysis, random fields and dynamic Monte Carlo methods*, Springer-Verlag, Berlin, 1995.

[19] O.C. Zienkiewicz, R.C. Taylor, *The finite element method, Vol. I, 4th Edition*. McGraw Hill, 1989.

[20] J.T. Oden, T. Belytschko, I. Babuska, T.J.R. Hughes, Research directions in computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, **192**, 913–922, 2003.

[21] J.H. Argyris, M. Papadrakakis, L. Karapitta, Elastoplastic analysis of shells with the triangular element TRIC. M. Papadrakakis, A. Samartin, E. Oñate eds. *4th International Colloquium on Computation of Shell and Spatial Structures (IASS-IACM 2000)*, Chania, Crete, Greece, June 4-7, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EFFICIENT SIZING OF STRUCTURES UNDER STRESS CONSTRAINTS

Zhi Hong, Mostafa M. Abdalla

Faculty of Aerospace Engineering, Delft University of Technology
Kluyverweg 1, 2629 HS Delft, Netherlands
e-mail: {Z.Hong, M.M.Abdalla}@tudelft.nl

Keywords: Sizing Optimization, Stress Constraints, Preconditioned Conjugate Gradient Method, Sensitivity Analysis

Abstract. *Optimisation algorithms used to automatically size structural members commonly involve stress constraints to avoid material failure. Therefore the cost of optimisation grows rapidly as the number of structural members is increased due to the corresponding increase in the number of constraints. In this work, an efficient method for large scale stress constrained structural sizing optimisation problems is proposed. A convex, separable, and scalable approximation for stress constraints which splits the approximation into a local fully stressed term and a global load distribution term is introduced. Predictor-corrector interior point method, an excellent option for large scale optimization problem, is chosen to solve the approximate subproblems. The core idea in this work is to achieve computational efficiency in the optimization procedure by avoiding the construction and the solution of the Schur complement system generated by the interior point method. Avoiding the Schur complement, and explicit sensitivity analysis, eliminates the high cost of solving stress constrained problems within the interior point optimisation. This is achieved using the preconditioned conjugate gradient method, and a new preconditioner is proposed specifically for stress constrained problems. The proposed method is applied to a number of beam sizing problems. Numerical results show that optimal complexity is achieved by the algorithm, the computational cost being linearly proportional to the number of sizing variables.*

1 INTRODUCTION

Optimisation methods have been extensively used to size structural components [1][2][3] [4][5][6] — mostly in a finite element environment. The use of optimisation algorithms allows for automated sizing and for a more thorough search of the design space leading more efficient structures. Strength constraints, which define the failure limits for the structure, have to be satisfied in every part of the structure during the sizing optimization. This is enforced numerically by applying strength constraints per finite element. For a large and complex structure with hundreds of thousands of degree of freedom, the computational cost of strength constraints becomes prohibitive for large scale optimization problems.

One simple optimization algorithm to handle stress constraints is the fully stressed method [7]. However, it turns out that the fully stressed design is not the minimum-weight design for statically indeterminate structures [8] and works well only with discrete structures. From then on, the structure optimization based on optimality condition starts to be investigated. In 1974, by introducing stress approximations, L.A. Schmit and B. Farshi [9] first achieve a significant improvement in computational efficiency of sizing optimization. Different approximation methods have been investigated thereafter [10][11][12][13][14][15]. Higher order approximations, even though they demonstrate high accuracy, are complicated when calculating the coefficients for higher order terms. After numerous attempts, first order and diagonal second order approximation are regarded as good options when both accuracy and computational complexity have been taken into account [10][14][16]. Therefore, they are commonly applied in structural optimization.

Interior point method stands out among gradient-based algorithms because of its robustness and efficiency. The convergence rate of interior point method is not sensitive to the size of optimization problem. Originally proposed for linear programming, it has been extended to different types of problems. Boggs et al.[17] use interior point method to solve sequential quadratic programming (SQP) for large scale nonlinear programming problems. Robert and David[18] extend the method to solve nonconvex nonlinear programming, and show the robustness and efficiency of the method. In [19], Mehrotra has shown that by the predictor-corrector pattern, the number of iterations decreases further to solve linear programming problem. By employing a Taylor polynomial of second order to approximate a primal-dual trajectory, Mehrotra[20] successfully improve the efficiency of the algorithm. Andersen, Roos and Terlaky[21], utilize Mehrotra's predictor-corrector interior point method to solve large-scale sparse conic quadratic optimization problems robustly and efficiently. Zillober[22] demonstrates the advantages of predictor-corrector interior point method over the dual approach in solving large scale nonlinear programming with applications to structural optimisation problems.

The Preconditioned Conjugate Gradient (PCG) method is an efficient iterative technique for the solution of large sets linear equations. This iterative method converges rapidly if the preconditioner manages to reduce the condition number of coefficient matrix or when a good initial guess is available. Schmit [23] employes the PCG to solve finite element method in structural optimization and discuss the potential parallel computation of the method. Another advantage of PCG is that the coefficient matrix does not need to be stored explicitly. This property will help greatly to save computational effort in the proposed algorithm.

The objective of this paper is to demonstrate an efficient optimization method to optimize statically indeterminate beam structures. The method is designed to be fast for typical sizing scenarios under stress constraints. The computational complexity of of the algorithm is reduced from approximately cubic dependence on the number of elements to approximately linear de-

pendence.

The structure in this paper is outlined as follows: In Section 2 the framework of the conventional optimization approach for stress constrained problems is described. A brief description of the Finite Element model, successive convex approximations for stress constraints, and sensitivity analysis is given. The predictor-corrector interior point method used in this paper is also described and the computational complexity of the conventional approach estimated. Section 3 introduces an efficient and cheap preconditioner to solve the dense linear system generated by the interior point method. The matrix-vector multiplications needed by the PCG method are reduced to an implicit sensitivity analysis leading to a reduction in the computational cost of sensitivity analysis . Two numerical examples will be used to demonstrate the efficiency of the optimization method in Section 4. Finally, Section 5 provides the conclusions.

2 Formulation

Before describing the proposed method, an overview of the optimisation process, stress approximation, sensitivity analysis and the iterative optimization is given in this section. The purpose is to familiarise the reader with the relevant concepts and formulations first, and to identify the computational bottlenecks in the procedure.

The, more or less, standard approach in structural optimisation is to use successive convex and separable approximations. The optimisation starts at a given design point, convex separable approximations of all involved response are constructed based on sensitivity information, then the approximate subproblem based on the approximations is solved to find the next design iterate. The procedure is repeated until a suitable convergence tolerance is achieved. The solution of the convex subproblem itself is obtained using an iterative convex optimisation technique. In this work, an predictor-corrector interior point method is used. The various elements of the procedure are explained in more details in the following sections.

2.1 Successive Convex Approximation Method

In this section, a cheap and efficient convex approximation for stress constraints is described. Vanderplaats [12, 13] has shown that, approximating the internal forces first as a *linear* function of the design variables, is more accurate than approximating the stress constraints directly. In this approach, the linear approximations are used to calculate the internal forces from which the stresses and hence the failure constraint(s) (e.g., von Mises stress) may be calculated. However, these so-called force approximations are not necessarily separable or convex. The motivation for the present research is to design a high quality approximation method which overcomes this disadvantage.

The basic idea is to use a zeroth order force approximation rather than a first order approximation. This zeroth order approximation is then corrected to attain first order accuracy by the addition of linear terms.

The model problem is an Euler-Bernoulli beam designed for minimum volume under material failure constraints. For simplicity a rectangular cross section, width w and height h, is considered. Furthermore, either only the width or only the height would be used as a design variable which is denoted by x. The stress constraint takes the form,

$$\frac{|\sigma|}{\sigma_{all}} - 1 \leq 0 \tag{1}$$

where σ_{all} is the allowable normal stress, and the normal stress σ is related to the intermediate

internal force M, representing the bending, moment by,

$$\sigma = \frac{|M|h}{2I}, \tag{2}$$

where I is the moment of inertia of the cross section given by

$$I = \frac{wh^3}{12}. \tag{3}$$

The continuous problem would be solved using methods of the calculus of variations. The numerical treatment starts by discretising the beam using n finite elements. Each element is assigned a design variable. The stress constraint for the ith element may be written as a function of M_i and the design variables,

$$r_i(M_i, x_i) = \pm\frac{6M_i}{\sigma_{all} w_i h_i^2} \leq 1., x_i = w_i \text{ or } h_i, \quad i = 1, 2, \ldots n \tag{4}$$

Note that the bending moment in any given element is generally a function of all the design variables $M_i = M_i(\mathbf{x})$. The volume of the beam is the integral over the length of the cross sectional area wh.

Let the current design iterate be denoted by $\mathbf{x}_k$. The proposed approximation then takes the form,

$$r_i \approx \frac{b_i^{(k)}}{x_i^q} + \sum_{j=1}^{n} A_{ji}^{(k)}(x_j - x_j^{(k)}). \tag{5}$$

For width optimisation $q = 1$ and $b_i^{(k)} = 6|M_i^{(k)}|/\sigma_{all}h^2$, while for height optimisation $q = 2$ and $b_i^{(k)} = 6|M_i^{(k)}|/\sigma_{all}w$. The first term represents the zeroth order force approximation for the stress constraint. This first term would be exact only for statically determinate structures where the bending moment is independent of the design variables. The second term provides a linear correction and represents the effect of *load redistribution* present in statically indeterminate structures. The coefficients $A_{ji}^{(k)}$ are obtained by matching the derivatives of the approximation Eq.(5) with respect to design variables $\mathsf{G}(\mathbf{x}_k)$ to the derivative values obtained from sensitivity analysis G_e at the approximation point $\mathbf{x}_k$.

$$\mathsf{G}(\mathbf{x}_k) = \mathsf{G}_s(\mathbf{x}_k) + A^{(k)} \tag{6}$$

$$\mathsf{G}_e(\mathbf{x}_k) = \mathsf{G}(\mathbf{x}_k) \tag{7}$$

Where,

$$\mathsf{G}_e(\mathbf{x}_k) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1} & \frac{\partial r_2}{\partial x_1} & \cdots & \frac{\partial r_n}{\partial x_1} \\ \frac{\partial r_1}{\partial x_2} & \frac{\partial r_2}{\partial x_2} & \cdots & \frac{\partial r_n}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial r_1}{\partial x_n} & \frac{\partial r_2}{\partial x_n} & \cdots & \frac{\partial r_n}{\partial x_n} \end{bmatrix}\Bigg|_{\mathbf{x}_k} \tag{8}$$

$$\mathsf{G}_s(\mathbf{x}_k) = \begin{bmatrix} -\frac{qb_1^{(k)}}{x_1^{q+1}} & 0 & \cdots & 0 \\ 0 & -\frac{qb_2^{(k)}}{x_2^{q+1}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{qb_n^{(k)}}{x_n^{q+1}} \end{bmatrix} \tag{9}$$

Therefore the coefficient $A^{(k)}$ is,

$$A^{(k)} = \mathsf{G}_e(\mathbf{x}_k) - \mathsf{G}_s(\mathbf{x}_k) \tag{10}$$

The proposed stress approximation here is thus convex and separable.

For the case under consideration, the volume of the beam is an exact linear function of the design variables. Thus, the objective function may be written as,

$$V = \sum_{i=1}^{n} c_i x_i. \tag{11}$$

2.2 Sensitivity Analysis

Sensitivity analysis is required first to provide the gradient of the objective and constraints in gradient-based optimisation. Then the gradient is employed to guide the optimisation method in the optimisation procedure, which pushes the design variables towards the optimal solution in the feasible domain. The well-established and efficient way to calculate sensitivity is through the Finite Element Method (FEM) which estimates the mechanical response of structures.

Consequently, FEM is employed to analyse structures and provide information required by the proposed optimization method. Due to the ease-to-implement and powerful capability of FEM to analyse complex structures, it ensures the propose to solve real engineering problems in the very beginning. According to minimum total potential energy principal, we can come to the equation for static analysis:

$$\mathsf{K}\mathbf{u} = \mathbf{f} \tag{12}$$

where K is the global stiffness matrix attained by assembling stiffness matrix of each element according to their degree of freedom. $\mathbf{u}$ is the global displacement of each node. $\mathbf{f}$ is the external force loaded on the structure.

Before we get the internal force for i^{th} element in finite element method, the displacement corresponding to the DOF of ith element in local coordinate system should be obtained first by,

$$\tilde{\mathbf{u}}_i = \mathsf{T}_i \mathbf{u} \tag{13}$$

Here, T_i is the transformation matrix which transfer the displacement of the ith element from global coordinate to local element coordinate system.

The equation for the internal force, i.e., the bending moment in this case, is,

$$M_i = \mathsf{D}(x_i)\mathsf{B}_i\tilde{\mathbf{u}}_i. \tag{14}$$

Here, $\mathsf{D}(x_i)$ is the material stiffness, B_i is the strain-displacement matrix, and $\mathbf{u}_i$ contains the nodal degrees of freedom connected to the ith element. In the case of beams under pure bending D_i for the ith element is scalar given by the bending stiffness $EI(x_i)$, and the strain-displacement matrix B_i reduced a row vector relating the average curvature to the nodal degrees of freedom.

Substituting Eq. (13) into Eq. (14), the sensitivity of the beding moment M_i with respect to the design variable x_j is given by,

$$\frac{\partial M_i}{\partial x_j} = \frac{\partial D_i}{\partial x_i}\delta_{ij}\mathsf{B}_i\mathsf{T}_i\mathbf{u}_i + \mathbf{z}_i^T\frac{\partial \mathbf{u}_i}{\partial x_j} \tag{15}$$

where δ_{ij} is Kronecker's delta, and $\mathbf{z}_i = \mathsf{T}_i^T\mathsf{B}_i^T\mathsf{D}_i^T$.

There are two common methods to proceed with the sensitivity analysis: the direct method and the adjoint method [24].

The direct method starts from the assembled finite element equations,

$$\mathsf{K}(\mathbf{x})\,\mathbf{u} = \mathbf{f}, \tag{16}$$

where K is the assembled stiffness matrix, $\mathbf{u}$ the vector of degrees of freedom and $\mathbf{f}$ is the vector of external forces.

Differentiating Eq.(16),

$$\mathsf{K}\frac{\partial \mathbf{u}}{\partial x_j} = \frac{\partial \mathbf{f}}{\partial x_j} - \frac{\partial \mathsf{K}}{\partial x_j}\,\mathbf{u}, \tag{17}$$

The solution to Eq.(17) allows the computation of the derivative of the displacements with respect to every design variable.

The adjoint method attempts to avoid the explicit calculation of the displacement derivatives. To this end, the vector $\mathbf{z}_i$ is pseudo-load, and the following adjoint system is introduced,

$$\mathsf{K}\,\mathbf{t}_i = \mathbf{z}_i. \tag{18}$$

Then the sensitivity of the bending moments is given by,

$$\frac{\partial M_i}{\partial x_j} = \frac{\partial D_i}{\partial x_i}\delta_{ij}\mathsf{B}_i\mathsf{T}_i\mathbf{u}_i - \mathbf{t}_i^T\frac{\partial \mathsf{K}}{\partial x_j}\mathbf{u} \tag{19}$$

The derivatives of the bending moment in all elements then follows from Eq.(15). The derivatives of the stress constraint itself are then easily computed as follows,

$$\frac{\partial r_i}{\partial x_j} = \left.\frac{\partial r_i}{\partial x_i}\right|_{M_i}\delta_{ij} + \left.\frac{\partial r_i}{\partial M_i}\right|_{\mathbf{x}_i}\frac{\partial M_i}{\partial x_j} \tag{20}$$

Using the direct method, Eq.(17) has to be solved once for each design variable. By contrast, Eq. (18) for the adjoint method needs to be solved only as many times as there are stress constraints.

2.3 Predictor-Corrector Interior Point Method

The predictor-corrector interior point method is a good option for large scale optimization problems because of its great efficiency [25]. In this paper, the interior point method is therefore employed in structural optimisation. In interior point methods a logarithmic penalty is used to enforce the constraints. The penalty parameter is then driven towards zero to obtain the optimal solution. At each step of optimisation Newton's method is used to update primal and dual (Lagrange multiplier) variables. The predictor corrector method of Mehrotra[20] is used to accelerate convergence.

When minimizing the volume of a beam with stress and design variables constrained, the problem can be formulated as follows.

objective:

$$\text{Min:} f_0(x) = \sum_{i=1}^{n} x_i h_i l_i \tag{21}$$

subject to:

$$f_i(x) \leq \sigma_{all} \quad i = 1 \ldots \text{m, m is the number of stress constraints}$$
$$x_j^l \leq x_j \leq x_j^u \quad j = 1 \ldots \text{n, n is the number of design variables}$$

In this paper, $m = n$ for the beam structure. x_i and l_i are the thickness and the length of the ith beam, respectively. x_j^l and x_j^u are the lower bound and upper bounds of the design variable. Since we confine the optimization problem to be sizing optimization, $x_j^l > 0$.

The Lagrangian function of the problem is

$$\begin{aligned} L(\mathbf{x}, \mathbf{s}, \mathbf{y}^s, \mathbf{s}^u, \mathbf{y}^u, \mathbf{s}^l, \mathbf{y}^l) =& f_0 + (\mathbf{y}^s)^T(\mathbf{f}(x) + \mathbf{s}) + (\mathbf{y}^u)^T(\mathbf{x} - \overline{\mathbf{x}} + \mathbf{s}^u) + (\mathbf{y}^l)^T(\underline{\mathbf{x}} - \mathbf{x} + \mathbf{s}^l) \\ & - \mu \sum_{i=1}^{n} \ln s_i - \mu \sum_{i=1}^{n} \ln s_i^u - \mu \sum_{i=1}^{n} \ln s_i^l \end{aligned} \tag{22}$$

f_0: normalized objective function; $\mathbf{f}(\mathbf{x})$: normalized stress constraints (stress: $f_i(\mathbf{x}) = sgn(\sigma_i)\sigma_i/\sigma_{allow} - 1, i = 1 \ldots n$); $\mathbf{s}$: the slack of the stress constraints which transform inequality constraints stress constraints into equality constraints; $\mathbf{s}^u$,$\mathbf{s}^l$: the slack of the design variables with respect to the upper bound and lower bound which transform inequality constraints of design variables into equality constraints; $\mathbf{y}^s$: Lagrange multipliers of the stress constraints; $\mathbf{y}^u$,$\mathbf{y}^l$: the Lagrange multipliers of the design variable with respect to the upper bound and the lower bound respectively; $\overline{\mathbf{x}}$: upper bound of the design variables ($\overline{x}_i = x_i^u$); $\underline{\mathbf{x}}$: lower bound of the design variables ($\underline{x}_i = x_i^l$); μ: penalty parameter.

According to Karush-Kuhn-Tucker conditions, a set of equations will be:

$$\frac{\partial L}{\partial \mathbf{x}} = \nabla \mathbf{f}(\mathbf{x}) + \mathsf{G}(\mathbf{x}) \cdot \mathbf{y}^s + \mathbf{y}^u - \mathbf{y}^l = \mathbf{0} \tag{23}$$
$$\frac{\partial L}{\partial \mathbf{y}^l} = \underline{\mathbf{x}} - \mathbf{x} + \mathbf{s}^l = \mathbf{0} \tag{24}$$
$$\frac{\partial L}{\partial \mathbf{y}^u} = \mathbf{x} - \overline{\mathbf{x}} + \mathbf{s}^u = \mathbf{0} \tag{25}$$
$$\frac{\partial L}{\partial \mathbf{y}^s} = \mathbf{f}(\mathbf{x}) + \mathbf{s} = \mathbf{0} \tag{26}$$
$$\frac{\partial L}{\partial \mathbf{s}} = \mathsf{Y}^s \cdot \mathbf{s} - \mu \cdot \overline{\mathbf{e}} = \mathbf{0} \tag{27}$$
$$\frac{\partial L}{\partial \mathbf{s}^u} = \mathsf{Y}^u \cdot \mathbf{s}^u - \mu \cdot \overline{\mathbf{e}} = \mathbf{0} \tag{28}$$
$$\frac{\partial L}{\partial \mathbf{s}^l} = \mathsf{Y}^l \cdot \mathbf{s}^l - \mu \cdot \overline{\mathbf{e}} = \mathbf{0} \tag{29}$$

Here $\mathsf{G}(\mathbf{x})$ is the gradient of the stress approximation (Eq.(5)); $\overline{\mathbf{e}}^T = \overbrace{[1, 1, \ldots, 1]}^{n}$; $\mathsf{Y}^s = diag(y_1^s, \cdots, y_n^s)$; Y^u and Y^l are similar.

In order to solve the above equations, Newton's method is employed. Then a linear system

of the subproblem is obtained:

$$\mathsf{G}(\mathbf{x}) \cdot \Delta\mathbf{y}^s + \mathsf{H}(\mathbf{x}) \cdot \mathbf{y}^s \cdot \Delta\mathbf{x} + \Delta\mathbf{y}^u - \Delta\mathbf{y}^l = -\nabla\mathbf{f} - \mathsf{G}(\mathrm{x}) \cdot \mathbf{y}^s - \mathbf{y}^u + \mathbf{y}^l \tag{30}$$

$$-\Delta\mathbf{x} + \Delta\mathbf{s}^l = \mathbf{x} - \underline{\mathbf{x}} - \mathbf{s}^l \tag{31}$$

$$\Delta\mathbf{x} + \Delta\mathbf{s}^u = \overline{\mathbf{x}} - \mathbf{x} - \mathbf{s}^u \tag{32}$$

$$\mathsf{G}^T(\mathbf{x})\Delta\mathbf{x} + \Delta\mathbf{s} = -\mathbf{f}(\mathbf{x}) - \mathbf{s} \tag{33}$$

$$\mathsf{S} \cdot \Delta\mathbf{y}^s + \mathsf{Y}^s \cdot \Delta\mathbf{s} = \mu \cdot \overline{\mathbf{e}} - \mathsf{Y}^s \cdot \mathbf{s} \tag{34}$$

$$\mathsf{S}^u \cdot \Delta\mathbf{y}^u + \mathsf{Y}^u \cdot \Delta\mathbf{s}^u = \mu \cdot \overline{\mathbf{e}} - \mathsf{Y}^u \cdot \mathbf{s}^u \tag{35}$$

$$\mathsf{S}^l \cdot \Delta\mathbf{y}^l + \mathsf{Y}^l \cdot \Delta\mathbf{s}^l = \mu \cdot \overline{\mathbf{e}} - \mathsf{Y}^l \cdot \mathbf{s}^l \tag{36}$$

Here $\mathsf{H}(\mathbf{x})$ are the Hessian matrixs of the stress approximation with respect to the design variables. Since the approximation is separable, $\mathsf{H}(\mathbf{x}) \cdot \mathbf{y}_s$ is diagonal matrix which is very cheap to obtain as shown in [22]. To signify the right hand side of the equations above, we have:

$$\begin{bmatrix} f_x \\ f_{y^l} \\ f_{y^u} \\ f_{y^s} \\ f_s \\ f_{s^u} \\ f_{s^l} \end{bmatrix} = \begin{bmatrix} -\nabla\mathbf{f} - \mathsf{G}(\mathbf{x}) \cdot \mathbf{y}^s - \mathbf{y}^u + \mathbf{y}^l \\ \mathbf{x} - \underline{\mathbf{x}} - \mathbf{s}^l \\ \overline{\mathbf{x}} - \mathbf{x} - \mathbf{s}^u \\ -\mathbf{f}(\mathbf{x}) - \mathbf{s} \\ \mu \cdot \overline{\mathbf{e}} - \mathsf{Y}^s \cdot s \\ \mu \cdot \overline{\mathbf{e}} - \mathsf{Y}^u \cdot s^u \\ \mu \cdot \overline{\mathbf{e}} - \mathsf{Y}^l \cdot s^l \end{bmatrix} \tag{37}$$

From Eq.(32) and Eq.(35), we can get:

$$\Delta\mathbf{s}^u = f_{y^s} - \mathsf{G}^T(x)\Delta\mathbf{x} \tag{38}$$

$$\Delta\mathbf{y}^u = (\mathsf{S}^u)^{-1} \cdot (\mathbf{f}_{s^u} - \mathsf{Y}^u \cdot \Delta\mathbf{s}^u) \tag{39}$$

From Eq.(31) and Eq.(36), we can get:

$$\Delta\mathbf{s}^l = f_{y^l} + \Delta\mathbf{x} \tag{40}$$

$$\Delta y^l = (\mathsf{S}^l)^{-1} \cdot (\mathbf{f}_{s^l} - \mathsf{Y}^l \cdot (\mathbf{f}_{y^l} + \Delta\mathbf{x})) \tag{41}$$

From Eq.(33), we can get:

$$\Delta\mathbf{s} = f_{y^s} - \mathsf{G}^T(x) \cdot \Delta\mathbf{x} \tag{42}$$

Substitute $\Delta\mathbf{s}^u$ into $\Delta\mathbf{y}^u$ and $\Delta\mathbf{s}^l$ into $\Delta\mathbf{y}^l$, then substitute $\Delta\mathbf{y}^u$ and $\Delta\mathbf{y}^l$ into Eq.(30) yields:

$$\mathsf{G}(x) \cdot \Delta\mathbf{y}^s + \mathsf{H}(x) \cdot \mathbf{y}^s \cdot \Delta\mathbf{x} + \mathsf{Y}^u \cdot (\mathsf{S}^u)^{-1} \cdot \Delta\mathbf{x} + \mathsf{Y}^l \cdot (\mathsf{S}^l)^{-1} \cdot \Delta\mathbf{x} = \mathbf{f}_{sys_x} \tag{43}$$

$$\mathbf{f}_{sys_x} = \mathbf{f}_x - (\mathsf{S}^u)^{-1} \cdot \mathbf{f}_{s^u} + Y^u \cdot (\mathsf{S}^u)^{-1} \cdot \mathbf{f}_{y^u} + (\mathsf{S}^l)^{-1} \cdot \mathbf{f}_{s^l} - Y^l \cdot (\mathsf{S}^l)^{-1} \cdot \mathbf{f}_{y^l}$$

with Eq.(34) substituting into Eq.(33) it comes to:

$$\mathsf{G}^T(x) \cdot \Delta\mathbf{x} + (\mathsf{Y}^s)^{-1} \cdot (\mathbf{f}_s - \mathsf{S} \cdot \Delta\mathbf{y}_s) = \mathbf{f}_{y^s} \tag{44}$$

Combining Eq.(43) and Eq.(44) we can get the final equation of the subproblem with only $\Delta\mathbf{y}^s$:

$$[(\mathsf{Y}^s)^{-1} \cdot \mathsf{S} + \mathsf{G}^T(x) \cdot \mathsf{C}^{-1} \cdot \mathsf{G}(x)]\Delta\mathbf{y}^s = \mathbf{f} \tag{45}$$

Here, $\mathsf{C} = \mathsf{H}(x) \cdot \mathbf{y}^s + Y^u \cdot (\mathsf{S}^u)^{-1} + \mathsf{Y}^l \cdot (\mathsf{S}^l)^{-1}$

Since all these three parts are diagonal matrix, C is a diagonal matrix.

$$\mathbf{f} = -\mathbf{f}_{y^s} + (\mathsf{Y}^s)^{-1} \cdot \mathbf{f}_s + \mathsf{G}^T(x) \cdot \mathsf{C}^{-1} \cdot \mathbf{f}_{sys_x} \tag{46}$$

From Eq.(45) we can get $\Delta\mathbf{y}^s$. Back substitute it into Eq.(30 – 36) we can get: $\Delta\mathbf{x}$, $\Delta\mathbf{s^u}$, $\Delta\mathbf{s}$, $\Delta\mathbf{s}^l$, $\Delta\mathbf{y}^u$, and $\Delta\mathbf{y}^l$.

Algorithm 1 Algorithm of Predictor-Corrector Interior Point Method

1: Set $k = 1$, initialize the parameters $\mu, \mathbf{x}, \mathbf{s}^u, \mathbf{s}^l, \mathbf{s}, \mathbf{y}^s, y^u, y^l$.
2: **while** $d_{gap} \geq tol$ **do** ▷ stop if duality gap $< tol$
3: Update Eq.(45) by setting $\mu = 0$
4: $\mu_{(k+1)} \leftarrow$ solve Eq.(45) in predictor step
5: Update Eq.(45) by setting $\mu = \mu_{(k+1)}$
6: $\mathbf{x}, \mathbf{s}, \mathbf{y}^s, \mathbf{s}^u, \mathbf{y}^u, \mathbf{s}^l, \mathbf{y}^l \leftarrow$ solve Eq.(45) in corrector step
7: Update $d_{(k+1)}$
8: **end while**

- Initialize the variables (Line 1) in the predictor-corrector interior point method:

  Set the design variables $\mathbf{x}$ a set of feasible design. Make $\mathbf{s}^u = \overline{\mathbf{x}}$, $\mathbf{s}^l = \underline{\mathbf{x}}$, $\mathbf{s} = \mathbf{1}$, $\mu = 1/(dim(\mathbf{y}^s) + 2dim(\mathbf{x}))$, $\mathbf{y}^s = \mu/\mathbf{s}$, $\mathbf{y}^u = \mu/\mathbf{s}^u$, $\mathbf{y}^l = \mu/\mathbf{s}^l$.

  $dim(\mathbf{x})$ is the length of vector $\mathbf{x}$; $dim(\mathbf{y}^s)$ is the length of vector $\mathbf{y}^s$

- The way to calculate penalty parameter μ (Line 4) in predictor step is as follows:

  First we solve Eq.(45) with $\mu = 0$, where we get slacks ($\Delta\mathbf{s}$, $\Delta\mathbf{s}^u$, $\Delta\mathbf{s}^l$) and Lagrange multipliers ($\Delta\mathbf{y}^s$, $\Delta\mathbf{y}^u$, $\Delta\mathbf{y}^l$). Then we update the duality gap by,

  $d\_gap_{(k+1)} = (\mathbf{s} + \delta_{primal} \cdot \Delta\mathbf{s})^T \cdot (\mathbf{y}^s + \delta_{dual} \cdot \Delta\mathbf{y}^s) + (\mathbf{s}^u + \delta_{primal} \cdot \Delta\mathbf{s}^u)^T \cdot (\mathbf{y}^u + \delta_{dual} \cdot \Delta\mathbf{y}^u) + (\mathbf{s}^l + \delta_{primal} \cdot \Delta\mathbf{s}^l)^T \cdot (\mathbf{y}^l + \delta_{dual} \cdot \Delta\mathbf{y}^l)$

  Meanwhile we calculate the current duality gap by,

  $d\_gap_k = \mathbf{s}^T \cdot \mathbf{y}^s + (\mathbf{s}^u)^T \cdot \mathbf{y}^u + (\mathbf{s}^l)^T \cdot \mathbf{y}^l$

  The penalty parameter is obtained by,

  $\mu_{(k+1)} = min\{max((d\_gap_{(k+1)}/d\_gap_k)^2, 0.1), 1\} \cdot \mu_k$

- Update the design variables, Lagrange multipliers and slacks in Line 6:

  $\mathbf{x} = \mathbf{x} + \delta_{primal} \cdot \Delta\mathbf{x}$; $\mathbf{s} = \mathbf{s} + \delta_{primal} \cdot \Delta\mathbf{s}$;

  $\mathbf{s}^u = \mathbf{s}^u + \delta_{primal} \cdot \Delta\mathbf{s}^u$; $\mathbf{s}^l = \mathbf{s}^l + \delta_{primal} \cdot \Delta\mathbf{s}^l$;

  $\mathbf{y}^s = \mathbf{y}^s + \delta_{dual} \cdot \Delta\mathbf{y}^s$; $\mathbf{y}^u = \mathbf{y}^u + \delta_{dual} \cdot \Delta\mathbf{y}^u$;

  $\mathbf{y}^l = \mathbf{y}^l + \delta_{dual} \cdot \Delta\mathbf{y}^l$;

The way to obtain δ_{primal}, δ_{dual} in the previous steps are all based on the following rule:
$\delta_{primal} = 0.95 \cdot min\{max\{-s_i^l/\delta s_i^l\}, max\{-s_i/\delta s_i\}, max\{-s_i^u/\delta s_i^u\}, 0.95\}$
$\delta_{dual} = 0.95 \cdot min\{max\{-y_i^l/\delta y_i^l\}, max\{-y_i^u/\delta y_i^u\}, max\{-y_i^s/\delta y_i^s\}, 0.95\}$

2.4 Computational Complexity

The computational complexity of a numerical method is directly relevant to the efficiency of the method. In sizing optimization, the computationally expensive operations are FEM analysis, stress sensitivity analysis and iterative optimization with the interior point method. For a numerical structure, the number of elements and/or nodes is denoted by n. The order of magnitude of each of these operations will be estimated as a function of n.

Since the stiffness matrix for FEM analysis is a symmetric and positive definite equation, the Cholesky decomposition is employed to save computational effort. The computational work for Cholesky decomposition of the finite element equations is:

$$O(nm^2)$$

where m is the bandwidth of the equations. For the beam problem under consideration the number of neighbours for each node is fixed and the bandwidth is constant, thus the computational cost for FEM is,

$$C_1 = O(n),$$

the computational cost for FEM analysis is linear with respect to the number of elements in the numerical model.

For stress sensitivity analysis, the most expensive part is to solve the Eq.(18) for the adjoint displacements. This stiffness matrix is the same as that for equilibrium equation in FEM analysis. The order of computational effort for each adjoint calculation is therefore $O(n)$. Nevertheless, in a beam structure with n elements, the number of stress constraints is n. Therefore Eq.(18) is solved n times to get G_e. The computational work for stress sensitivity analysis is:

$$C_2 = O(n^2)$$

The number of iteration inside the interior point method does not vary much when the size of optimization problem increases. The largest cost in each interior point iteration is the construction and solution of the system of equations Eq.(45) repeated here,

$$[(\mathsf{Y}^s)^{-1}\mathsf{S} + \mathsf{G}^T(x)\mathsf{C}^{-1}\mathsf{G}(x)]\Delta\mathbf{y}^s = \mathbf{f}$$

Since Y_s and S are diagonal, the cost to obtain $\mathsf{Y}_s^{-1}\mathsf{S}$ is $O(n)$. The cost to obtain $\mathsf{G}^T(\mathbf{x})\mathsf{C}^{-1}\mathsf{G}(\mathbf{x})$ is $O(n^3)$ because of the cost of full matrix multiplication. The solution of Eq.(45) can be obtained using Cholesky decomposition. The matrix is full, thus the computational cost is $O(n^3)$. As a result, the computational complexity of the interior point method is,

$$C_3 = O(n^3)$$

Since the number of iterations for the method of successive approximations (successive optimization) is almost independent of the problem size. The computational complexity for the entire optimisation is dominated by the interior point method and would be,

$$C = O(n^3)$$

Consequently reducing the computational complexity of the optimisation should start by addressing the cost of Eq.(45). If this is reduced, the second target would be reducing the cost of the adjoint sensitivity analysis. The main contribution of the paper is in reducing both of them to $O(n)$ operations.

3 Numerical Improvements

Since the computational work for the above optimisation method increases sharply while the size of optimisation increases. Methods to reduce the computational complexity of the optimisation procedure will be introduced in this section.

The first priority is to accelerating the iterative optimisation which dominates the overall computational cost. In the proposed method, to alleviate the main cost for the iterative optimisation, the preconditioned conjugate gradient method (PCG) is used to solve system of equations Eq.(45) for the dual variables. The PCG is an efficient solver for large scale linear equations. The storage of the full coefficient matrix is not require as well. Therefore the cost for the matrix matrix multiplication in building the coefficient matrix of Eq.(45) is avoided. Additionally, since the efficiency of the method is crucially dependent on the quality of the preconditioner, a diagonal preconditioner based on the concept of fully stressed design is proposed for the optimisation problem in Section 3.1. This is an optimal preconditioner from the point of view of the computational cost of using the preconditioner for one step ($O(n)$). Numerical results (Section 4) will confirm that it is also optimal from the point of view of the number of the PCG iterations.

The PCG method requires the ability to form the matrix-vector product of the coefficient matrix with an arbitrary vector. To avoid having to explicitly construct the gradient matrix G, a method is proposed to merge the stress sensitivity analysis into the optimisation procedure, which leads to an implicit sensitivity analysis as shown in Section 3.2. Thus, the computational cost of sensitivity analysis is significantly reduced by creating an direct inline link between the FEM and the optimiser. With the implicit sensitivity analysis, the most expensive cost, the production of the gradient matrix and an arbitrary vector, in the PCG can be accordingly reduced to $O(n)$. Combined with PCG and the interior point method, the entire optimisation cost are finally reduced to $O(n)$ from $O(n^3)$. The proposed method therefore achieves powerful computational savings for large scale optimisation problems.

3.1 Preconditioner for Conjugate Gradient Method

The efficiency of the PCG depends greatly on the effectiveness of the preconditioner. A preconditioner M is a matrix that approximates the coefficient matrix of the equations to be solved. It is important to make sure that the preconditioner is cheap to invert and can effectively reduce the condition number of the preconditioned coefficient matrix. In this section, an effective preconditioner for the subproblem Eq.(45) satisfying the two requirements is developed.

Let us recap the system of equations in Eq.(45),

$$\left[(\mathsf{Y}^s)^{-1}\mathsf{S} + \mathsf{G}^T\mathsf{C}^{-1}\mathsf{G}\right]\Delta\mathbf{y} = \mathbf{f}. \tag{47}$$

The gradient matrix G is, in general, a fully populated dense matrix. Thus the coefficient matrix is also dense. From Eq.(5), the gradient matrix G takes the form,

$$\mathsf{G} = \mathsf{G}_s(\mathbf{x}) + A^{(k)}, \tag{48}$$

Where G_s is a diagonal matrix containing the gradient of the statically determinate part of the stress approximation. $A^{(k)}$ is the gradient matrix of the linear part representing load redistribution. As we have derived in Section 2.1, $A^{(k)}$ is a full matrix for statically indeterminate structures. It is expensive to obtain because of the calculation procedure to obtain $\mathsf{G}_e(\mathbf{x}_k)$.

However to create a cheap preconditioner the following approximation is invoked,

$$\mathsf{G} \approx \mathsf{G}_s. \tag{49}$$

Physically this is to assume that the effect of load redistribution is negligible. In other words, that the internal loads in the structure are *frozen* at the value at the latest design point. With these assumption the preconditioner takes the form,

$$\mathsf{M} = (\mathsf{Y}^s)^{-1}\mathsf{S} + \mathsf{G}_s^T\mathsf{C}^{-1}\mathsf{G}_s, \tag{50}$$

where M is a diagonal matrix which is easy to invert. The effectiveness of the preconditioner in terms of the condition number cannot be demonstrated theoretically but will be assessed in the numerical experiments in Section 4.

3.2 Implicit Sensitivity Analysis

While solving the subproblem (Eq.(45)) with the PCG iteratively in the interior point method, it requires the forming of products of the coefficient matrix of Eq.(45) by an arbitrary vector $\mathbf{v}$,

$$\left[(\mathsf{Y}^s)^{-1}\mathsf{S} + \mathsf{G}^T(\mathbf{x})\mathsf{C}^{-1}\mathsf{G}(\mathbf{x})\right]\mathbf{v}. \tag{51}$$

The first term of the coefficient matrix, $(\mathsf{Y}^s)^{-1}\mathsf{S}$ is diagonal and forming products of the form $\mathbf{w}_f = (\mathsf{Y}^s)^{-1}\mathsf{S}\cdot\mathbf{v}$ is cheap ($O(n)$).

However, the second term, for $\mathsf{G}\cdot\mathbf{v}$ and $\mathsf{G}^T\cdot\mathbf{v}$, is extremely expensive for the large scale optimisation problems because of the computational work to obtain the gradient matrix G. As we have shown the cost to form stress sensitivity G_e, which forms part of G, increases quadratically when n increases in Section 2.4. Therefore we attempts to avoid the cost of calculating the gradient matrix G explicitly before the optimisation starts. Instead we merge the gradient matrix inside the interior point method by substituting the derivative of stress approximation Eq.(5) into $\mathsf{G}\cdot\mathbf{v}$, $\mathsf{G}^T\cdot\mathbf{v}$ directly. This forms an implicit sensitivity analysis which will be described below.

The product $\mathbf{w}_s = \mathsf{G}^T\mathsf{C}^{-1}\mathsf{G}\cdot\mathbf{v}$ can be computed in three steps.

$$\begin{aligned}\mathbf{w}_1 &= \mathsf{G}(\mathbf{x})\mathbf{v},\\ \mathbf{w}_2 &= \mathsf{C}^{-1}\mathbf{w}_1,\\ \mathbf{w}_s &= \mathsf{G}^T(\mathbf{x})\mathbf{w}_2.\end{aligned} \tag{52}$$

The matrix C is diagonal, hence the matrix vector product involving it is $O(n)$. Thus the cost is dominated by products of the form $\mathsf{G}\mathbf{v}$ and $\mathsf{G}^T\mathbf{v}$. Note that the matrix $\mathsf{G}(\mathbf{x})$ contains the gradients of the stress approximations at the current iterate of the interior point method which is different from the gradient matrix evaluated at the approximation point $\mathsf{G}(\mathbf{x}_k)$ ($\mathsf{G}_e(\mathbf{x}_k)$). Recalling the decomposition of $\mathsf{G}(x)$, Eq.(6) and Eq.(10), G is expressed as,

$$\mathsf{G}(\mathbf{x}) = \mathsf{G}_s(\mathbf{x}) - \mathsf{G}_s(\mathbf{x}_k) + \mathsf{G}(\mathbf{x}_k). \tag{53}$$

Noting that $\mathsf{G}_s(\mathbf{x})$ and $\mathsf{G}_s(\mathbf{x}_k)$ are diagonal and cheap to explicitly construct and store, then the cost is dominated by products of the form $\mathsf{G}(\mathbf{x}_k)\cdot\mathbf{v}$ and $\mathsf{G}^T(\mathbf{x}_k)\cdot\mathbf{v}$. The purpose of this section is to show that these products may be formed without the explicit calculation and storage of the matrix $\mathsf{G}(\mathbf{x}_k)$ and in $O(n)$ operations.

To compute $\mathsf{G}(\mathbf{x}_k)\cdot\mathbf{v}$ start from Eq.(20) for the stress sensitivity to obtain,

$$\mathsf{G}(\mathbf{x}_k)\cdot\mathbf{v} = \begin{bmatrix}\sum_i (\mathsf{D}_i\mathsf{B}_i\mathsf{T}_i)\, v_i(-\mathsf{K}^{-1})\frac{\partial \mathsf{K}}{\partial x_1}\mathbf{u}\\ \sum_i (\mathsf{D}_i\mathsf{B}_i\mathsf{T}_i)\, v_i(-\mathsf{K}^{-1})\frac{\partial \mathsf{K}}{\partial x_2}\mathbf{u}\\ \vdots\\ \sum_i (\mathsf{D}_i\mathsf{B}_i\mathsf{T}_i)\, v_i(-\mathsf{K}^{-1})\frac{\partial \mathsf{K}}{\partial x_n}\mathbf{u}\end{bmatrix}\Bigg|_{\mathbf{x}_k} \tag{54}$$

The first part $\sum_i (\mathsf{D}_i\mathsf{B}_i\mathsf{T}_i)\, v_i(-\mathsf{K}^{-1})$ is the same in each term of the above equation. Since the stiffness matrix K is symmetric, this term can be obtained by the adjoint method.

$$\mathsf{K}\mathbf{t} = \mathbf{z}, \tag{55}$$

where, $\mathbf{z} = \sum_i (\mathsf{D}_i\mathsf{B}_i\mathsf{T}_i)^T v_i$ is the dummy load.

The final equation becomes,

$$G(\mathbf{x}_k)\cdot\mathbf{v} = \begin{bmatrix} \mathbf{t}^T\frac{\partial \mathsf{K}}{\partial x_1}\mathbf{u} \\ \mathbf{t}^T\frac{\partial \mathsf{K}}{\partial x_2}\mathbf{u} \\ \vdots \\ \mathbf{t}^T\frac{\partial \mathsf{K}}{\partial x_n}\mathbf{u} \end{bmatrix}\Bigg|_{\mathbf{x}_k} \tag{56}$$

The number nonzero elements in $\frac{\partial \mathsf{K}}{\partial x_i}$ depends on the number of degrees of freedom of the element which is a small *fixed* number. Hence, $\frac{\partial \mathsf{K}}{\partial x_i}\mathbf{u}$ has a fixed cost to compute. Therefore the most expensive cost for the above equation is the solution of the adjoint system. The computational cost correspondingly is $O(n)$ to evaluate $\mathsf{G}(\mathbf{x}_k)\cdot\mathbf{v}$ once.

Next calculating $\mathsf{G}^T(\mathbf{x}_k)\cdot\mathbf{v}$ is considered.

$$\mathsf{G}^T(\mathbf{x}_k)\cdot\mathbf{v} = \begin{bmatrix} (\mathsf{D}_1\mathsf{B}_1\mathsf{T}_1)\,(-\mathsf{K}^{-1})\sum_i v_i\frac{\partial \mathsf{K}}{\partial x_i}\mathbf{u} \\ (\mathsf{D}_2\mathsf{B}_2\mathsf{T}_2)\,(-\mathsf{K}^{-1})\sum_i v_i\frac{\partial \mathsf{K}}{\partial x_i}\mathbf{u} \\ \vdots \\ (\mathsf{D}_n\mathsf{B}_n\mathsf{T}_n)\,(-\mathsf{K}^{-1})\sum_i v_i\frac{\partial \mathsf{K}}{\partial x_i}\mathbf{u} \end{bmatrix}\Bigg|_{\mathbf{x}_k} \tag{57}$$

It is important to notice $(-\mathsf{K}^{-1})\sum_i v_i\frac{\partial \mathsf{K}}{\partial x_i}\mathbf{u}$ is the same in each component of the vector. Actually, we may consider the scalar v_i as the perturbation of design variable x_i. Consequently we can assemble the $\sum_i v_i\frac{\partial \mathsf{K}}{\partial x_i}$ into a single matrix $\Delta\mathsf{K}$. $(-\mathsf{K}^{-1})\sum_i v_i\frac{\partial \mathsf{K}}{\partial x_i}\mathbf{u}$ is actually $(-\mathsf{K}^{-1})\Delta\mathsf{K}\mathbf{u}$. Then by reanalysis method [24],

$$\Delta\mathsf{K}\mathbf{u} + \mathsf{K}\Delta\mathbf{u} = \mathbf{0}, \tag{58}$$

from which the displacement increment $\Delta\mathbf{u}$ may be calculated.

Consequently, Eq.(57) is estimating $\Delta r_i = \mathsf{D}_i\mathsf{B}_i\mathsf{T}_i\Delta\mathbf{u}$:

$$\mathsf{G}^T(\mathbf{x}_k)\cdot\mathbf{v} = \begin{bmatrix} \Delta r_1 \\ \Delta r_2 \\ \vdots \\ \Delta r_n \end{bmatrix}\Bigg|_{\mathbf{x}_k} \tag{59}$$

The most expensive part in this computation is to inverse the stiffness matrix K which the computational cost is $O(n)$. In this method, we need to inverse K only *once* in the reanalysis method Eq.(58) to obtain Eq.(59). Therefore the computational work is $O(n)$ to calculate $\mathsf{G}^T\cdot\mathbf{v}$ once.

Both the cost for $\mathsf{G}\cdot\mathbf{v}$ and $\mathsf{G}^T\cdot\mathbf{v}$ is $O(n)$ now. By contrast, previously we have to repeat n times to inverse K in adjoint method to get stress sensitivity first and then to complete gradient matrix $\mathsf{G}(\mathbf{x})$. The computational work is $O((n)^2)$. Therefore the cost to get $\mathsf{G}\cdot\mathbf{v}$ and $\mathsf{G}^T\cdot\mathbf{v}$ has been decreased effectively.

In the procedure to set up the subproblem and stress approximate, the implicit sensitivity analysis procedure mentioned above can also be applied when $\mathsf{G}(\mathbf{x})\cdot\mathbf{v}$ or $\mathsf{G}(\mathbf{x})^T\cdot\mathbf{v}$ are required.

The explicit stress sensitivity analysis which is a very expensive part is successfully removed inside the optimisation framework. The structure of the proposed method are simplified to only FEM and iterative optimisation with the implicit sensitivity analysis. Rather than completely separated, the FEM and the iterative optimisation are linked by the reanalysis and adjoint method in the implicit sensitivity analysis. Inside the optimisation, adjoint method and the reanalysis employ the stiffness matrix of FEM to update the Lagrange multipliers and design variables, the stress approximation updates the mechanical responses $(\mathbf{u}, \sigma)$ by reanalysis method with the updated design.

3.3 Summary of the Optimization Algorithm

The computational cost of the iterative optimisation has been efficiently reduced by the PCG with the proposed preconditioner. Meanwhile, the computational efficiency for the sensitivity analysis has also been improved by the implicit sensitivity analysis. Therefore the entire optimisation procedure can be accelerated. In this subsection, the overall optimisation algorithm is summarised.

The outermost loop has the design point changed according to the method of successive approximations.

Algorithm 2 Algorithm for the proposed optimisation method

1: Initialize the design variable $\mathbf{x_0}$
2: **while** $vol \geq tol$ **do** ▷ stop if the volume variation tol
3: Finite element analysis Eq. (12)
4: Update $\mathbf{x} \leftarrow$ the solution of predictor-corrector interior point method ▷ see Section 2.3
5: **end while**

In the proposed optimisation method, there are only two main operations, the FEM and the iterative optimisation. The procedure iterates until the relative change of the volume of the structure is smaller than the predefined tolerance. The explicit sensitivity analysis module is removed from the optimisation procedure. The solution of the approximate subproblem remains the same as in Section 2.3. The only difference is that the PCG method is employed to solve the system of equations in Eq. (45) iteratively in the predictor and corrector steps of the algorithm. Since no explicit gradient matrix is available in the proposed method, the implicit sensitivity analysis is deployed every iteration step in the interior point method (Eqs. (56) and (59)).

The number of iteration steps for both the interior point method and PCG is small and not influenced by the size of optimization problems [20][25]. The computational cost for the optimisation will be reduced significantly.

In the next section, the computational complexity of the proposed method will be demonstrated.

3.4 Computational Complexity for the Efficient Optimization Method

The computational complexity of the proposed efficient optimization method will be analysed in this section to estimate how much improvement can be obtained. The most expensive parts in the optimisation module are the implicit sensitivity analysis and the PCG iterations.

For the implicit sensitivity analysis, the most expensive cost is for the reanalysis method and the adjoint method (Eq. (56,59)). Thus the cost of implicit sensitivity analyses is,

$$C_2' = O(n),$$

per iteration.

In the PCG, since the full coefficient matrix is not required, the computational work $(O(n^3))$ for the matrix matrix product is saved. The most costly part is the use of implicit sensitivity analysis, already discussed, and the inversion of the preconditioner. Since the preconditioner is a diagonal matrix then the cost of one PCG iteration is,

$$C_3' = O(n).$$

The number of iteration steps for both the interior point method and the PCG is small and is *assumed* not to depend on the size of optimization problem. Therefore, the total number of iterations of the PCG method $n_{iter} \ll n$ when n is large enough. The computational cost for the complete optimisation is therefore $O(n)$. The proposed method, successfully reduces the computational complexity for stress constrained optimization problems from $O(n^3)$ to $O(n)$.

4 Computational Result

In this section, two numerical cases will be utilized to show how much do we achieve with the proposed method. The target structures are statically indeterminate beams. In both cases, we repeat the optimization procedure with increasing number of elements in FEM model. All tests are implemented on an DEll pc with 8.0 GB installed memory, Intel Core i5-4670 CPU @ 3.40 GHz 3.40 GHz. The CPU time for the second test optimization procedure is recorded and compared between the original method and the proposed method in this paper. The numerical cases show that the proposed method not only produce feasible optimal solution but also successfully decreases the CPU time of optimization procedure by one order.

4.1 Two-span Continuous Beam

We employ the proposed method to complete two numerical cases. The first numerical case is the two-span continuous beam in Figure 1. The width w in the rectangular cross-section A-A is the design variable while the height h of the cross-section is fixed in the optimization problem. In this numerical case, apart from the stress constraints, the displacement in y direction at $7370mm$ is also constrained apart from the stress constraints and the lower and upper bound of the design variables.

The optimization problem is defined as:

Objective:

$$\text{Min:} f_0(\mathbf{w}) = h \sum_{i=1}^{n} l_i w_i \tag{60}$$

subject to:

$$\begin{aligned} f_i(\mathbf{w}) \leq \sigma_{all} \quad & i = 1 \ldots \text{n, n is the number of stress constraints} \\ w_l \leq w_i \leq w_u \quad & i = 1 \ldots \text{n, n is the number of design variables} \\ u_{con}(\mathbf{w}) \leq u_a \quad & u_{con}(\mathbf{w}) \text{ is the displacement at 7370mm from the left} \end{aligned}$$

The parameters of the optimization problem is listed in Table 1.

Table 1: Parameters of Two-span continuous beam

Young's Modulus	$E = 20Gpa$
Allowable stress	$\sigma_{all} = 5Mpa$
Allowable displacement/span ratio	$r = 1/3500$
Initial design width	$t_i = 1000mm$
Minimum width	$t_l = 100mm$
Maximum width	$t_u = 1000mm$
Distributed load	$p1 = 35KN/m, p2 = 25KN/m$

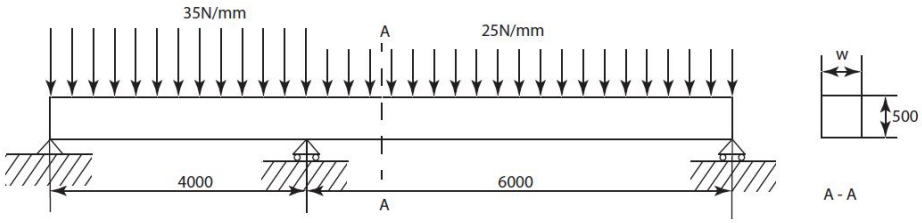


Figure 1: Two-span Continuous Beam

In this optimisation, the relative error of PCG is $err_{PCG} = 1e^{-4}$, the tolerance of duality gap is $tol_d = 1e^{-5}$, the tolerance of the volume variation is $tol_v = 1e^{-3}$. The optimal solution of the width is shown in Figure 2. It keeps at the minimum where both stress and the displacements constraints are not active. Then compared with Figure 34, the width of the beam at $1272mm$ and $4000mm$ increases because of the active of stress constraints. Then the width of the beam increases again at $7334mm$. However the increment of width in this part is due to the active of displacement constraints as demonstrated in Figure 5.

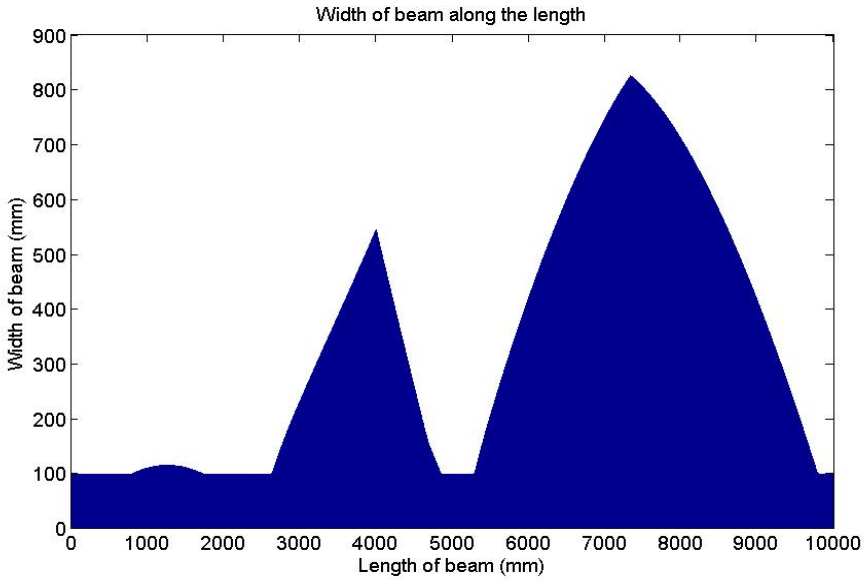


Figure 2: The width along the length of the optimal beam

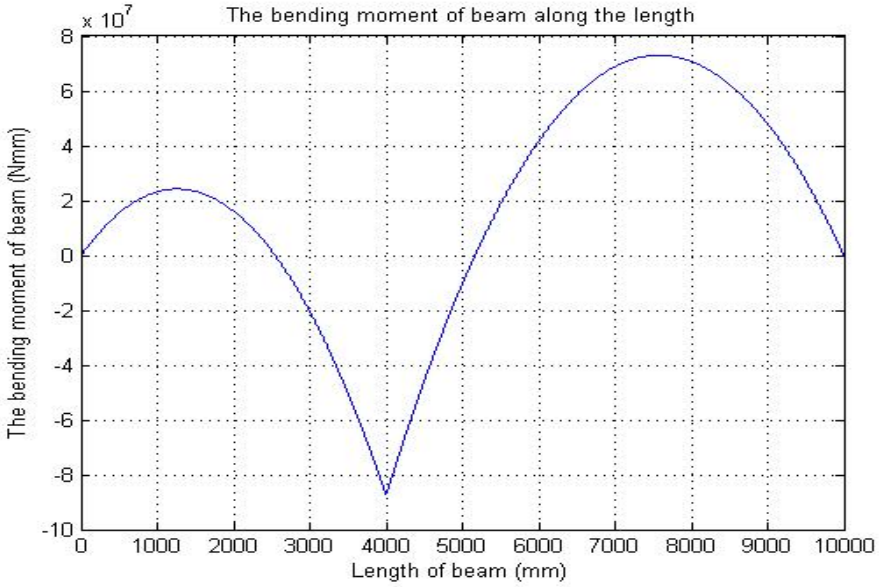


Figure 3: the bending moment in the optimal beam

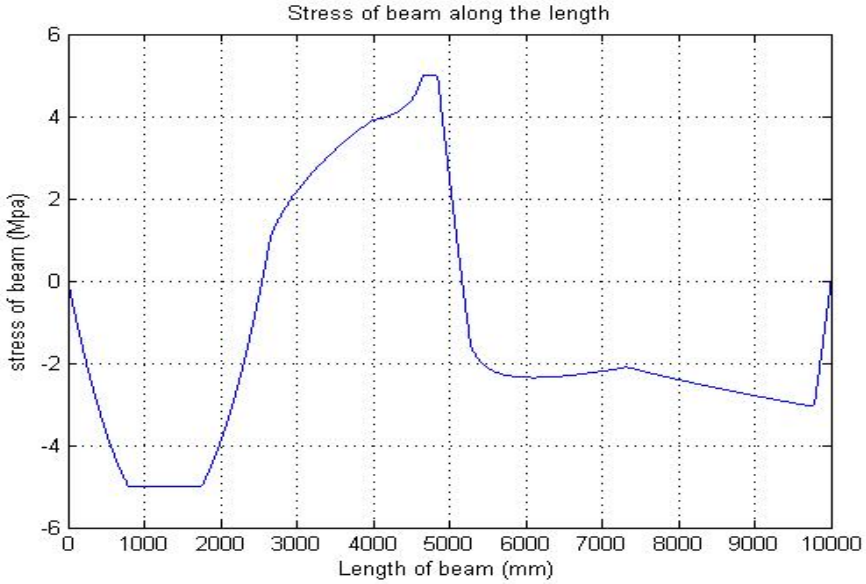


Figure 4: the normal stress on the top of the optimal beam

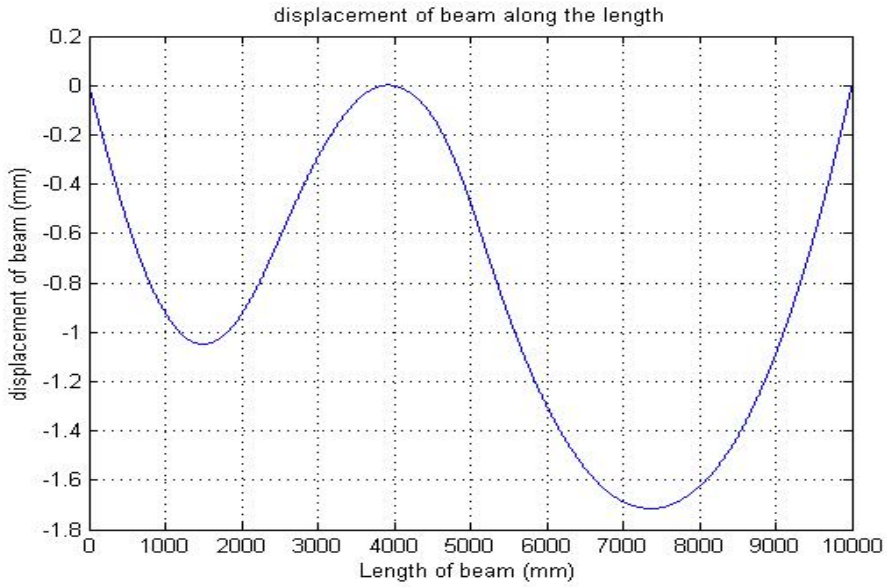


Figure 5: the displacement of the optimal beam

The volume of the beam decreases during the optimization as shown in Figure 6. It drops rapidly in the first three steps before it converges at eighth steps.

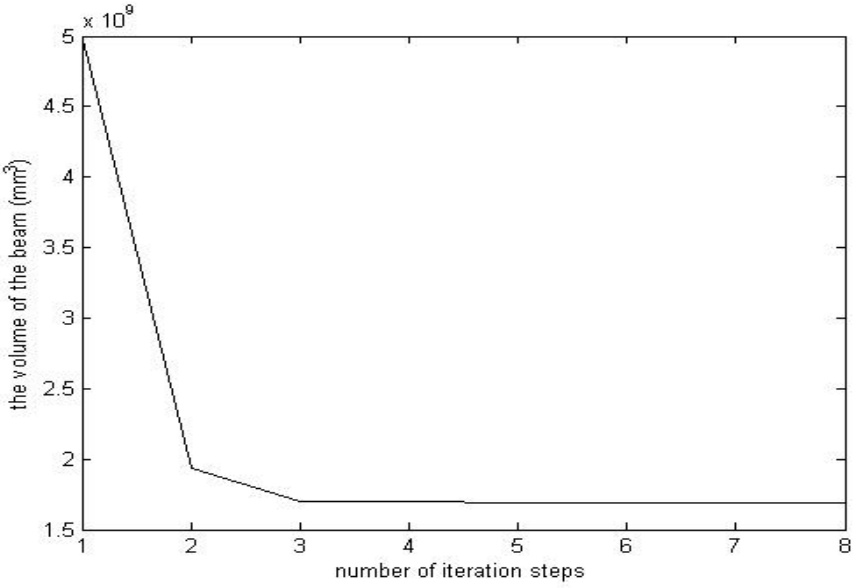


Figure 6: The volume of the beam in each iteration

4.2 Hollow Beam Structure

In this case, the design variables are the thickness t of the hollow beam shown in Figure 7. The left edge of the beam is fixed and the right edge is simply supported. Uniform distributed load is applied on the top. While design the beam with minimum volume of material, the stress within each element is constrained and the design variables are confined within the predefined range. The optimization model is defined:

Objective:

$$\textit{Min:} f_0(t) = S\sum_{i=1}^{n} l_i t_i \quad S \text{ is the perimeter of the cross-section area} \tag{61}$$

subject to:

$$f_i(t) \leq \sigma_{all} \quad i = 1 \ldots \text{n, n is the number of stress constraints}$$
$$t_l \leq t_i \leq t_u \quad i = 1 \ldots \text{n, n is the number of design variables}$$

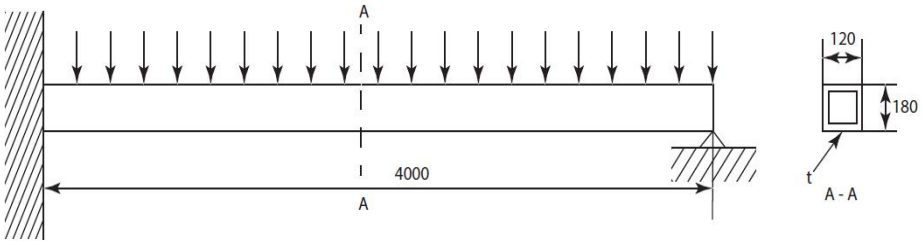


Figure 7: Hollow Beam Structure

The optimization starts with uniform thickness distribution. The parameters of the optimization problem is listed in Table 2

Table 2: Parameters of hollow beam structure

Young's Modulus	$E = 200Gpa$
Allowable stress	$\sigma_{all} = 172.36Mpa$
Initial design thickness	$t_i = 20mm$
Minimum thickness	$t_l = 3mm$
Maximum thickness	$t_u = 20mm$
Distributed load	$p = 40N/mm$

In this case, the relative error of PCG is $err_{PCG} = 1e^{-4}$, the tolerance of duality gap is $tol_d = 1e^{-5}$, the tolerance of the volume variation is $tol_v = 1e^{-3}$. The optimal thickness along the length of the beam is depicted in Figure 8. It first decreases from $16.7mm$ from the left edge to $3mm$ at around $1000mm$. Then it increases to $7.2mm$ at around $2600mm$ before it reduces to $3mm$ again.

The internal force of the beam, i.e., the bending moment here, is illustrated in Figure 9. On the left edge, it reaches $-9.31 \times 10^7 Nmm$, then it increases to $0Nmm$ and then reaches the $4 \times 10^7 Nmm$ at about $2500mm$. It goes back to $0Nmm$ again on the right edge.

Therefore the distribution of the thickness matches the trend of bending moment to confine the stress. From Figure 10, which shows the normal stress on the top edge of the beam along the length, the normal stress reaches the allowable value where the stress constraints are active.

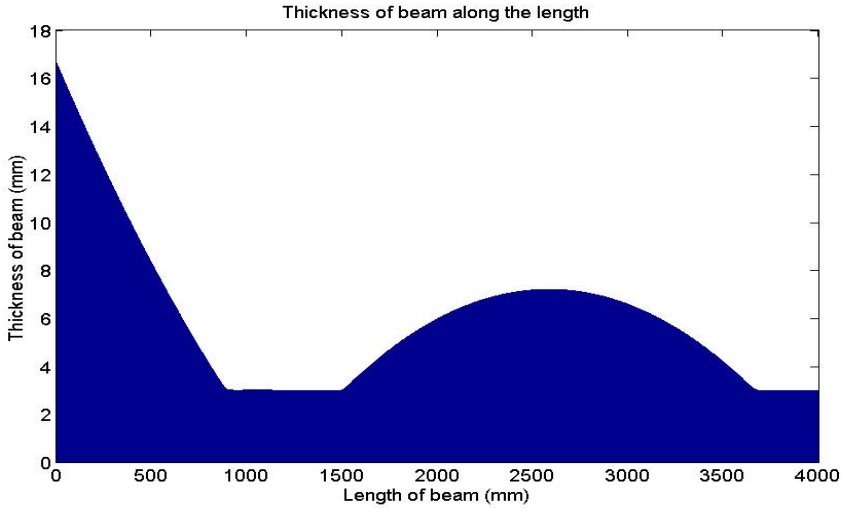


Figure 8: thickness along the length of the optimal beam

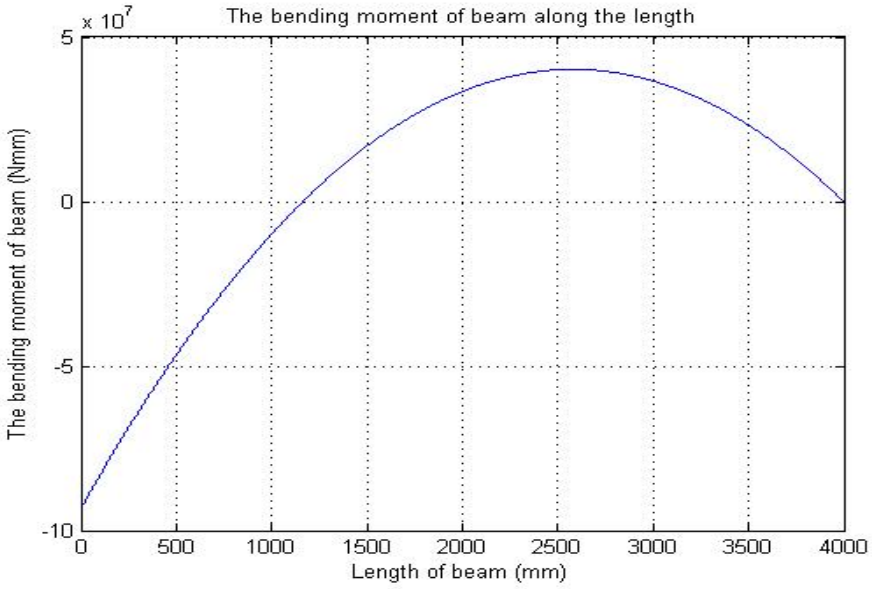


Figure 9: the bending moment in the optimal beam

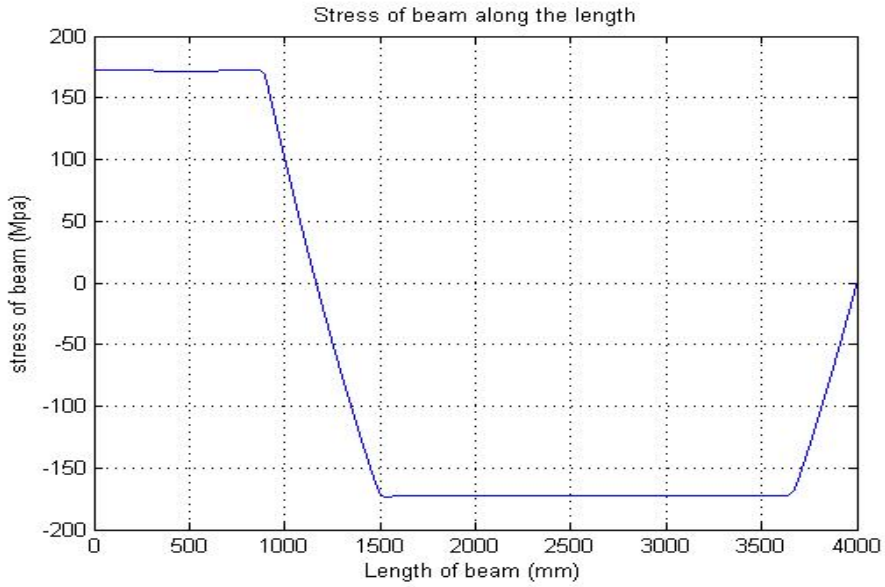


Figure 10: the normal stress on the top of the optimal beam

The variation of the volume during optimization is shown in Figure 11. The volume drops rapidly in the first two loops. It takes merely 6 steps in total to converge to the optimal solution.

In this case, we record the number of iteration steps in the PCG, the interior point method and the entire optimisation when the structure is discretised by different number of elements. We start the procedure from 20 elements and double the number of elements each time. From Figure 12, there is no obvious change for the number of steps in the entire optimization iteration, the interior point and the PCG when the number of elements increases from 20 to 10240. The number of steps for the PCG is the averaged total number of the PCG iterations per interior point method step. The CPU time for the optimization procedure with the original method and the proposed method is portrayed in Figure 13. The dashed line in the figure demonstrates the linear trend between the CPU time and the number of elements. As we can see from the figure, the trend of the solid line is almost parallel to the dashed line. It proves that computational complexity for the whole procedure has been decreased from more than $O(n^2)$ to $O(n)$. Therefore great efficiency is achieved by the proposed method is for large scale stress constrained optimization.

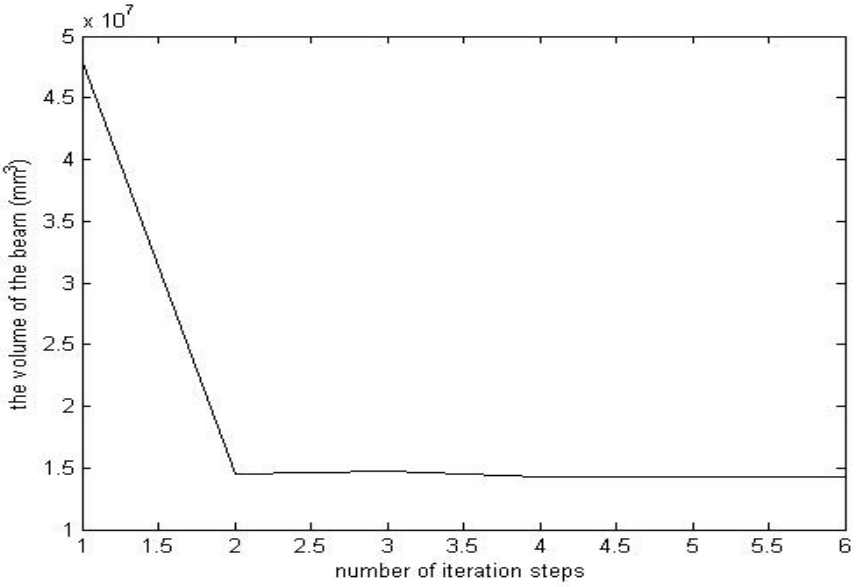


Figure 11: The volume of the beam in each iteration

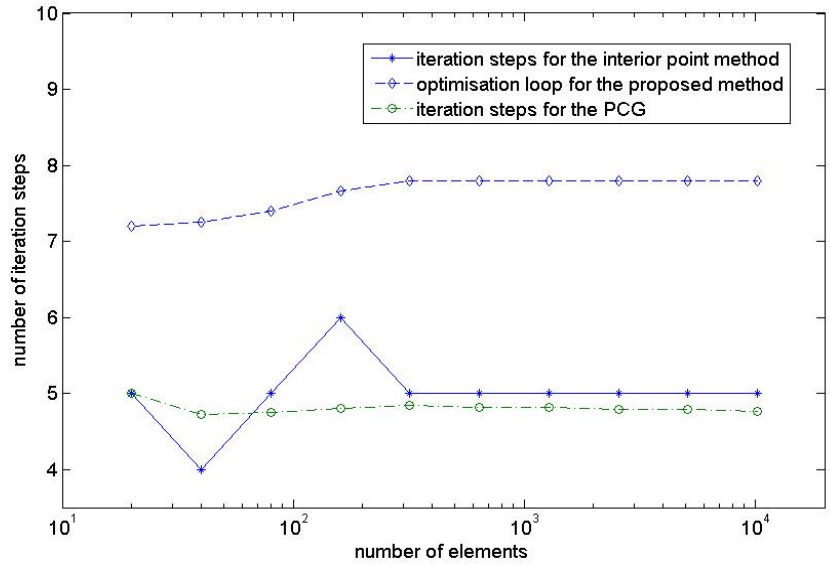


Figure 12: the Number of steps for entire optimization,interior point method and PCG

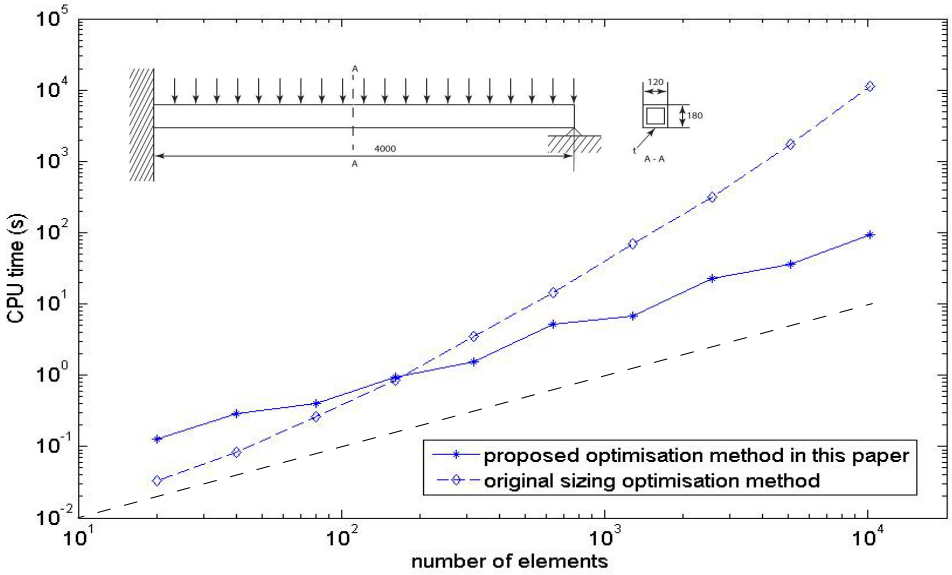


Figure 13: the log-log form of the CPU time vs number of elements for the optimization with original method and the proposed method

5 Conclusion

An efficient method for large-scale beam sizing under stress constraints is introduced in this paper. A novel successive convex stress approximation which is homogenous and separable is described first. This proposed approximation not only provides sufficient accuracy but also makes the subproblem cheap to build in the optimisation module.

The PCG is employed to enhance the efficiency of solving the subproblem in the optimisation module for the large scale optimisation problems. An effective preconditioner based on the fully stressed concept is introduced. Numerical test in Section 4.2 proves that the PCG converges only a few steps to solve the linear equation with the proposed preconditioner.

An efficient implicit sensitivity analysis method is proposed in this method to replace the expensive sensitivity analysis by the adjoint method. The structure of the optimisation framework therefore is simplified to only FEM analysis and iterative optimisation with the interior point method. The computational effort for sensitivity related calculation has been decreased from $O(n^2)$ to $O(n)$.

The implicit sensitivity analysis is deployed each time the gradient matrix G is required in the iterative optimisation procedure. Since the total number of iterations of the PCG method in the interior point method is $n_{iter} \ll n$ when n is large enough. The computational cost for the entire optimisation drops from $O(n^3)$ to $O(n)$.

Great efficiency for stress constrained large scale optimisation is achieved with the proposed method. Numerical tests have validated the conclusion from the theoretical analysis of computational complexity. For the future research, the method is supposed to be extended to plate elements to strength the problem solving capacity of the proposed method.

REFERENCES

[1] Geoffrey G. Pope, Optimum Design of Stressed Skin Structures. *AIAA Journal*, **11**, 1545-1552, 1973.

[2] Raphael T. Haftka, Simultaneous Analysis and Design. *AIAA Journal*, **23**, 1099-1103,1985.

[3] Attila P. Nagy, Mostafa M. Abdalla, Zafer Gürdal, Isogeometric sizing and shape optimisation of beam structures. *Computer Methods in Applied Mechanics and Engineering*, **199**, 1216-1230, 2010.

[4] R. Sedaghati, A. Suleman, and B. Tabarrok, Structural Optimization with Frequency Constraints Using the Finite Element Force Method. *AIAA Journal*, **40**, 382-388,2002.

[5] L.J. Li and Z.B. Huang and F. Liu and Q.H. Wu, A heuristic particle swarm optimizer for optimization of pin connected structures. *Computers & Structures*, **85**, 340-349, 2007.

[6] Charles Camp, Shahram Pezeshk, Guozhong Cao, Optimized Design of Two-Dimensional Structures Using a Genetic Algorithm. *Journal of Structural Engineering*, **124**, 551-559, 1998.

[7] A.G.M. Michell M.C.E., LVIII. The limits of economy of material in frame-structures. *Philosophical Magazine Series 6*, **8 (47)**, 589-597, 1904.

[8] RAZANI, R, Behavior of fully stressed design of structures and its relationshipto minimum-weight design. *AIAA Journal*, **3 (12)**, 2262 - 2268, 1965.

[9] L.A. Schmit, B. Farshi, Some Approximation Concepts for Structural Synthesis. *AIAA Journal*, **12 (5)**, 692 - 699, 1974.

[10] Fadel, GM and Riley, MF and Barthelemy, JM, Two point exponential approximation method for structural optimization. *Structural Optimization*, **2 (2)**, 117–124, 1990.

[11] Zhou, M and Xia, RW, Two-level approximation concept in structural synthesis. *International Journal for Numerical Methods in Engineering*, **29 (8)**, 1681–1699, 1990.

[12] VANDERPLAATS and GARRET N, New approximation method for stress constraints in structural synthesis. *AIAA Journal*, **27 (3)**, 352–358, 1989.

[13] Vanderplaats, G.N. and Thomas, H.L., An improved approximation for stress constraints in plate structures. *Structural optimization*, **6 (1)**, 1-6, 1993.

[14] Etman, LFP and Groenwold, Albert A and Rooda, JE, On diagonal QP subproblems for sequential approximate optimization. *eighth world congress on structural and multidisciplinary optimization*, Lisboa, Portugal, 2009.

[15] Roberts Jr, Ronald W and Canfield, Robert A, Accuracy of Enriched Multipoint Cubic Approximations for Large-Scale Optimization. *51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference 18th AIAA/ASME/AHS Adaptive Structures Conference 12th*, 2010.

[16] Fleury, Claude, First and second order convex approximation strategies in structural optimization. *Structural Optimization*, **1 (1)**, 3–10, 1989.

[17] Boggs, Paul T and Kearsley, Anthony J and Tolle, Jon W, A practical algorithm for general large scale nonlinear optimization problems. *SIAM Journal on Optimization*, **9 (3)**, 755–778, 1999.

[18] Vanderbei, Robert J and Shanno, David F, An interior-point algorithm for nonconvex nonlinear programming. *Computational Optimization and Applications*, **13 (1-3)**, 231–252, 1999.

[19] Sanjay Mehrotra, On finding a vertex solution using interior point methods. *Linear Algebra and its Applications*, **152**, 233 - 253, 1991.

[20] Sanjay Mehrotra, On the Implementation of a Primal-Dual Interior Point Method. *SIAM Journal on Optimization*, **2 (4)**, 575-601, 1992.

[21] E. D. Andersen and C. Roos and T. Terlaky, On implementing a primal-dual interior-point method for conic quadratic optimization. *MATHEMATICAL PROGRAMMING SER. B*, **95 (95)**, 249–277, 2000.

[22] Zillober, Christian, A Combined Convex ApproximationInterior Point Approach for Large Scale Nonlinear Programming. *Optimization and Engineering*, **2 (1)**, 51-73, 2001.

[23] Schmit, Lucien A and Lai, Ye-Chen, International journal for numerical methods in engineering. *Wiley Online Library*, **37 (6)**, 943–964, 1994.

[24] R. T. Haftka, Z. Gürdal, Elements of structural optimization. Vol. 11, *Springer Science & Business Media*, 2012.

[25] I. J. Lustig, R. E. Marsten, D. F. Shanno, On implementing mehrotra's predictor-corrector interior-point method for linear programming. *SIAM Journal on Optimization*, **2 (3)**, 435–449, 1992.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOPOLOGY OPTIMIZATION OF WAVE BARRIERS FOR RAILWAY INDUCED VIBRATIONS IN BUILDINGS

Cédric Van hoorickx[1], Mattias Schevenels[2], and Geert Lombaert[1]

[1]KU Leuven, Department of Civil Engineering, Structural Mechanics Section
Kasteelpark Arenberg 40, 3001 Leuven, Belgium
e-mail: {cedric.vanhoorickx,geert.lombaert}@bwk.kuleuven.be

[2] KU Leuven, Department of Architecture, Architectural Engineering
Kasteelpark Arenberg 1, 3001 Leuven, Belgium
e-mail: mattias.schevenels@asro.kuleuven.be

Keywords: Topology optimization, Railway induced vibration, Elastodynamic wave propagation, Wave barriers.

Abstract. *In order to reduce environmental ground vibration due to railway traffic, mitigation measures at the source, the transmission path and the receiver can be applied. Mitigation measures on the transmission path are particularly appealing in situations with existing track and buildings. In this contribution, the optimal design of stiff wave barriers is considered.*

Up to now, wall barriers with a rectangular cross section have mostly been studied. However, current construction methods of wave barriers provide a wide flexibility in design geometry. Therefore further improvement is feasible. Topology optimization is used to discover novel, improved design geometries.

Stiffer material is inserted into a design domain located between the source and a building. The objective is to minimize the amount of stiffer material in the design domain while satisfying the required vibration criteria in the building. The optimization problem is solved using a gradient based method and the adjoint method is used to enable an efficient calculation of the sensitivities.

In this way, an optimal geometry is determined that considerably reduces the required amount of stiffer material. Topology optimized stiff wave barriers are therefore effective in reducing vibration levels, outperforming the rectangular wave barriers. The design contains, however, small features, making it sensitive to geometric imperfections, and a worst case robust optimization is applied. The resulting smaller sensitivity with respect to geometric imperfections also ensures that a posteriori design simplifications have a minimal impact on the barriers performance.

1 INTRODUCTION

Railway induced vibrations may lead to the malfunctioning of sensitive equipment, discomfort and structural damage at high vibration levels. Furthermore, noise can be re-radiated from floors and walls. This paper focuses on malfunctioning of sensitive equipment. Extensive research is performed to search for effective vibration mitigation measures. Three categories of mitigation measures are distinguished, namely at the source [1], the transmission path [2, 3] and the receiver [4]. These refer respectively to the railway track, the soil and the building. The present work considers wave barriers to hinder ground vibration propagation from source to receiver.

Currently, only a limited number of simple design geometries for mitigation measures have been investigated. However, current flexibility in construction methods, such as jet grouting, provides opportunities for further improvement. To discover novel design geometries, topology optimization is applied. Topology optimization, originally developed for static mechanical problems, is used for a variety of applications, including problems governed by wave propagation [5]. Examples include electromagnetic, elastomagnetic and acoustic applications, such as photonic crystal waveguide designs [6], designs of phononic band-gap materials [7] and noise barriers [8].

This study uses topology optimization to design barriers impeding elastodynamic wave transmission. Wave barrier effectiveness is quantified by comparing the calculated one-third octave band velocity spectra in specific receiver points in the building to generic vibration criteria for sensitive equipment [9]. Since full control of the built geometry is difficult in the construction of underground structures, the optimized design should be robust with respect to geometrical imperfections [10]. A worst case approach is applied to improve the robustness of the design [11, 12]. The resulting designs can be used to realize simplified design solutions, which are suited for construction and are robust with respect to geometric imperfections.

The paper starts with the introduction of the optimization problem and a brief review of the topology optimization approach. Next, results for optimized wave barriers are discussed. Finally, the sensitivity to geometric imperfections is analyzed and a simplified design is presented.

2 PROBLEM DESCRIPTION AND METHODOLOGY

2.1 Problem description

The considered problem is shown in figure 1 and consists of a building located at the surface of a homogeneous elastic halfspace, representing the soil. The soil (subscript '1', as abbreviation for material 1) has a mass density ρ_1 of $2000\,\mathrm{kg/m^3}$, a longitudinal wave velocity C_{p1} of $400\,\mathrm{m/s}$, and a shear wave velocity C_{s1} of $200\,\mathrm{m/s}$. The building (subscript 'bu') is a four-storey frame consisting of four floors and two spans, with a total width w_{bu} of $12\,\mathrm{m}$ and a total height h_{bu} of $12\,\mathrm{m}$. The thickness of the walls $t_{\mathrm{bu,w}}$ is equal to $0.25\,\mathrm{m}$, while the thickness of the slabs $t_{\mathrm{bu,s}}$ is equal to $0.20\,\mathrm{m}$. The walls and slabs are made of reinforced concrete with a mass density ρ_{bu} of $2500\,\mathrm{kg/m^3}$, a Young's modulus E_{bu} of $30\,\mathrm{GPa}$, and a Poisson's ratio ν_{bu} of 0.25.

The halfspace is excited at the surface by a vertical load at a distance of $25\,\mathrm{m}$ from the building. A line load with a uniform spectrum in the frequency range between $4\,\mathrm{Hz}$ and $8\,\mathrm{Hz}$ is applied. The problem can be modeled as a two-dimensional problem and is shown in figure 2. The aim is to reduce the vibration levels at specific receiver points in the building, in this case on the first floor in the span closest to the source. A horizontal receiver is considered in the middle of the floor, while two vertical receivers are considered at one fourth and in the middle

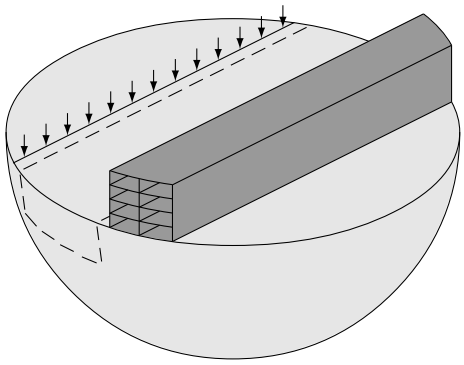


Figure 1: Semi-infinite halfspace excited at the surface close to a building.

of the floor span.

To reduce the vibration levels in these receiver points, a design domain with dimensions $8 \times 20\,\mathrm{m}^2$ is considered in the elastic halfspace, located between the source and the building. In this design domain a second stiffer material is introduced. This material (subscript '2') has a mass density ρ_2 of $2000\,\mathrm{kg/m}^3$, a longitudinal wave velocity $C_{\mathrm{p}2}$ of $950\,\mathrm{m/s}$, and a shear wave velocity $C_{\mathrm{s}2}$ of $550\,\mathrm{m/s}$.

The elastodynamic problem is solved using the finite element method with two-dimensional four-node elements in plane strain. For the mesh, an element size of $0.25\,\mathrm{m}$ is used, corresponding to ten elements per shear wavelength $\lambda_{\mathrm{s}1}$ at a frequency of 80Hz, the upper limit considered in this paper. In the design domain, however, the mesh size is reduced to $0.125\,\mathrm{m}$ to achieve a design with sufficient detail. The building is modeled using beam elements adapted to plane strain with an element size of $0.25\,\mathrm{m}$. The finite element equilibrium equations are written as:

$$\hat{\mathbf{K}}\hat{\mathbf{u}} = \hat{\mathbf{p}} \tag{1}$$

where $\hat{\mathbf{p}}$ is the load vector, $\hat{\mathbf{u}}$ is the displacement vector, and $\hat{\mathbf{K}}$ is the dynamic stiffness matrix, which depends on the frequency ω. At the boundaries of the finite element mesh, appropriate radiation boundary conditions have to prevent spurious wave reflections. The Perfectly Matched Layers (PML) of Harari and Albocher [13] are applied.

The one-third octave band velocity spectra are computed from the velocity spectra in the frequency domain:

$$\hat{v}_{m,i}^{\mathrm{RMS}} = \sqrt{\frac{1}{N_{\mathrm{f}}} \sum_{n=1}^{N_{\mathrm{f}}} \left|\hat{v}_{m,n,i}\right|^2} \tag{2}$$

where N_{f} is the number of frequencies in the one-third octave band with index m and i is the position of the receiver. This can be expressed as follows (since $\hat{v}_{m,n,i} = \mathrm{i}\omega\hat{u}_{m,n,i}$ with i the imaginary unit):

$$\hat{v}_{m,i}^{\mathrm{RMS}} = \sqrt{\frac{1}{N_{\mathrm{f}}} \sum_{n=1}^{N_{\mathrm{f}}} \omega_{m,n}^2 \left|\hat{u}_{m,n,i}\right|^2} = \sqrt{\frac{1}{N_{\mathrm{f}}} \sum_{n=1}^{N_{\mathrm{f}}} \omega_{m,n}^2 \hat{\mathbf{u}}_{m,n}^H \mathbf{L}_i \hat{\mathbf{u}}_{m,n}} \tag{3}$$

where $\mathbf{L}_i$ is a sparse selection matrix, equal to one at the diagonal element corresponding to the degree of freedom of receiver i.

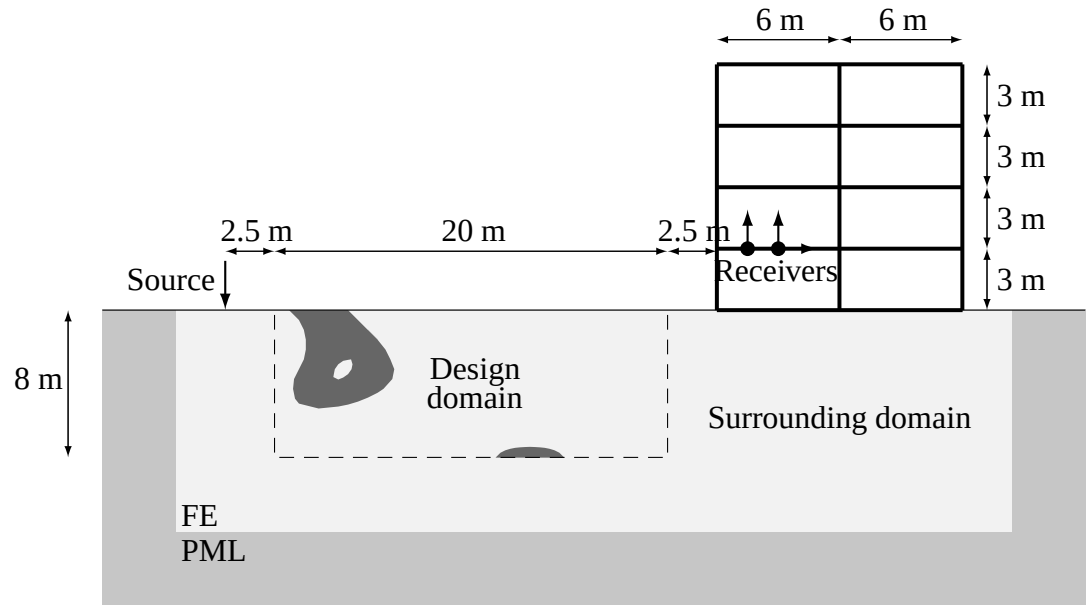


Figure 2: The two-dimensional optimization problem.

2.2 Topology optimization

Topology optimization [5] is used to optimally distribute the stiffer material in the design domain. The material distribution is parameterized using element densities $\bar{\rho}_e$ for each element e in the design domain. The value $\bar{\rho}_e = 0$ corresponds to the properties of the homogeneous halfspace, while the value $\bar{\rho}_e = 1$ corresponds to the properties of the stiffer material. Using a continuous interpolation of the densities between 0 and 1 makes it possible to solve the optimization problem with a gradient based method.

A Solid Isotropic Material with Penalization (SIMP) interpolation [14, 15] is used to interpolate the material properties α:

$$\alpha = \alpha_1 + \bar{\rho}_e^p \left(\alpha_2 - \alpha_1\right) \tag{4}$$

where α_1 and α_2 are the properties of the original and stiffer material, respectively. The penalization factor $p \geq 1$ avoids so-called gray designs with intermediate densities. A value $p = 1$ is used for the mass density ρ while $p = 3$ is used for the constrained modulus ρC_{p}^2 and the shear modulus ρC_{s}^2.

The (physical) element densities $\bar{\rho}_e$ are obtained from the design variables ρ_e of the optimization problem by applying a projection filter [16]:

$$\bar{\rho}_e = \frac{\tanh\left(\beta\eta\right) + \tanh\left(\beta(\tilde{\rho}_e - \eta)\right)}{\tanh\left(\beta\eta\right) + \tanh\left(\beta(1 - \eta)\right)} \tag{5}$$

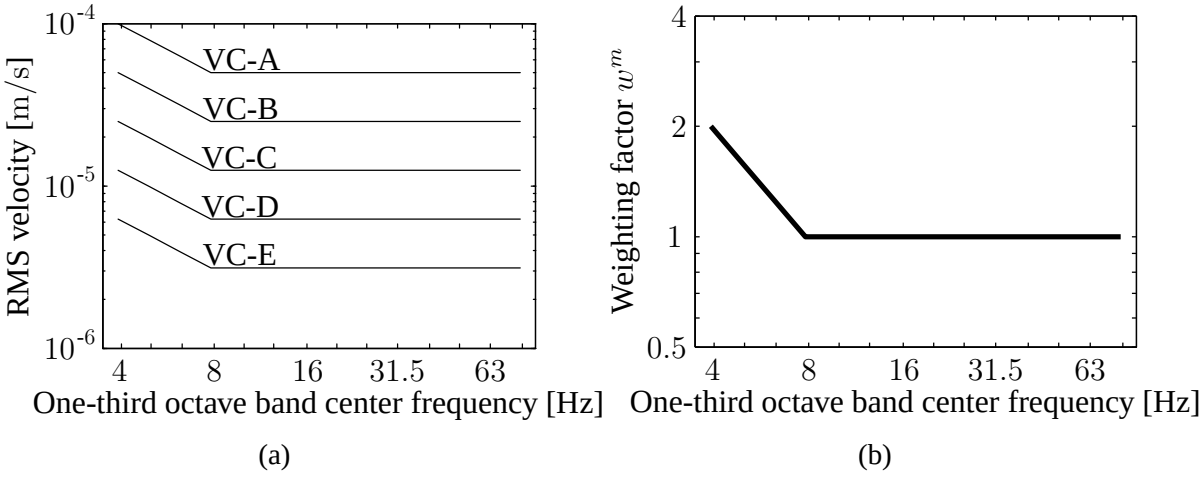
with β a sharpness parameter, controlling the smoothness of the projection, η the projection threshold, and $\tilde{\rho}_e$ the intermediate densities obtained from:

$$\tilde{\rho}_e = \frac{\sum_{i=1}^{N} w_{ei} v_i \rho_i}{\sum_{i=1}^{N} w_{ei} v_i} \tag{6}$$

where v_i is the volume of element i and the weight $w_{ei} = \max(R - r_{ei})$ depends on the filter radius R and the center-to-center distance r_{ei} between the elements e and i. In the present work, the filter radius is taken to be $R = 6$ elements or $1.5\,\mathrm{m}$, the projection threshold value is set to $\eta = 0.5$ and the sharpness parameter β has an initial value equal to 1 and is continuously increased (multiplied by a factor 1.0116)such that the value of 32 is reached after 300 iterations.

2.3 Optimization problem

To quantify the effectiveness of the wave barriers, the calculated one-third octave band velocity spectra are compared with generic vibration criteria (VC) for sensitive equipment [9]. These criteria are specified in terms of the maximum allowable RMS velocity in the one-third octave bands between 4 and 80 Hz, as shown in figure 3a. The VC criteria impose less stringent vibration limits in the frequency range between 4 to 8 Hz, where the limit is that of constant acceleration instead of constant velocity.



(a) (b)

Figure 3: (a) Gordon's generic vibration criteria for sensitive equipment [9] and (b) the weighting function w_m proportional to the VC curves.

The aim is to reduce the vibration levels to meet the intended vibration criteria. However, since there is a cost in the amount of material that is stiffened, the volume of the wave barrier is minimized. The problem is reformulated as a minimization problem, where the volume is minimized and constraints are set on the vibration levels:

$$\begin{aligned} \min_{\rho_e} \quad & \sum_{e=1}^{N} v_e \bar{\rho}_e \\ \text{s. t.} \quad & \hat{v}_{m,i}^{\mathrm{RMS}}(\bar{\rho}_e) \leq w_m s, \quad m = 1 \cdots N_{\mathrm{band}} \\ & \qquad\qquad\qquad\quad\; i = 1 \cdots N_L \\ & 0 \leq \rho_e \leq 1, \qquad\quad e = 1 \cdots N_e \end{aligned} \tag{7}$$

with N_{band} the number of one-third octave bands considered, N_L the number of receiver points and N_e the number of elements in the design domain. The value s is the maximum allowable RMS velocity in the one-third octave bands between $8\,\mathrm{Hz}$ and $80\,\mathrm{Hz}$, for example equal to $12.5\,\mu\mathrm{m/s}$ for VC-C. The function w_m provides a correction for the one-third octave bands between $4\,\mathrm{Hz}$ and $8\,\mathrm{Hz}$, as is shown in figure 3b.

The optimization problem is solved using the method of moving asymptotes (MMA) [17]. Since this method is gradient-based, the derivatives of the objective function have to be calculated. The sensitivities are calculated using the chain rule:

$$\frac{\partial v_{m,i}^{\mathrm{RMS}}}{\partial \bar{\rho}_e} = \frac{1}{2N_{\mathrm{f}} v_{m,i}^{\mathrm{RMS}}} \sum_{n=1}^{N_{\mathrm{f}}} \omega_{m,n}^2 \frac{\partial \left[\hat{\mathbf{u}}_{m,n}^H \mathbf{L}_i \hat{\mathbf{u}}_{m,n}\right]}{\partial \bar{\rho}_e} \tag{8}$$

To efficiently calculate the derivative of the squared displacement modulus $\hat{\mathbf{u}}_{m,n}^H \mathbf{L}_i \hat{\mathbf{u}}_{m,n}$ to the physical densities $\bar{\rho}_e$, the adjoint method is used (see [5, p. 17]), resulting in:

$$\frac{\partial \left[\hat{\mathbf{u}}_{m,n}^H \mathbf{L}_i \hat{\mathbf{u}}_{m,n}\right]}{\partial \bar{\rho}_e} = 2\,\mathrm{Re}\left\{ \boldsymbol{\lambda}_{m,n,i}^T \frac{\partial \hat{\mathbf{K}}_{m,n}}{\partial \bar{\rho}_e} \hat{\mathbf{u}}_{m,n} \right\} \tag{9}$$

where the vector $\boldsymbol{\lambda}_{m,n,i}$ is computed from the adjoint equation, which can be written in the following simplified form:

$$\hat{\mathbf{K}}_{m,n} \boldsymbol{\lambda}_{m,n,i} = -\left(\frac{\partial \left[\hat{\mathbf{u}}_{m,n}^H \mathbf{L}_i \hat{\mathbf{u}}_{m,n}\right]}{\partial \hat{\mathbf{u}}_{m,n}} \right)^T = -\mathbf{L}_i \hat{\mathbf{u}}_{m,n}^* \tag{10}$$

3 OPTIMIZATION RESULTS

To show the effectiveness of a rectangular wave barrier, the influence of its thickness is analyzed. The rectangular wave barrier has a depth of $8\,\mathrm{m}$ and is located in the design domain as close as possible to the source, implying that the left boundary of the wave barrier is at a distance of $2.5\,\mathrm{m}$ from the source as follows from figure 2. The thickness is varied from $0\,\mathrm{m}$ to $20\,\mathrm{m}$, corresponding to the homogeneous halfspace and a fully filled design domain, respectively. This is done in steps of $0.5\,\mathrm{m}$.

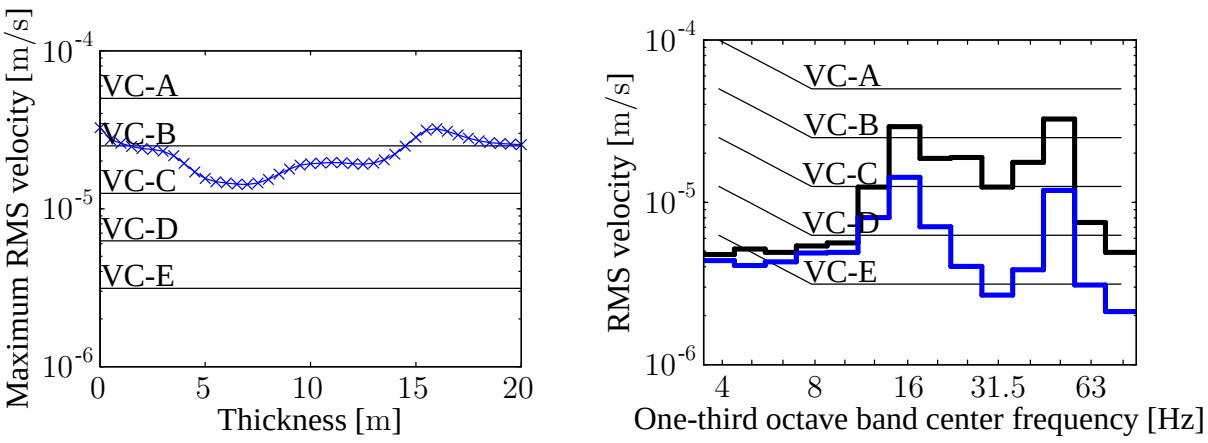


Figure 4: (a) Influence of the thickness of a rectangular wave barrier with a depth of $8\mathrm{m}$ on the maximum RMS velocity and (b) the maximum one-third octave band RMS spectra of the velocity at the receivers for the original homogeneous halfspace (black) and after the introduction of a rectangular wave barrier with a depth of $8\,\mathrm{m}$ and an optimal thickness of $7\,\mathrm{m}$ (blue).

Figure 4a shows the maximum RMS velocity for the different calculated thicknesses. It is clear that the problem is not convex, and has a global minimum around a thickness of $7\,\mathrm{m}$ in the

range considered. Figure 4b compares the one-third octave band RMS spectra of the velocity for the homogeneous halfspace and the rectangular wave barrier with the optimal thickness of $7\,\mathrm{m}$. The rectangular wave barrier provides a considerable reduction for the one-third octave bands with a center frequency higher than $12.5\,\mathrm{Hz}$, where vibration levels are reduced with a factor larger than 2 as compared to the homogeneous halfspace. However, if the VC-C criterion is aimed for, this cannot be achieved by placing a rectangular wave barrier. Moreover, the area of the cross section of the wave barrier leading to the optimal reduction is rather large, namely $8\,\mathrm{m} \times 7\,\mathrm{m} = 56\,\mathrm{m}^2$.

To overcome these limitations, the topology optimization problem in Eq. (7) is solved, where the volume is minimized and constraints are imposed on the vibration levels to meet the VC-C criterion. This is done by setting s equal to $12.5\,\mu\mathrm{m/s}$. The optimized design is shown in figure 5a. The design has a total cross sectional area equal to $6.8\,\%$ of the design domain, or $10.9\,\mathrm{m}^2$ in total.

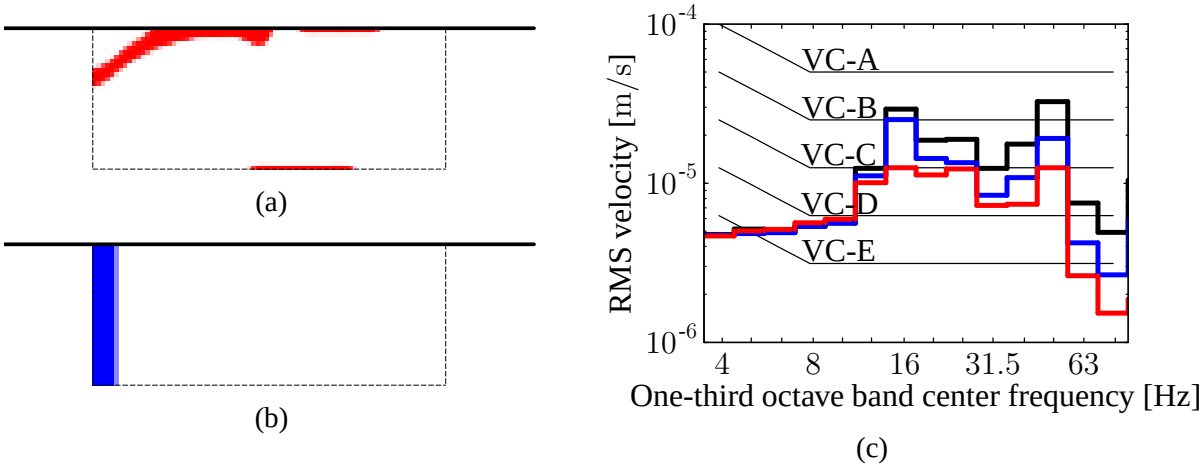


Figure 5: (a) Topology optimized design, (b) rectangular design with a depth of $8\,\mathrm{m}$ and the same volume as the optimized design, and (c) the maximum one-third octave band RMS spectra of the velocity at the receivers for the original homogeneous halfspace (black), after the introduction of the rectangular design in (b) (blue), and after the introduction of the optimal design in (a) (red).

In the optimized design of figure 5a, stiffer material is introduced in three main areas: (1) an inclined plate at the top left, (2) a horizontal plate at the surface, and (3) a horizontal plate at the bottom. The incoming waves are partly reflected by the stiffer material, partly redirected downwards into the soil by the inclined top part (1), and partly guided through the horizontal top part (2). At frequencies around $50\,\mathrm{Hz}$, a part of the waves redirected into the soil is reflected by the bottom part (3) and destructively interferes with the waves transmitted by the top part.

As the objective is to minimize the volume of the wave barrier, this optimized design is compared with a rectangular wave barrier with a depth of $8\,\mathrm{m}$ and the same volume as the optimized design. This rectangular design is shown in figure 5b. The maximum one-third octave band RMS spectra of the velocity at the receivers is compared for the two designs and for the homogeneous halfspace in figure 5c. For the homogeneous halfspace, the VC-B criterion is visibly exceeded, and only the VC-A criterion is met. Also for the rectangular design, the VC-B criterion is (slightly) exceeded in the one-third octave band with a center frequency of $16\,\mathrm{Hz}$. The optimized design, however, does not exceed the VC-C criterion, as was imposed

by the constraints. Therefore, the optimized design is not only cost efficient when compared to rectangular wave barriers, but is also more effective in reducing the vibration levels.

4 GEOMETRIC IMPERFECTIONS

The optimized design in figure 5a may be very sensitive to deviations in the geometry. The influence of geometric deviations can be modeled by varying the projection threshold η in Eq. (5), as explained in Wang et al. [12].

Figure 6 shows the influence of the projection threshold η for the optimized design in figure 5a. For low values of the projection threshold, for example $\eta = 0.25$, the projection of the filtered densities to the physical densities also includes lower values, increasing the dimensions of the stiffer material; these designs are called dilated. For high values of the projection threshold, for example $\eta = 0.75$, the projection of the filtered densities to the physical densities only includes the higher values, decreasing the dimensions of the stiffer material; these designs are called eroded.

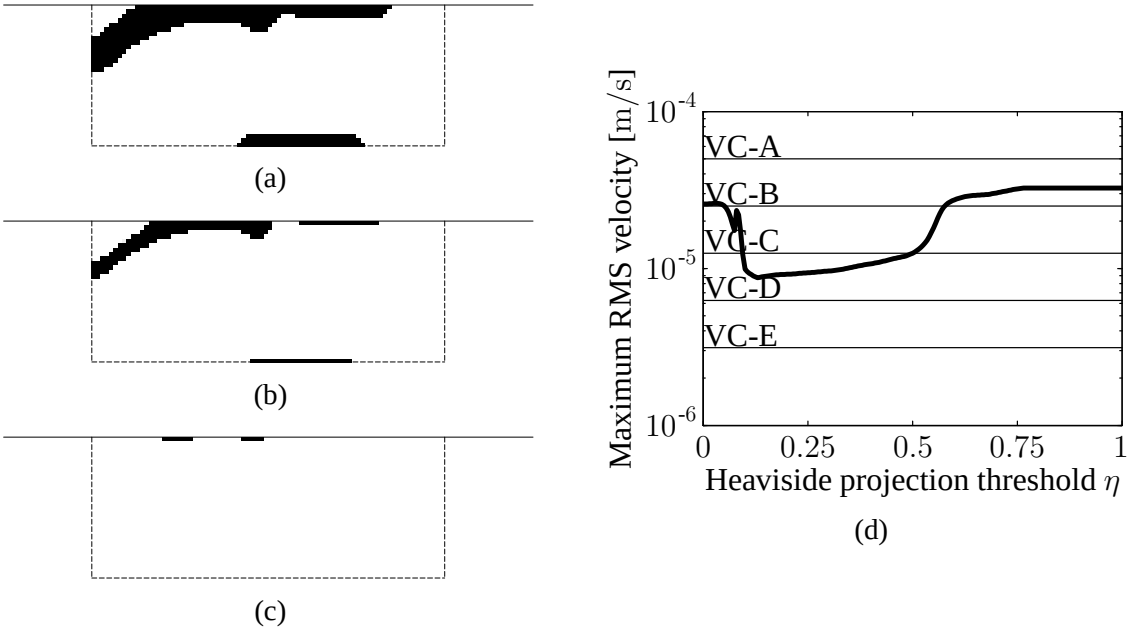


Figure 6: The (a) dilated ($\eta = 0.25$), (b) intermediate ($\eta = 0.5$), and (c) eroded ($\eta = 0.75$) version of the optimized design and (d) the influence of the projection threshold η on the maximum RMS velocity over the one-third octave bands.

Figure 6d shows that decreasing the projection threshold for example from $\eta = 0.5$ to $\eta = 0.25$ leads to a small reduction of the maximum RMS velocity over the one-third octave bands. However, increasing the projection threshold for example from $\eta = 0.5$ to $\eta = 0.75$ largely increases the maximum RMS velocity. As can be seen by comparing figures 6b and 6c, the optimized design contains only very small features which are removed when increasing the projection threshold. As a result, almost no stiffer material is left in the design domain and the performance is close to that of the homogeneous halfspace.

To obtain a design which is less sensitive to this type of geometric imperfections, a robust topology optimization approach is used. A worst case formulation is adopted where the worst performance of some (extreme) cases is optimized. Because of the smoothness of the curve

in figure 6d, good results are expected when optimizing the worst performance of only three designs with different projection thresholds ($\eta^d = 0.25$, $\eta^i = 0.5$, and $\eta^e = 0.75$). The robust optimization problem is then formulated as follows:

$$\begin{aligned} \min_{\rho_e} \quad & \sum_{e=1}^{N} v_e \bar{\rho}_e \left(\eta^i\right) \\ \text{s.t.} \quad & \hat{v}_{m,i}^{\mathrm{RMS}} \left(\bar{\rho}_e \left(\eta^q\right)\right) \leq w_m s, \quad m = 1 \cdots N_{\mathrm{band}} \\ & \qquad\qquad\qquad\qquad\quad i = 1 \cdots N_L \\ & \qquad\qquad\qquad\qquad\quad q = [d, i, e] \\ & 0 \leq \rho_e \leq 1, \qquad\qquad\quad e = 1 \cdots N_e \end{aligned} \tag{11}$$

where constraints are added for the dilated and eroded design.

The resulting dilated, intermediate and eroded design are shown in figure 7. Again, decreasing the projection threshold leads to a dilated design and increasing the projection threshold leads to an eroded design. However, as can be seen in figure 7d, the maximum RMS velocity does not exceed the VC-C criterion in the target interval $[0.25, 0.75]$. The volume, however, has increased as compared with the deterministic optimized design in figure 5a. The robust design has a total cross sectional area equal to $13.1\,\%$ of the design domain or $20.9\,\mathrm{m}^2$.

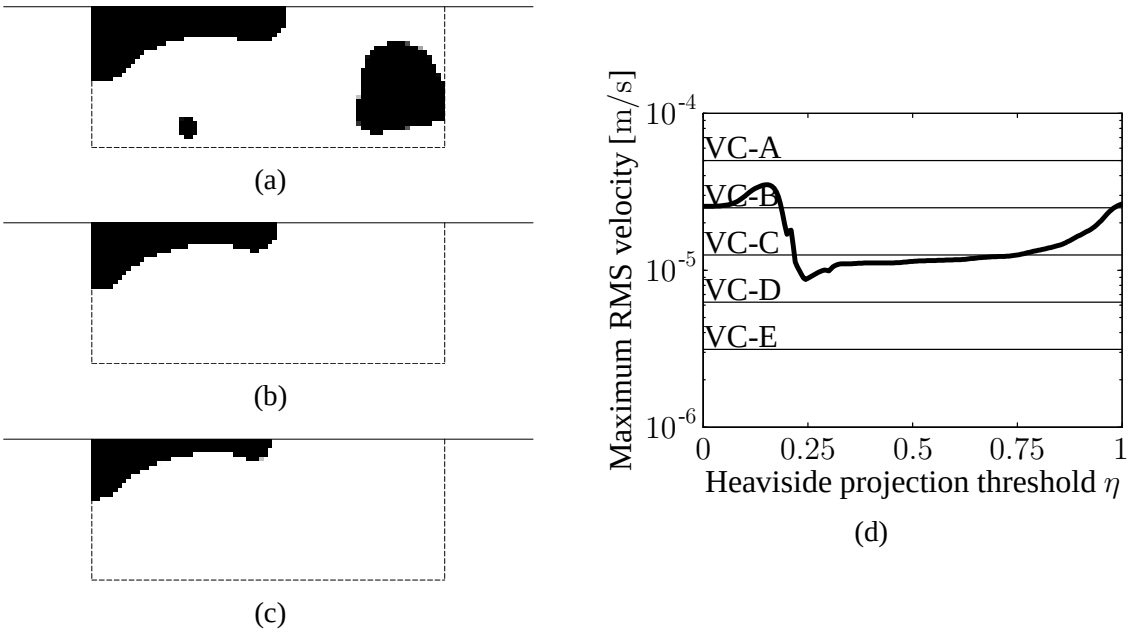


Figure 7: The (a) dilated ($\eta = 0.25$), (b) intermediate ($\eta = 0.5$), and (c) eroded ($\eta = 0.75$) version of the robust design and (d) the influence of the projection threshold η on the maximum RMS velocity over the one-third octave bands.

The performance of the intermediate robust design is compared in figure 8 with a rectangular wave barrier with a depth of $8\,\mathrm{m}$ and the same volume as the robust optimized design. While in the original halfspace, only the VC-A criterion was met, the rectangular design just meets the VC-B criterion and the optimized design meets the VC-C criterion, as was imposed by the constraints in Eq. (11).

For practical buildability considerations, it may be desirable to simplify the design in figure 8a. Since it is robust to thickness variation, simplifying the geometry is not expected to

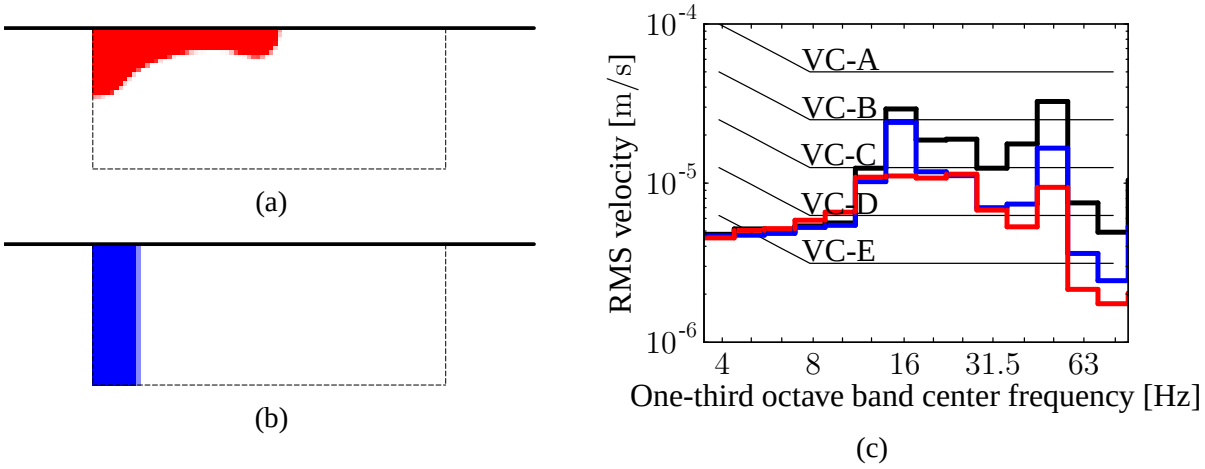


Figure 8: (a) Robust topology optimized design, (b) rectangular design with a depth of $8\,\mathrm{m}$ and the same volume as the robust design, and (c) the maximum one-third octave band RMS spectra of the velocity at the receivers for the original homogeneous halfspace (black), after the introduction of the rectangular design in (b) (blue), and after the introduction of the robust design in (a) (red).

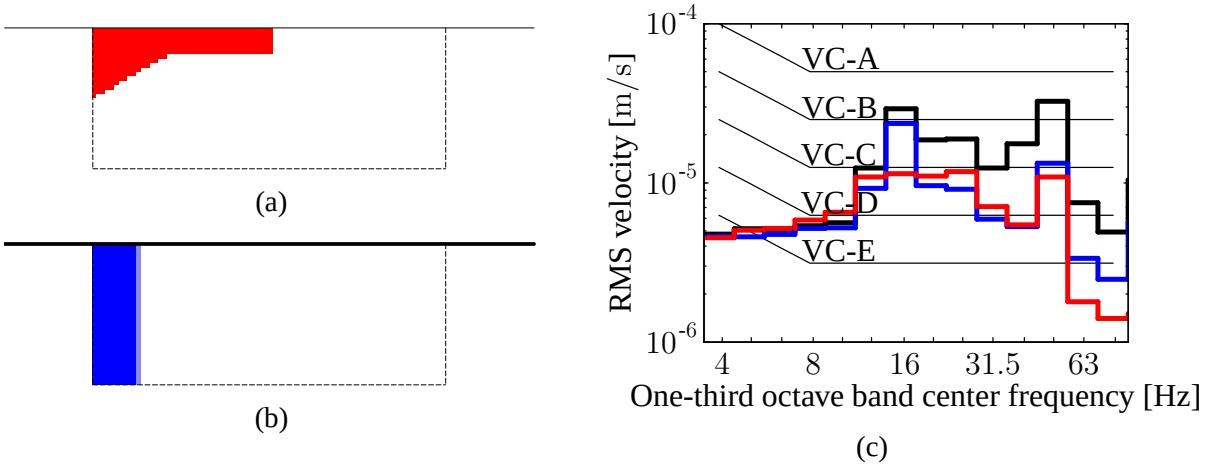


Figure 9: (a) Simplified design, (b) rectangular design with a depth of $8\,\mathrm{m}$ and the same volume as the simplified design, and (c) the maximum one-third octave band RMS spectra of the velocity at the receivers for the original homogeneous halfspace (black), after the introduction of the rectangular design in (b) (blue), and after the introduction of the simplified design in (a) (red).

significantly affect the design performance. By intuitively positioning a simplified shape with straight edges at the location where the stiffer material is concentrated in the optimized design, the design in figure 9a is obtained. This simplified design has approximately the same volume as the robust design (a cross sectional area of $20.9\,\mathrm{m}^2$) and the performance is similar, as can be seen in figure 9c. The VC-C criterion is still met.

5 CONCLUSION

In this paper, topology optimization is used to design wave barriers for the reduction of elastodynamic wave transmission. The wave barriers are located between the source, for example

a railway track, and the receivers, located in a building near the track. The volume of the wave barriers is minimized while the vibration levels have to meet stringent vibration criteria.

It is shown that rectangular wave barriers in the considered design domain are not able to meet the predefined vibration criteria. However, topology optimization leads to designs which not only meet the vibration criteria, but also need a low amount of material. The resulting design therefore outperforms the rectangular wave barrier.

Since the optimized design contains many small features, it is sensitive to geometric imperfections. A worst case approach is adopted. A larger amount of material is needed, but the resulting design is robust with respect to geometric imperfections, making it possible to simplify the topology with almost no deterioration of performance. The resulting design is buildable, cost efficient and effective in reducing vibration levels.

ACKNOWLEDGEMENTS

The first author is a doctoral fellow of the Research Foundation Flanders (FWO). The financial support is gratefully acknowledged. The authors are members of KU Leuven-BOF PFV/10/002 OPTEC-Optimization in Engineering Center.

REFERENCES

[1] G. Lombaert, G. Degrande, B. Vanhauwere, B. Vandeborght, S. François, The control of ground-borne vibrations from railway traffic by means of continuous floating slabs. *Journal of Sound and Vibration*, **297**, 946–961, 2006.

[2] L. Andersen, S.R.K. Nielsen, Reduction of ground vibration by means of barriers or soil improvement along a railway track, *Soil Dynamics and Earthquake Engineering*, **25**, 701–716, 2005.

[3] P. Coulier, S. François, G. Degrande, G. Lombaert, Subgrade stiffening next to the track as a wave impeding barrier for railway induced vibrations. *Soil Dynamics and Earthquake Engineering*, **48**, 119–131, 2013.

[4] J.P. Talbot, H.E.M. Hunt, A generic model for evaluating the performance of base-isolated buildings, *Journal of Low Frequency Noise, Vibration and Active Control*, **22**, 149–160, 2003.

[5] M.P. Bendsøe, O. Sigmund, *Topology optimization: theory, methods and applications*, Springer, Berlin, 2003.

[6] J.S. Jensen, O. Sigmund, Topology optimization of photonic crystal structures: a high-bandwidth low-loss T-junction waveguide, *Journal of the Optical Society of America B*, **22**, 1191–1198, 2005.

[7] O. Sigmund, J.S. Jensen, Systematic design of phononic band-gap materials and structures by topology optimization, *Philosophical Transactions of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences*, **361**, 1001–1019, 2003.

[8] M. Dühring, J. Jensen, O. Sigmund, Acoustic design by topology optimization, *Journal of Sound and Vibration*, **317**, 557–575, 2008.

[9] C.G. Gordon, Generic vibration criteria for vibration-sensitive equipment. *Proceedings of the SPIE Conference on Vibration Control and Metrology*, **1619**, 71–85, San Jose, California, USA, November, 1991.

[10] M. Jansen, G. Lombaert, M. Schevenels, Robust topoplogy optimization of structures with imperfect geometry based on geometric nonlinear analysis. *Computer Methods in Applied Mechanics and Engineering*, **285**, 452–467, 2015.

[11] O. Sigmund, Manufacturing tolerant topology optimization, *Acta Mechanica Sinica*, **25**, 227–239, 2009.

[12] F. Wang, B. Lazarov, O. Sigmund, On projection methods, convergence and robust formulations in topology optimization, *Structural and Multidisciplinary Optimization*, **43**, 767–784, 2011.

[13] I. Harari, U. Albocher, Studies of FE/PML for exterior problems of time-harmonic elastic waves. *Computer Methods in Applied Mechanics and Engineering*, **195**, 3854–3879, 2006.

[14] M.P. Bendsøe, Optimal shape design as a material distribution problem, *Structural optimization* **1**, 193–202, 1989.

[15] G. Rozvany, M. Zhou, T. Birker, Generalized shape optimization without homogenization, *Structural optimization* **4**, 250–252, 1992.

[16] J.K. Guest, J.H. Prévost, T. Belytschko, Achieving minimum length scale in topology optimization using nodal design variables and projection functions. *International Journal for Numerical Methods in Engineering*, **61**, 238–254, 2004.

[17] K. Svanberg, The method of moving asymptotes – a new method for structural optimization. *International Journal for Numerical Methods in Engineering*, **24**, 359–373, 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MULTIOBJECTIVE OPTIMIZATION USING GENETIC ALGORITHMS IN TIME-COST CONSTRUCTION PROJECT SCHEDULING PROBLEM

J. Magalhães-Mendes[1]

[1] Departamento de Engenharia Civil
CIDEM
ISEP - Instituto Superior de Engenharia – Instituto Politécnico do Porto
Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto
PORTUGAL
e-mail: jjm@isep.ipp.pt

Keywords: Multiobjective optimization, genetic algorithms, project management, construction management.

Abstract. *Multiobjective formulations are realistic models for many complex engineering optimization problems [1]. In a construction project, there are two main factors, such as project duration and project cost. The activity duration is a function of resources (i.e. crew size, equipments and materials) availability. On the other hand, resources demand direct costs. Therefore, the relationship between project time and direct cost of each activity is a monotonously decreasing curve. It means if activity duration is compressed then that leads to an increase in resources and so that direct costs. But, project indirect costs increase with the project duration. In general, for a project, the total cost is the sum of direct and indirect costs and exists an optimum duration for the least cost. Hence, relationship between project time and cost is trade-off [2].*

There are two general approaches to multiple-objective optimization. One is to combine the individual objective functions into a single composite function. Determination of a single objective is possible with methods such as utility theory, weighted sum method, etc., but the problem lies in the correct selection of the weights or utility functions to characterize the decision-makers preferences [1].

The main difficulty with the single composite function is selecting a weight vector for each run. To overcome this drawback a GA based-approach to solving the time-cost optimization problem has been proposed. The idea is transforming weights in the objective function by genes obtained from the genetic algorithm for each run.

The present approach provides a powerful alternative for the solution of the time-cost construction project scheduling problems when compared against others approaches including evolutionary algorithms.

1 INTRODUCTION

The time and cost are usually two objectives which are often tradeoff in project practices. When the construction time is shortened, the project cost should be added. The main objective of construction project management is to execute the project within the anticipated time while satisfying the minimum cost.

Several approaches to solve the time-cost optimization (TCO) problem have been proposed in the last years: mathematical, heuristic and search methods.

Several mathematical models such as linear programming (Hendrickson and Au [5]; Pagnoni [3]), integer programming, or dynamic programming (Robinson [8]; P. De et al. [23]) and LP/IP hybrid (Liu et al. [20]; Burns et al. [25]) and Meyer and Shaffer [27] use mixed integer programming. However, for large number of activity in network and complex problem, integer programming needs a lot of computation effort (Feng et al. [6]).

Heuristic algorithms are not considered to be in the category of optimization methods. They are algorithms developed to find an acceptable near optimum solution. Heuristic methods are usually algorithms easy to understand which can be applied to larger problems and typically provide acceptable solutions (Hegazy [26]). However, they have lack mathematical consistency and accuracy and are specific to certain instances of the problem (Fondahl [19]; Siemens [22]) are some of the research studies that have utilized heuristic methods for solving TCO problems.

Some researchers have tried to introduce evolutionary algorithms to find global optima such as genetic algorithm (GA) (Feng et al. [6]; Gen and Cheng [21]; Zheng et al. [10]; Zheng and Ng [9]; Mendes [16, 18] and Parveen and Saha [32]) the particle swarm optimization algorithm (Yang [10]), ant colony optimization (ACO) (Xiong and Kuang [28]; Ng and Zhang [24]; Afshar et al. [2]) and harmony search (HS) (Geem [29]). In this paper, the optimal time and cost generated by the GA techniques are compared with those produced by other techniques through some problems obtained from literature.

This paper is organized as follows. Section 2 describes the multiobjective optimization problem. Section 3 presents the approach. The case study and results are presented in Section 4. Finally, conclusions and future work are outlined in Section 5.

2 MULTIOBJECTIVE OPTIMIZATION

Multiobjective optimization deals with solving optimization problems which involve multiple objectives. We can say that there are two types of methods for solving problems with multi-objective optimization: the classical methods and methods based on evolutionary algorithms.

The disadvantages of the classical methods are shown in [31]:

- Only one non-dominated solution is obtained by each execution of the algorithm. It means that in order to get a set of solutions, it should be run many times.
- Some of them require some kind of information of the problem treated.
- Some of them are sensitive to the shape of the Pareto frontier, so in non-convex ones, they cannot find solutions.
- The dispersion of the founded Pareto solutions depends on the efficiency of the monocriteria optimizator.
- In problems that contain stochasticities, classical methods are not appropriate.
- Problems with discrete domain cannot be solved by classical methods, neither in the multiobjective case. Consequently, the problem treated in the present article, discrete, could not be solved by this kind of methods.

All this disadvantages are overcome with evolutionary multiobjective methods such as genetic algorithms (MOGA) [30].

With evolutionary algorithms being used for single-objective optimization for over two decades, the incorporation of more than one objective in the fitness function has finally gained popularity in the research [33].

The approach presented in this paper is based on a random key based genetic algorithm to perform its optimization process, so this approach aims to stipulate multiple search directions at each generation without using any additional parameters.

2.1 Linear Combination of Weights

The classical approach to solve a multi-objective optimization problem is to assign a weight w_i to each normalized objective function $z_i^{'}(x)$ so that the problem is converted to a single objective problem with a scalar objective function as follows [35]:

$$\min z = w_1 * z_1^{'}(x) + w_2 * z_2^{'}(x) + ... + w_n * z_n^{'}(x) \tag{1}$$

where $z_i^{'}(x)$ is the normalized objective function of $z_i(x)$ and $\Sigma w_i = 1$. The main difficulty with this approach is selecting a weight vector for each run.

2.2 Linear Combination of Genes

Linear Combination of Weights also called the weighted sum method (WSM) is the simplest approach and probably the most widely used classical method. This method scalarizes the set of objectives into a single objective by multiplying each objective with a user supplied weight.

Based on the WSM idea, this paper proposes a linear combination of genes where $z_i^{'}(x)$ is the normalized objective function of $z_i(x)$, with the following formulation:

$$\min z = gene_1 * z_1^{'}(x) + gene_2 * z_2^{'}(x) + ... + gene_n * z_n^{'}(x) \tag{2}$$

where $z_i^{'}(x)$ is the normalized objective function of $z_i(x)$ and $0 < \sum gene_i < n$, each $gene_i$ is randomly generated for individual solution $\boldsymbol{x}$ during the selection phase at each generation.

3 THE GA-BASED APPROACH

The approach combines a genetic algorithm, a schedule generation scheme and a local search procedure. The genetic algorithm is responsible for evolving the chromosomes which represent the priorities of the activities, delay times and objective function.

For each chromosome the following four phases are applied [16]:

1) *Transition parameters* - this phase is responsible for the process transition between first level and second level;
2) *Schedule parameters* - this phase is responsible for transforming the chromosome supplied by the genetic algorithm into the priorities of the activities and delay time;
3) *Schedule generation* - this phase makes use of the priorities and the delay time and constructs schedules;
4) *Schedule improvement* - this phase makes use of a local search procedure to improve the solution obtained in the schedule generation phase.

This study considers both project cost and time. For effective time-cost optimization the approach proposes an objective function with the following formulation:

$$Z_1^{'} = Z_{time}^{'} = \frac{(Z_t^{\max} - Z_t)}{(Z_t^{\max} - Z_t^{\min})} \tag{3}$$

$$Z_2^{'} = Z_{\cos t}^{'} = \frac{(Z_c^{\max} - Z_c)}{(Z_c^{\max} - Z_c^{\min})} \tag{4}$$

and finally:

$$\min f(x) = Gene_t \frac{(Z_t^{\max} - Z_t)}{(Z_t^{\max} - Z_t^{\min})} + Gene_c \frac{(Z_c^{\max} - Z_c)}{(Z_c^{\max} - Z_c^{\min})} \tag{5}$$

where,

$Z_c^{\max}$ = maximal value for total cost in the current chromosome;
$Z_t^{\max}$ = maximal value for time in the current chromosome;
$Z_c^{\min}$ = minimal value for total cost in the initial population;
$Z_t^{\min}$ = minimal value for time in the initial population;
Z_c = represents the total cost of the xth solution in current chromosome;
Z_t = represents the time of the xth solution in current chromosome.

3.1 GA-Decoding

Each chromosome represents a solution to the problem and it is encoded as a vector of random keys (random numbers). Each solution encoded as initial chromosome (first level) is made of 2+*mn*+*n* genes where *n* is the number of activities and *m* is the number of execution modes, see Figure 1 [16, 18].

Genes for Activity 1	Activity 1	Mode 1	Gene $_{11}$
		Mode 2	Gene $_{12}$
		…	…
		Mode m	Gene $_{1m}$
		Delay 1	Gene $_{1m+1}$
Genes for Activity 2	Activity 2	Mode 1	Gene $_{21}$
		Mode 2	Gene $_{22}$
		…	…
		Mode m	Gene $_{2m}$
		Delay 2	Gene $_{2m+1}$
… .	…		…
Genes for Activity n	Activity n	Mode 1	Gene $_{n1}$
		Mode 2	Gene $_{n2}$
		…	…
		Mode m	Gene $_{nm}$
		Delay n	Gene $_{nm+1}$
Genes for Objective Function	OF	Time	Gene $_{t}$
		Cost	Gene $_{c}$

Figure 1: Chromosome structure.

To decode each chromosome a schedule generation scheme (SGS) based on the idea of parameterized active schedules is applied [11, 14, 16]. This type of schedule consists of schedules in which no resource is kept idle for more than a predefined period if it could start processing some activity [16] and employs operators described in [15, 16].

3.2 Evolutionary Strategy

The GA based-approach uses an evolutionary strategy identical to the one proposed by Goldberg [7]. To breed good solutions, the population is operated by a genetic algorithm. There are many variations of genetic algorithms obtained by altering the reproduction, crossover, and mutation operators.

In this approach reproduction is accomplished by first copying some of the best individuals from one generation to the next, in what is called an elitist strategy.

The fitness proportionate selection, also known as roulette-wheel selection, is the genetic operator for selecting potentially useful solutions for reproduction. The characteristic of the roulette wheel selection is stochastic sampling.

The fitness value is used to associate a probability of selection with each individual chromosome. If f_i is the fitness of individual i in the population, its probability of being selected is,

$$p_i = \frac{f_i}{\sum_{i=1}^{N} f_i} \quad , \qquad i = 1, \ldots, n \tag{6}$$

A roulette wheel model is established to represent the survival probabilities for all the individuals in the population. Then the roulette wheel is rotated for several times.

After selecting, crossover may proceed in two steps. First, members of the newly selected (reproduced) chromosomes in the mating pool are mated at random. Second, each pair of chromosomes undergoes crossover as follows: an integer position k along the chromosome is selected uniformly at random between *1* and the chromosome length l. Two new chromosomes are created swapping all the genes between k+1 and l, see Mendes [17].

The mutation operator preserves diversification in the search. This operator is applied to each offspring in the population with a predetermined probability. We assume that the probability of the mutation in this paper is 5% [17].

4 CASE STUDY

The GA based-approach is to minimize the project overall time and cost and RKTCO is applied in the case study of a project of eighteen activities originally introduced by Feng et al. [6], see Figure 2. The activity relationship for the model project consists of 18 activities and three modes of construction for each activity and their associated time and cost are presented in [6, 16]. Indirect cost rate was $1000/day.

The Table 1 shows the results for several mathematical and evolutionary-based methods. The algorithm RKTCO obtains better solution than the other GA-based approaches. Furthermore, the algorithm RKTCO reaches the optimal solution quickly, i.e., in five seconds.

The results of the RKTCO illustrates that evolutionary methods based on genetic algorithms can obtain the better solutions and in very reasonable computational time.

Based on Feng et al. [6] five new problems are generated: three experiments using the parallel expansion of the 18-activity problem and two using the serial expansion. The parallel and serial problems consisting of 36, 54 and 180 activities are described in Golzarpoor [4].

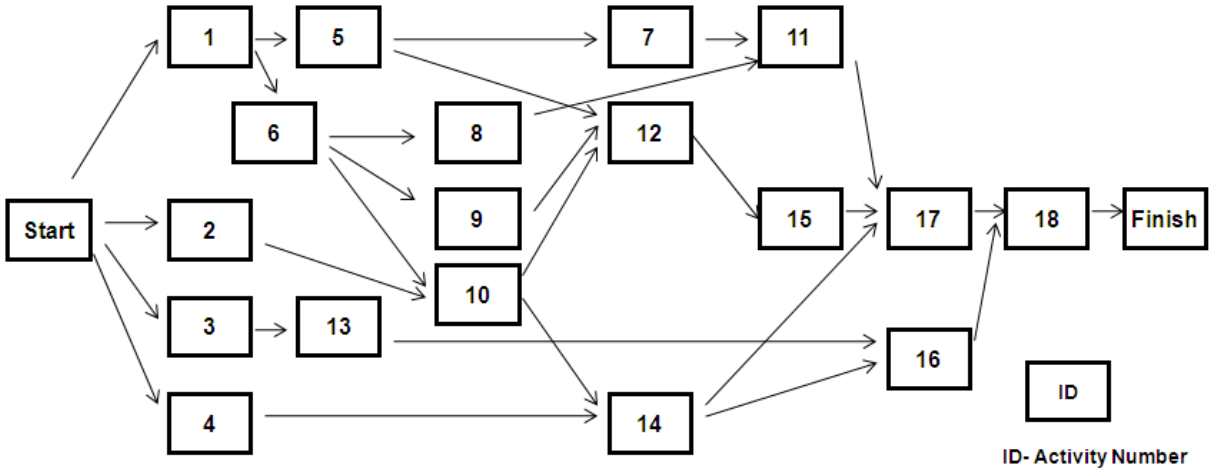


Figure 2: Project network of original problem with eighteen activities [6].

Approaches	Description	Deviation**	Criteria		Calculation Time
			Time	Cost ($)	
Optimal Solution	-	0%	110	216,270	-
Excel Solver*	Easy-to-use mathematical optimization tool	18%	110	254,620	2 min.
Risk Solver Platform Standard SLGRG Nonlinear*	Risk analysis, simulation, and optimization tools	0%	110	216,270	1.5 min.
Risk Solver Platform Standard Large-scale GRG Solver*	Risk analysis, simulation, and optimization tools	0%	110	216,270	1.5 min.
TCT Optimization Using Evolver (includes an evolutionary engine)*	GA-based optimization tool effective in optimizing complex and large-scale models	10%	110	238,070	30 min.
Risk Solver Platform Standard Evolutionary Solver*	Risk analysis, simulation, and optimization tools	27%	110	275,320	18 min.
Optimization Results using CPLEX CP Optimizer*	A solver for optimization based on constraint programming technique	0%	110	216,270	9 min.
Constraint Programming method using IBM ILOG Optimization Studio*	Constraint Programming method	0%	110	216,270	9 min.
This paper (RKTCO)	GA-based	0%	110	216,270	5 sec. for 50 generations

* Reported by Golzarpoor [4] **Percentage of deviation of the result from optimal solution

Table 1: Results for Project of 18 activities.

Table 2 shows the results for Project of 36 activities (parallel expansion). The Table 2 shows that the results of the RKTCO are the best when compared with the other methods reported by Golzarpoor [4].

Approaches	Description	Deviation**	Criteria		Calculation Time
			Time	Cost ($)	
Optimal Solution	-	0%	110	322,540	-
Excel Solver*	Easy-to-use mathematical optimization tool	27%	110	409,840	5 min
Risk Solver Platform Standard SLGRG Nonlinear*	Risk analysis, simulation, and optimization tools	20%	110	385,890	14 min
Risk Solver Platform Standard Large-scale GRG Solver*	Risk analysis, simulation, and optimization tools	18%	110	379,990	18 min
TCT Optimization Using Evolver (includes an evolutionary engine)*	GA-based optimization tool effective in optimizing complex and large-scale models	16%	109	373,790	30 min
Risk Solver Platform Standard Evolutionary Solver*	Risk analysis, simulation, and optimization tools	36%	110	438,440	15 min
Optimization Results using CPLEX CP Optimizer*	A solver for optimization based on constraint programming technique	0%	110	322,540	10 min
Constraint Programming method using IBM ILOG Optimization Studio*	Constraint Programming method	0%	110	322,540	10 min
This paper (RKTCO)	GA-based	0%	110	322,540	12 sec. 50generations

* Reported by Golzarpoor [3] **Percentage of deviation of the result from optimal solution

Table 2: Results for Project of 36 activities (parallel expansion).

Table 3 shows the results for Project of 36 activities (serial expansion). The RKTCO presents the best results when compared with the other methods reported by Golzarpoor [4].

Approaches	Description	Deviation**	Criteria		Calculation Time
			Time	Cost ($)	
Optimal Solution	-	0%	220	432,540	-
Excel Solver*	Easy-to-use mathematical optimization tool	0%	220	433,094	30 min
Risk Solver Platform Standard SLGRG Nonlinear*	Risk analysis, simulation, and optimization tools	11%	220	478,549	30 min
Risk Solver Platform Standard Large-scale GRG Solver*	Risk analysis, simulation, and optimization tools	19%	220	513,369	30 min
TCT Optimization Using Evolver (includes an evolutionary engine)*	GA-based optimization tool effective in optimizing complex and large-scale models	12%	220	484,640	30 min
Risk Solver Platform Standard Evolutionary Solver*	Risk analysis, simulation, and optimization tools	31%	219	566,640	1 min
Optimization Results using CPLEX CP Optimizer*	A solver for optimization based on constraint programming technique	0%	220	432,540	10 min
Constraint Programming method using IBM ILOG Optimization Studio*	Constraint Programming method	0%	220	432,540	10 min
This paper (RKTCO)	GA-based	0%	220	432,540	12 sec. 50generations

* Reported by Golzarpoor [4] **Percentage of deviation of the result from optimal solution

Table 3: Results for Project of 36 activities (serial expansion).

Table 4 shows the results for Project of 54 activities (parallel expansion) and the results of the RKTCO are the best when compared with the other methods reported by Golzarpoor [4].

Approaches	Description	Deviation**	Criteria		Calculation Time
			Time	Cost ($)	
Optimal Solution	-	0%	110	428,810	-
Excel Solver*	Easy-to-use mathematical optimization tool	42%	110	607,435	30 min
Risk Solver Platform Standard SLGRG Nonlinear*	Risk analysis, simulation, and optimization tools	6%	110	454,198	26 min
Risk Solver Platform Standard Large-scale GRG Solver*	Risk analysis, simulation, and optimization tools	6%	110	454,198	26 min
TCT Optimization Using Evolver (includes evolutionary engine)*	GA-based optimization tool effective in optimizing complex and large-scale models	17%	110	500,610	30 min
Risk Solver Platform Standard Evolutionary Solver*	Risk analysis, simulation, and optimization tools	44%	110	618,260	1 min
Optimization Results using CPLEX CP Optimizer*	A solver for optimization based on constraint programming technique	0%	110	428,810	10 min
Constraint Programming method using IBM ILOG Optimization Studio*	Constraint Programming method	0%	110	428,810	10 min
This paper (RKTCO)	GA-based	0%	110	428,810	14 sec. 50generations

* Reported by Golzarpoor [4] **Percentage of deviation of the result from optimal solution

Table 4: Results for Project of 54 activities (parallel expansion).

Table 5 shows the results for Project of 54 activities (serial expansion). The RKTCO presents the best results when compared with the other methods reported by Golzarpoor [4].

Approaches	Description	Deviation**	Criteria		Calculation Time
			Time	Cost ($)	
Optimal Solution	-	0%	330	648,810	-
Excel Solver*	Easy-to-use mathematical optimization tool	19%	330	769,460	13 min
Risk Solver Platform Standard SLGRG Nonlinear*	Risk analysis, simulation, and optimization tools	8%	330	698,851	30 min
Risk Solver Platform Standard Large-scale GRG Solver*	Risk analysis, simulation, and optimization tools	10%	330	714,551	30 min
TCT Optimization Using Evolver (includes an evolutionary engine)*	GA-based optimization tool effective in optimizing complex and large-scale models	8%	330	700,410	30 min
Risk Solver Platform Standard Evolutionary Solver*	Risk analysis, simulation, and optimization tools	57%	330	1.018,260	1 min
Optimization Results using CPLEX CP Optimizer*	A solver for optimization based on constraint programming technique	0%	330	648,810	10 min
Constraint Programming method using IBM ILOG Optimization Studio*	Constraint Programming method	0%	330	648,810	10 min
This paper (RKTCO)	GA-based	0%	330	648,810	15 sec. 50generations

* Reported by Golzarpoor [4] **Percentage of deviation of the result from optimal solution

Table 5: Results for Project of 54 activities (serial expansion).

Table 6 shows the results for Project of 180 activities (parallel expansion). The RKTCO presents the best results when compared with the other methods reported by Golzarpoor [4].

Approaches	Description	Deviation**	Criteria		Calculation Time
			Time	Cost ($)	
Optimal Solution	-	0%	110	1.172,700	-
Excel Solver*	Easy-to-use mathematical optimization tool	N/A	Too many variable cells		
Risk Solver Platform Standard SLGRG Nonlinear*	Risk analysis, simulation, and optimization tools	14%	110	1.336,900	30 min.
Risk Solver Platform Standard Large-scale GRG Solver*	Risk analysis, simulation, and optimization tools	35%	110	1.583,095	30 min.
TCT Optimization Using Evolver (includes an evolutionary engine)*	GA-based optimization tool effective in optimizing complex and large-scale models	54%	109	1.801,700	21 min.
Risk Solver Platform Standard Evolutionary Solver*	Risk analysis, simulation, and optimization tools	54%	110	1.807,000	21 min.
Optimization Results using CPLEX CP Optimizer*	A solver for optimization based on constraint programming technique	0%	110	1.172,700	13 min
Constraint Programming method using IBM ILOG Optimization Studio*	Constraint Programming method	0%	110	1.172,700	13 min
This paper (RKTCO)	GA-based	0%	110	1.172,700	180 sec. for 1000 generations

* Reported by Golzarpoor [4] **Percentage of deviation of the result from optimal solution

Table 6: Results for Project of 180 activities (parallel expansion).

From the above results it is clear that no algorithm dominates RKTCO. The CPLEX CP optimizer and Constraint Programming method using IBM ILOG Optimization Studio seems to have similar performance, but with higher computational time than RKTCO.

The proposed RKTVO has the best performance between all evolutionary–based algorithms (EA).

The time necessary by RKTCO to obtain the optimal solution is highly promising and shows that a good implementation can be critical to the success of the genetic algorithms.

This computational experience has been performed on a computer with an Intel Core 2 Duo CPU T7250 @2.33 GHz and 1,95 GB of RAM. The algorithm proposed in this work has been coded in VBA under Microsoft Windows NT.

5 CONCLUSIONS AND FURTHER RESEARCH

A new GA based-approach to solving the time-cost optimization for construction projects has been proposed. The project activities have various construction modes, which reflect different ways of performing the activity, each mode having a different impact on the duration and cost of the project. The present approach provides an attractive alternative for the solution of the construction multi-objective optimization problems. The developments made in this paper provide guidelines for designing and implementing practical GA applications in the civil engineering domain and gives to construction managers a tool to balance critical construction resources in the competitive construction industry.

Further research can be extended to others multiobjective problems with relationships between cost, time and quality in the construction industry.

ACKNOWLEDGMENTS

This work has been partially supported by the CIDEM (Centre for Research & Development in Mechanical Engineering). CIDEM is an investigation unit of FCT – Portuguese Foundation for the Science and Technology located at the School of Engineering of Polytechnic of Porto.

REFERENCES

[1] A. David, W. Coit, A. E. Smith, Multi-objective optimization using genetic algorithms: A tutorial, *Reliability Engineering & System Safety*, 91(9), 992–1007, 2006.

[2] A.Afshar, A. Ziaraty, A. Kaveh, F. Sharifi, Nondominated Archiving Multicolony Ant Algorithm in Time–Cost Trade-Off Optimization, *J. Constr. Eng. Manage*., 135(7), 668-674, 2009.

[3] A. Pagnoni, *Project engineering: computer-oriented planning and operational decision making*. Springer-Verlag, 1990.

[4] B. Golzarpoor, *Time-Cost Optimization of Large-Scale Construction Projects Using Constraint Programming*, Master Thesis, University of Waterloo, Ontario, Canada, 2012.

[5] C. Hendrickson, T. Au, *Project management for construction: fundamental concepts for owners, engineers, architects, and builders*, Prentice-Hall International Series in Civil Engineering and Engineering Mechanics, 1989.

[6] C.-W. Feng , L. Liu, S.A. Burns, Using genetic algorithms to solve construction time-cost trade-off problems, *J. Comp. Civ.Engrg.,* ASCE, 11(3), 184-189, 1997.

[7] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley: Reading, MA, 1989.

[8] D. R. Robinson, A dynamic programming solution to cost-time tradeoff for CPM, *Management Science*, 22(2), 158-166, 1975.

[9] D. X. M. Zheng, S. T. Ng, Stochastic time–cost optimization model incorporating fuzzy sets theory and nonreplaceable front, *Journal of Construction Engineering and Management*, ASCE, 131(2), 176-186, 2005.

[10] D. X.M. Zheng, S.T. Ng, M.M. Kumaraswamy, Applying a genetic algorithm-based multiobjective approach for time-cost optimization, *Journal of Construction Engineering and Management*, 130(2), 168-176, 2004.

[11] J.F. Gonçalves, J.J. M. Mendes, M.C.G. Resende. A hybrid genetic algorithm for the job shop scheduling problem. *European Journal of Operational Research*, 167, 77-95, 2005.

[12] J. Magalhães-Mendes, Project scheduling under multiple resources constraints using a genetic algorithm, *WSEAS Transactions on Business and Economics*, World Scientific and Engineering Academy and Society, USA, 11, 487-496, 2008.

[13] J. Magalhães-Mendes, A two-level genetic algorithm for the multi-mode resource-constrained project scheduling problem, *International Journal of Systems Applications, Engineering & Development*, 5(3), 271-278, 2011.

[14] J. Magalhães-Mendes, Active, Parameterized Active, and Non-Delay Schedules for Project Scheduling with Multi-Modes, *Proceedings of the 16th WSEAS International Conference on Applied Mathematics*, Montreaux, Switzerland, December 29-31, 134-139, 2011.

[15] J. Magalhães-Mendes, A hybrid genetic algorithm for the multi-mode resource-constrained project scheduling, *Proceedings of the 6th European Congress on Computational Methods in Applied Sciences and Engineering* (ECCOMAS 2012), Vienna University of Technology, Austria, e-Book Full Papers, 8871-8882, 2012.

[16] J. Magalhães-Mendes, *Multiobjective Approach for Time-Cost Optimization using a Multi-Mode Hybrid Genetic Algorithm*, In J. Magalhaes-Mendes and D. Greiner (Editors), Evolutionary Algorithms and Metaheuristics in Civil Engineering and Construction Management, Springer International Publishing, 39, 33-49, 2015.

[17] J. Magalhães-Mendes, A Comparative Study of Crossover Operators for Genetic Algorithms to Solve the Job Shop Scheduling Problem, *WSEAS Transactions on Computers*, World Scientific and Engineering Academy and Society, USA, 12(4), 164-173, 2013.

[18] J. Magalhães-Mendes, Time-Cost Optimization of Construction Projects: A Genetic Algorithm-Based Approach, in Y. Tsompanakis, J. Kruis, B.H.V. Topping, (Editors), *Proceedings of the Fourth International Conference on Soft Computing Technology in Civil, Structural and Environmental Engineering*, Civil-Comp Press, Stirlingshire, UK, Paper 11, 2015. doi:10.4203/ccp.109.11

[19] J. W. Fondahl, *A non-computer approach to the critical path method for the construction industry*, Technical Report No. 9, The Construction Institute, Department of Civil Engineering, Stamford University., 1961.

[20] L. Liu, S. Burns, C. Feng, Construction time-cost trade-off analysis using LP/IP hybrid method, *J. Constr. Eng. Manage.,* 121(4), 446-454, 1995.

[21] M. Gen, R. Cheng, *Genetic algorithms & engineering optimization*, Wiley-Interscience, New York, 2000.

[22] N. Siemens, A simple CPM time-cost trade-off algorithm, *Management Science*, 17(6), B354-B363, 1971.

[23] P. De, E.J. Dunne, J.B. Ghosh, C. E. Wells, The discrete time-cost trade-off problem revisited, *European Journal of Operational Research*, 81(2), 225-238, 1995.

[24] S. T. Ng, Y. Zhang, Optimizing construction time and cost using ant colony optimization approach, *Journal of Construction Engineering and Management*, ASCE, 134(9), 721-728, 2008.

[25] S.A. Burns, L. Liu, C. W. Feng, The LP/IP hybrid method for construction time-cost trade-off analysis, *Constr. Manag. Econ.,* 14(3), 265-276, 1996.

[26] T. Hegazy, Optimization of construction time-cost trade-off analysis using genetic algorithms, *Canadian Journal of Civil Engineering*, 26(6), 685-697, 1999.

[27] W. L. Mayer, L.R. Shaffer, Extension of the critical path method through the application of integer programming, *Civ. Engrg. Constr. Res. Ser. 2*, Univ. of Illinois, Urbana, III, 1963.

[28] Y. Xiong, Y. Kuang, Applying an ant colony optimization algorithm-based multiobjective approach for time–cost trade-off, *Journal of Construction Engineering and Management*, ASCE, 134(2), 153-156, 2008.

[29] Z. W. Geem, Multiobjective optimization of time-cost trade-off using harmony search, *Journal of Construction Engineering and Management*, ASCE, 136(6), 711-716, 2010.

[30] D. Greiner, J.M. Emperador, G. Winter, Multiobjective Optimisation of Bar Structures by Pareto GA, *Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering* (ECCOMAS 2000), CIMNE, Barcelona, Spain, 2000.

[31] K. Deb, *Evolutionary Algorithms for Multicriterion Optimization in Engineering Design*. Genetic Algorithms and Evolution Strategies in Engineering and Computer Science, J. Wiley & Sons, 1999.

[32] S. Parveen, S. K. Saha, GA Based Multi-Objective Time-Cost Optimization in a Project with Resources Consideration, *International Journal of Modern Engineering Research* (IJMER), 2(6), 4352-4359, 2012.

[33] C. C. Coello, An updated survey of GA-based multiobjective optimization techniques, *ACM Comput. Surv*., 32(2), 109–143, 2000.

[34] J.H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

APPLICATION OF GENETIC ALGORITHMS TO STRUTTED SHEET PILE WALL DESIGN OPTIMIZATION

Mohamed A. Eid[1], **and Remon Isaac**[2]

[1] Assistant Professor – Faculty of Engineering – Minia University, Egypt
M.Eid@mu.edu.eg

[2] Assistant Professor – Faculty of Engineering – Minia University, Egypt
RemonIsaac@gmail.com

Keywords: Strutted Sheet Pile Wall, Optimization, Genetic Algorithms, Finite Element Analysis, Deep Excavation.

Abstract. *Strutted sheet pile wall represents an economic commonly used solution for vertical or near vertical deep excavation, when open cuts with side slopes are not applicable. It is mainly used to minimize/ control ground deformations to ensure safety of adjacent structures. Thus, the strutted sheet pile wall design process should present elements with sufficient strength/ stiffness to resist the lateral earth pressure without losing the economic advantage. Although the choice of strut and wall configuration is very precise and sensitive process to optimize the design, it still depends on empirical formulas and designers' experience. Application of optimization to this process can bring out some scientific based design rules.*

In this research, a heuristic optimization technique, Genetic Algorithms, is applied to the strutted sheet pile wall design. The optimization process aims to minimize the construction cost expressed with elements dimensions considering both deformation and stress constraints for the ground soil and construction material. The Genetic Algorithms technique is combined with Finite Element Analysis to find the optimal values for the design variables (strut cross section and position and sheet pile wall thickness and embedded depth) for different soft soil types. Results can be used to draw a map for the design process.

1 INTRODUCTION

Development projects in urban congested areas such as transportation tunnels, underground parking garages, basements and utilities…etc. necessitate the use of the underground space and hence the use of side support systems to retain their deep excavations.

Generally, excavation support systems for deep excavations consist of two main components: a wall, and its supporting measures. Walls supporting deep excavations may be classified into the following three major categories according to the form of supporting measures provided for them: (1) Cantilevered wall (usually for relatively shallow excavation); (2) Strutted/braced wall; and (3) Tied-back or anchored wall.

Many factors affect the excavation-induced deformations such as: wall stiffness, stiffness of supporting measures, ground conditions, groundwater condition and control measures, excavation depth, construction sequences and workmanship. Strutted sheet pile wall represents an economic commonly used solution for vertical or near vertical deep excavation as it minimizes ground deformations to ensure safety of adjacent structures. Design optimization of strutted wall systems will have a significant impact on cost especially for infrastructure projects with deep excavation.

In 1981, Mana & Clough made few trials on the impact of one effective design parameter in an attempt to reach a better design for the strutted sheet pile wall [1]. They showed that increasing the stiffness of the strut/anchor/raker reduces the deformation 40% as shown in Figure 1.

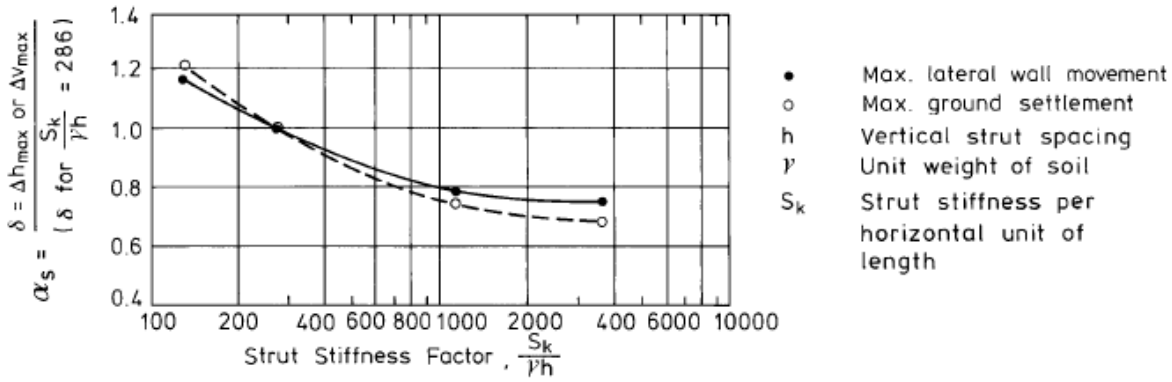


Figure 1. The effect of the strut stiffness on the maximum lateral deformation of the wall and the maximum settlement [1].

Recently, in 2012, Chowdhury, Deb, and Sengupta carried out a parametric study (16 design cases of diaphragm wall systems with different strut arrangements) to investigate the influence of different design parameters, such as strut stiffness, wall thickness, strut arrangement and the embedded depth of the wall on strut force, maximum moment developed in the wall, maximum lateral displacement of the wall, and maximum vertical displacement of ground surface. The parametric study was used to derive a design guideline to optimize the side support system. It was observed that, for a particular wall thickness and strut stiffness, different strut arrangements produced different results for maximum strut force, maximum moment, maximum horizontal wall displacement, and maximum vertical ground surface displacement [2].

A real optimization process cannot be performed for such problem through deterministic search methods. The calculus based analytical methods depend on the existence of mathematical formula and continues functions to express the studied problems, which are not available in this case. Application of enumerative search method shall consume a very long time to

reach the optimal solution, regarding the large search space [3]. Thus, stochastic search methods shall be more efficient in such optimization problem. This research aims to demonstrate how a heuristic optimization technique, Genetic Algorithms, can be applied to reach an optimal or a quasi optimal solution with reasonable computation cost.

2 GENETIC ALGORITHMS

Genetic Algorithms (GA_s) are search and optimization procedures that are motivated by the principles of natural genetics and natural selection. They are also referred to as stochastic optimization techniques different from usual mathematical programming [4]. Being considered as stochastic methods, the GA_s do not need specific information to guide the search, and require only an evaluation of the objective function value for each decision variable set in order to proceed. They typically work with a coding of the decision variables, not with the decision variables themselves. They search simultaneously using a population of decision variable sets, not a single set of decision variables [5].

As in a biological system submitted to external constraints, the fittest members of the population are selected to survive and given better chances of reproducing and transmitting part of their genetic heritage to the next generation. A new population is then created by recombination of parental genes. It is expected that some members of this new population will have acquired the best characteristics of both parents and, being better adapted to the environmental conditions, will provide an improved solution to the considered problem. After it has replaced the original population, the new group is submitted to the same evaluation procedure, and later generates its own offsprings. The process is repeated many times, until elite members of a given generation share the same genetic heritage. These members, who are often quite different from their ancestors, possess genetic information that corresponds to the best solution to the optimization problem [6].

3 OPTIMIZATION PROBLEM DESCRIPTION

The optimization problem addressed herein is to find out the optimal wall embedded depth, sheet wall section and struts number, section and positions those lead to pre-specified safe lateral deformation during excavation and acceptable induced stresses in soil and structural elements to have minimum system cost. Figure 2 shows a half section for the structural system and geometrical variables considered.

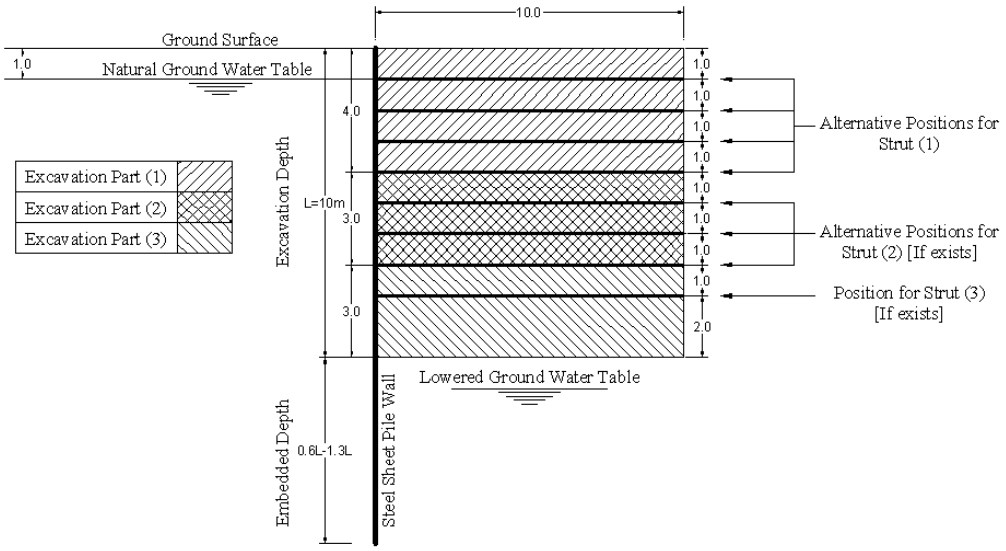


Figure 2: Problem layout and geometrical variables.

Hence, the objective function can be stated as:

$$\text{find } X \in R_k \text{ to}$$
$$\text{minimize } f(X)$$
$$\text{subject to } g_i(X) \leq 0,\ i = 1, 2, \ldots, n \text{ and}$$
$$X_j^L \leq X \leq X_j^U, \qquad j = 1, 2, \ldots\ldots, k$$

where X is the vector of design variables; f (X) is the objective function; $g_i(X)$ is the performance constraints; and X_j^L and X_j^U refer to the lower and upper bounds on the design variables respectively. The objective function here is the weight of the system and can be expressed as:

$$\min W = \gamma_s (L_w S_w + L_s S_s N_s) \tag{1}$$

where γ_s is steel density, L_w is the sheet wall total length, S_w is the sheet wall cross-sectional area, L_s is the strut length, S_s is the strut cross-sectional area and N_s is the struts number.

3.1 Design Variables

Six design variables are considered in this optimization problem:

1) Sheet pile wall section: the section is selected from 8 different alternatives shown in Table 1 [7].
2) Sheet pile wall embedded depth: a range from 0.6 L to 1.3 L, where L is the excavation depth, with 0.1 L step is examined.
3) Strut section: 8 different pipes from Egyptian standard steel sections, shown in Table 2, are selected.
4) Upper strut position: 4 different alternative positions for upper strut are considered; at 1, 2, 3 or 4 m depth from ground surface.
5) Middle strut existence/position: this one has an optional existence with 3 alternative positions; at 5, 6 or 7 m from ground surface.
6) Lower strut existence: it has an optional existence at a certain position; 8 m from ground surface. So, it is a yes/no variable.

Section No.	Section Name	Width (mm)	Height (mm)	Back thickness (mm)	Web thickness (mm)	Inertia (cm^4/m)
1	Larssen 600	600	150	10.0	9.9	4050
2	Larssen 601	600	310	8.0	6.8	12245
3	Larssen 602	600	310	8.7	8.4	13640
4	Larssen 603	600	310	10.2	8.5	19375
5	Larssen 604	600	380	10.5	9.2	31675
6	Larssen 605	600	420	13.0	9.2	43890
7	Larssen 606	600	435	14.9	9.4	55900
8	Larssen 607	600	452	19.5	10.8	73900

Table 1: Alternative sections for sheet pile wall.

Section No.	Pipe No.	Diameter (mm)	Thickness (mm)	Area (cm^2)	Radius of gyration (cm)
1	325	325	8	79.7	11.2
2	325	325	10	99.0	11.1
3	368	368	8	90.5	12.7
4	368	368	10	112.0	12.7
5	419	419	10	128.0	14.5
6	419	419	12	153.0	14.4
7	529	529	9	147.0	18.4
8	529	529	10	163.0	18.4

Table 2: Alternative sections for struts pipes.

3.2 Constraints

The performance constraints include concern for two types of safety conditions:

1) Stress condition: the induced stresses in soil, sheet wall and struts, including buckling effect in struts, should be within allowable limits.

$$g_{1i} = |\sigma_i| - \sigma_{all} \leq 0, \qquad i = 1, 2, ...m \tag{2}$$

2) Deformation condition: lateral displacements in sheet wall during excavation should not exceed the pre-specified limit (0.005 L = 5 cm).

$$g_{2i} = |\Delta_i| - \Delta_{all} \leq 0, \qquad i = 1, 2, ...m \tag{3}$$

4 NUMERICAL MODELING

In this paper, conventional numerical model with plane-strain analysis is used. Figure 3 shows the finite elements mesh used for simulation with close views for excavation area before and after excavation.

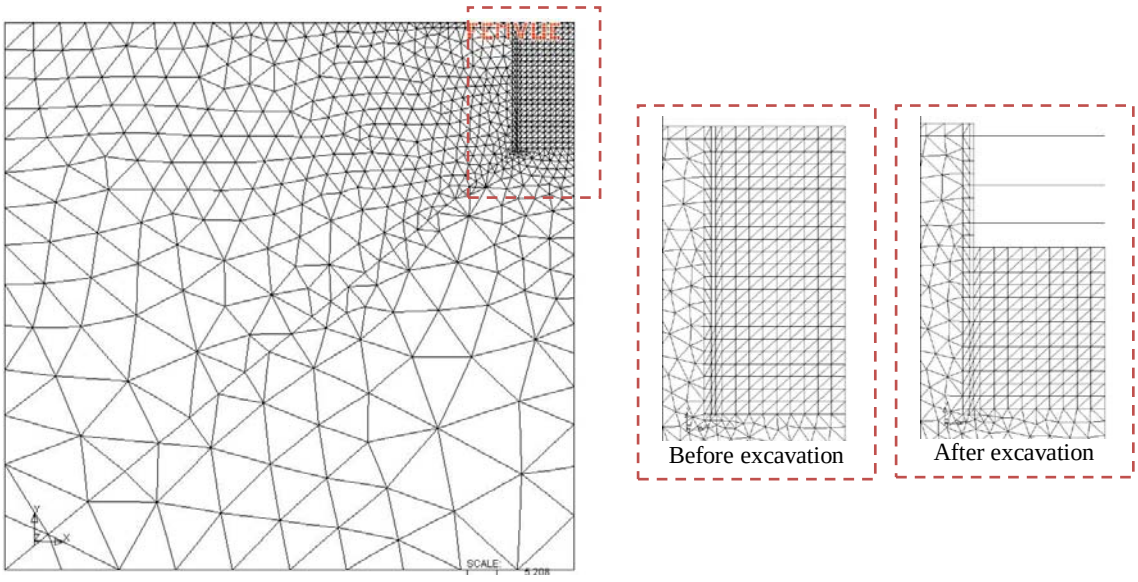


Figure 3: Finite elements simulation mesh.

For the modeling process in the Finite Element program, FINAL package [8], the soil media is modeled using six-node linearly varying strain triangular finite elements (L.S.T), the sheet pile wall is presented by six-node beam elements (Beam6) and the strut is modeled by two-node link member (Beam2). Ground water is presented with its net lateral load. A half-section mesh is used in the analysis to reduce computation time. Sufficient mesh depth and width, to model soil infinite body, are used. For boundary condition, vertical and horizontal movements are prevented at the bottom of the model while only the horizontal movements are prevented at both sides.

4.1 Geotechnical Parameters

Two different soil types are examined, in two different models, in order to check the impact of soil parameter on optimal values. Characteristic and mechanical properties for these soil types are given in Tables 3 and 4.

Density	19 kN/m^3
Elastic Modulus	20 MPa
Poisson Ratio	0.3
Angle of Internal Friction	35 ^{o}
Cohesive Strength	0 kPa

Table 3: Mechanical properties for sandy soil.

Density	20 kN/m^3
Elastic Modulus	15 MPa
Poisson Ratio	0.45
Angle of Internal Friction	0 ^{o}
Cohesive Strength	50 kPa

Table 4: Mechanical properties for clayey soil.

5 COMPUTATION PROCEDURE

5.1 Enumerative Search Process

The main target, for current research, is to check the feasibility and efficiency of GA_s application in the studied geotechnical problem. So, an enumerative search process was performed to investigate all possible solutions to find the global optimal solution and the associated design parameters. For each soil type, the search space was divided to four equal parts. Four moderate speed PCs run continuously for about 48 hours to finish this automated search process. Tables 5 and 6 summarize the search results.

	Sandy Soil	Clayey Soil
Number of safe solutions	10629	11206
Number of unsafe solutions	5755	5178
Maximum safe weight (ton/m)	6.472	6.472
Minimum safe weight [optimal value] (ton/m)	1.872	1.759

Table 5: Enumerative search general statistics.

	Sandy Soil	Clayey Soil
Sheet pile wall section	Larssen 602	Larssen 601
Sheet pile wall embedded depth	0.6 L = 6 m	0.6 L = 6 m
Strut section	Pipe 325 (section 1)	Pipe 325 (section 1)
Upper strut position	4 m from ground surface	4 m from ground surface
Middle strut existence/position	None	None
Lower strut existence	None	None

Table 6: Global optimal solution.

5.2 Optimization Process

According to the assumed alternatives for every variable, the total number of possible solutions (chromosomes) = 8*8*8*4*4*2 = 16384, for each soil. An initial population of 8 chromosomes is considered. Each individual is sent to the FINAL package to check the stresses in structural elements and lateral displacement and compare them with the allowable values. Unsafe solutions get penalty function; their evaluation value (system total weight) is multiplied by 10. Then, all solutions are encoded to binary form to facilitate the application of mating operators. Genetic Algorithm's mating operators are crossover and mutation. Each two solutions are mated together to produce two children solutions. Like their parents, produced solutions have to be checked. Unsafe produced solutions, also, get the penalty function. Both parents and children solutions are collected in one pool and sorted in an ascending order. The last 8 solutions are discarded and the first 8 solutions form the parents' population for the next generation. Processing optimization operators and repeating them through generated population leads to convergence toward global optimum.

The termination criteria is defined with limited number of generations; 200 generations. During these generations production, the GA_s computer program explores about 1600 solution which is less than 10% of the search space. Difficulty of having optimal or quasi optimal solution increases as convergence rate increases. Thus, the mutation probability is increased from 20% in the first 100 generations to 40% in the second 100 generations. Figure 4 shows the structural overall weight progression through generations for different analysis times.

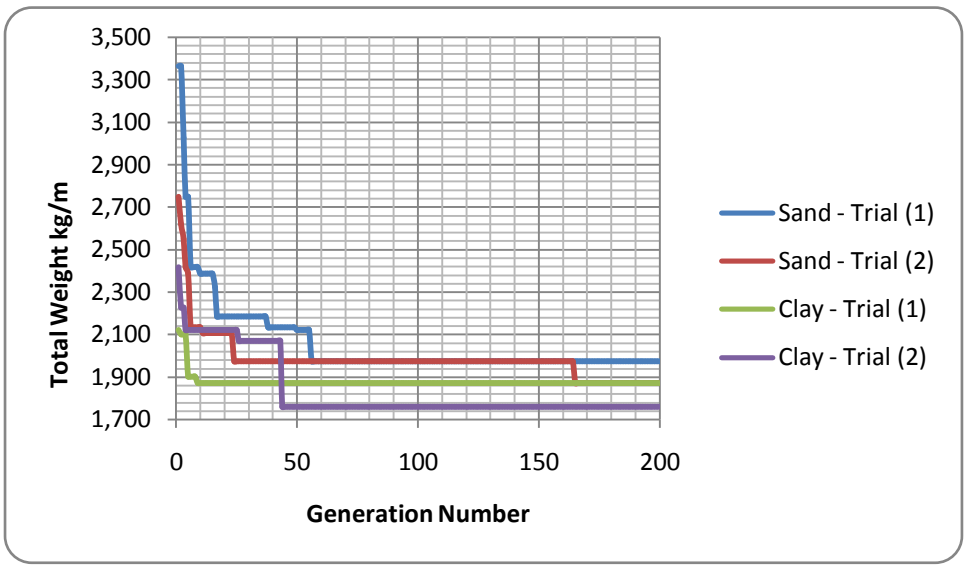


Figure 4: Overall system weight progression through generations.

In different analysis times, the program managed to find the global optimal or a quasi optimal solution (less than 7% higher than optimal system weight) before termination. The analysis consumes 18 hours on a moderate computer specifications (Core2 Duo processor, 2.0 GHz speed and 2GB RAM).

6 CONCLUSIONS

- GA_s have proven to be successful in the presented optimization problem.
- Despite the large search space, the program managed to reach the optimal or near optimal solution in reasonable time and iterations number.
- Increasing the mutation probability did not help enhancing the progression.
- Repeating the application for different soil and geometric conditions can draw a map for this design process.

REFERENCES

[1] A. I. Mana, G. W. Clough, G.W. 1981. Prediction of movements for braced cuts in clay. *Journal of Geotechnical Engineering, American Society of Civil Engineers*, vol. 107, pp. 759-777, 1981.

[2] S.S. Chowdhury, K. Deb, A. Sengupta, Estimation of Design Parameters for Braced Excavation: Numerical Study. *International Journal of Geomechanics*, 13(3), pp.234-247, 2012.

[3] L. Rutkowski, *Computational Intelligence: Methods and Techniques*. Springer, 2008.

[4] J. Holland, *Adaptation in Natural and Artificial Systems (2nd ed.)*. The University of Michigan Press: Michigan, 1975; MIT Press: Cambridge, MA, 1992.

[5] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, Inc., NY., 1989.

[6] N. Turkkan, Discrete Optimization of Structures Using a Floating Point Genetic Algorithm. *Annual Conference of the Canadian Society for Civil Engineering*, Moncton, Nouveau - Brunswick, Canada, GCM-134/1-8, 2003.

[7] ThyssenKrupp GfT Bautechnik, *Sheet piling Handbook*. ThyssenKrupp Services, 2007.

[8] G. Swoboda, *Programmsystem FINAL, Finite Element Analysis Linearer und Nichtlinearer Stukturen, Version 7.1*, Druck Universität Innsbruck, Austria, 2001.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

NASH EVOLUTIONARY ALGORITHMS: TESTING PROBLEM SIZE IN RECONSTRUCTION PROBLEMS IN FRAME STRUCTURES

D. Greiner[1], J. Periaux[2], J.M. Emperador[1], B. Galván[1] and G. Winter[1]

[1] Institute of Intelligent Systems and Numerical Applications in Engineering (SIANI), Universidad de Las Palmas de Gran Canaria (ULPGC), 35017, Spain;
email: david.greiner@ulpgc.es

[2] Mathematical Information Technology Department (MIT) - University of Jyvaskyla, Finland & International Center for Numerical Methods in Engineering (CIMNE) – Universidad Politécnica de Cataluña, Spain;
email: jperiaux@gmail.com

Keywords: Structural Design, Frame Optimization, Evolutionary Algorithms, Nash Equilibrium, Game Strategies, Structural Optimization.

Abstract. *The use of evolutionary algorithms has been enhanced in recent years for solving real engineering problems, where the requirements of intense computational calculations are needed, especially when computational engineering simulations are involved (use of finite element method, boundary element method, etc). The coupling of game-theory concepts in evolutionary algorithms has been a recent line of research which could enhance the efficiency of the optimum design procedure and the quality of the design solutions achieved. They have been applied in several fields of engineering and sciences, mainly, in aeronautical and structural engineering (e.g: in computational fluid dynamics and solid mechanics problems). Among them, Nash-evolutionary algorithms (Nash-EAs) have been recently applied in the single-objective reconstruction inverse design problem in structural engineering (aiming to obtain the structure whose maximum stresses match those stresses considered as references), with successful speed-up of the structural optimum search. Several test cases of different search space size bar structures are handled here, with bar sized structures up to 105 bar elements. Particularly, frames -bar structures with rigid nodes where bending moment and shear effort should also be taken into consideration- are handled here. Influence of the structural size in the comparative performance of Nash-EAs will be investigated and tested. The performance of Nash-EAs improves significantly the one of the standard panmictic evolutionary algorithms. According to the results shown here, this advantage is greater when the problem size increases.*

1 INTRODUCTION

Evolutionary algorithms have been successfully spread their use among computational engineering applications for optimum design in the last decades (see e.g.: [1,2,3,4]). Among the strategies to increase the efficiency of those algorithms, parallelization and hybridization with game theory concepts have been proven useful. Particularly, the use of Nash-Evolutionary Algorithms (Nash-EAs) has been demonstrated as acceleration tools increasing convergence speed and/or quality of solutions in computational mechanics applications [5]. In this work, a study of performance of Nash-EAs versus standard panmictic EAs is developed, considering two structural test cases of different sizes in the reconstruction inverse problem, and comparing the influence of the number of bars (size of the problem search space).

In section 2, Nash EAs are described; the reconstruction inverse structural problem handled is explained in section 3, and the frame structural test cases are detailed and justified in section 4. This paper continues with the results and discussion in section 5, and finalizes with the conclusions in section 6.

2 NASH – EVOLUTIONARY ALGORITHMS

Nash-EAs were introduced in Sefrioui and Periaux [6] for solving computational fluid dynamics problems. They are based in hybridizing the mathematical concepts of Nash equilibrium (Nash, 1950-51) [7,8] (competitive game theory where players maximize their payoffs while taking into account the strategies of their competitors) in the evolutionary search: A set of subpopulations co-evolve simultaneously each of which deals only with a partition of the search variables. These subpopulations interact to evolve towards the equilibrium; when dealing with a single objective problem, a virtual Nash game approach has been applied in inverse shape optimization computational fluid dynamics and computational solid mechanics problems as an improvement technique versus the standard panmictic evolutionary algorithms. This approach has been successfully applied in the case of inverse problems where the fitness function objective is a sum of separable terms (such as the case of many shape optimization problems) [9,10].

3 STRUCTURAL PROBLEM

The aim of the structural reconstruction problem is to achieve the structure which fits most the maximum reference stresses. The optimum structural bar design is defined as a design in which some allocation of every bar in the structure has a maximum stress value as accurately equal as the maximum reference stress for that bar. Equation (1) shows the fitness function (FF) to be minimized (reconstruction problem).

$$Fitness\ \ Function = \sqrt{\sum_{i=1}^{Nbars} (\sigma_{MAX-i} - \sigma_{MAX-Ri})^2} \tag{1}$$

where $\sigma_{MAX\text{-}i}$ is the maximum calculated stress and $\sigma_{MAX\text{-}Ri}$ the maximum reference stress, in bar i. A null value of the fitness function in the reconstruction problem means a perfect match of stresses, and a location of the aimed structural design.

4 TEST CASES

4.1 Test Case 1: 55 bar sized frame structure

A fifty-five bar sized frame structural test case [11] is shown in figure 1, particularly the case of discrete cross-section type variables is handled here. The frame structure of reference

considered is the one corresponding to IPE330 cross section type in all beams (being the search space an interval between IPE080 and IPE500) and HEB450 cross section type in all columns (being the search space an interval between HEB100 and HEB450). Details about the corresponding maximum stress in each bar are available in [12].

A two player (two subpopulations) splitting territory approach will be used. Two Domain Decomposition (DD) player territories are tested in section 5. The distribution of bars in every case are shown as follows in figure 2 (black and cyan colors indicate membership to each territory, respectively): a Bottom-Up DD and a Left-Right DD. Territories divide the variable search space in two subsets of 27 and 28 bars.

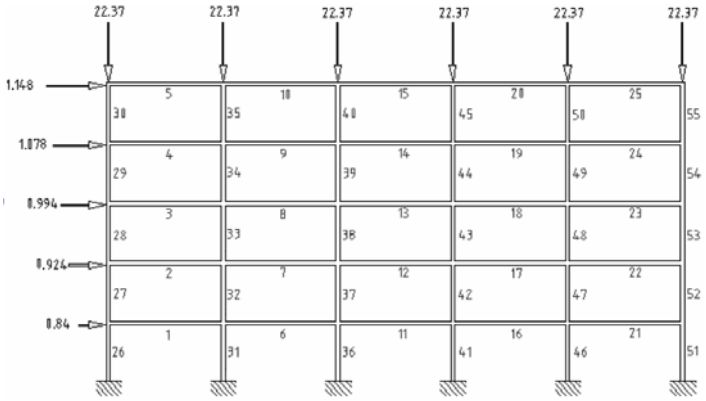


Figure 1: Computational domain, boundary conditions and loadings available in [11], 55 bar sized.

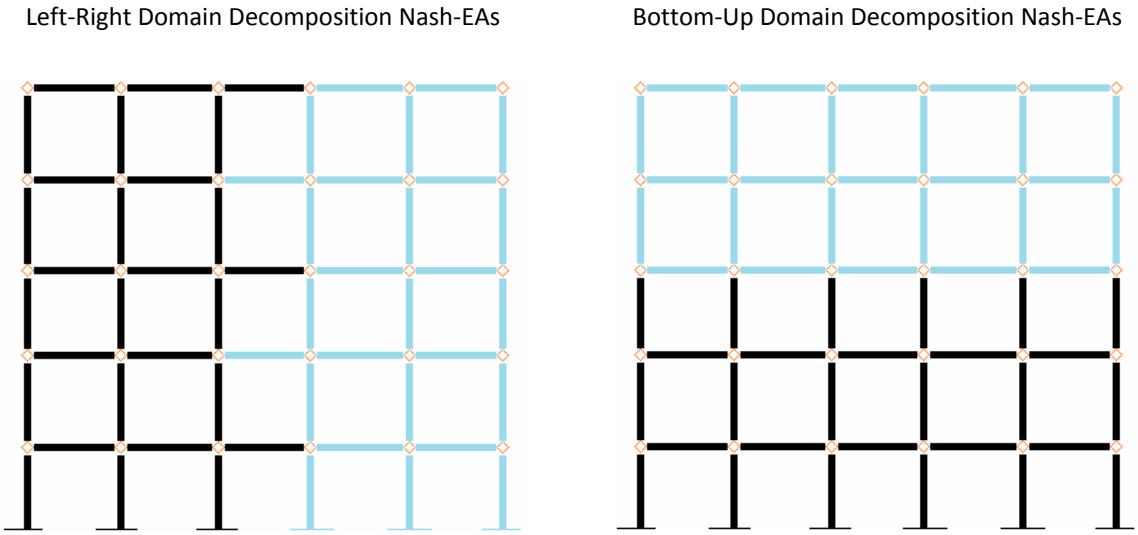


Figure 2: Nash EAs domain decompositions, 55 bar sized structure.

4.2 Test Case 2: 105 bar sized frame structure

This test case is designed purposely to study the influence of problem size in the performance of Nash-EAs versus panmictic EA. A one-hundred-five bar sized frame structural test case is shown in figure 3, particularly the case of discrete cross-section type variables is handled here. This test case is an horizontal extension of the one shown in section 4.1, with identical computational domain, boundary conditions and loadings, as well as identical cross section search space.

Also in this extended test case, a two player (two subpopulations) splitting territory approach will be used. Two Domain Decomposition (DD) player territories are tested in section 5. The distribution of bars in every case are shown as follows in figure 4 (black and cyan colors indicate membership to each territory, respectively): a Bottom-Up DD and a Left-Right DD. Territories divide the variable search space in two subsets of 52 and 53 bars.

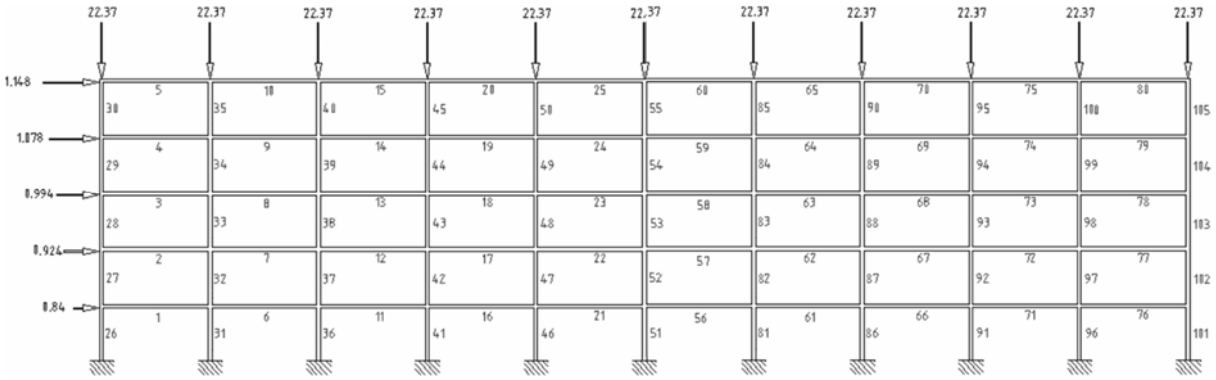
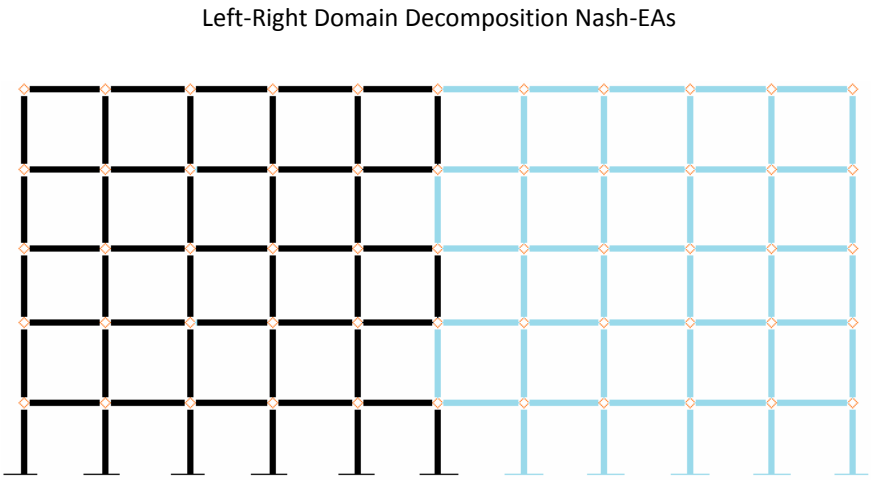


Figure 3: Computational domain, boundary conditions and loadings as in Figure 1, 105 bar sized (horizontal extension of test case 1).



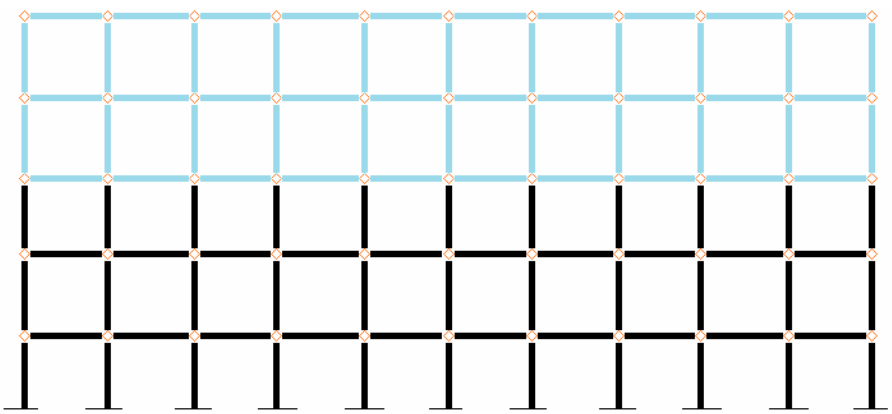


Figure 4: Nash EAs domain decompositions, 105 bar sized structure.

5 RESULTS AND DISCUSSION

5.1 Test Case 1: 55 bar sized frame structure

A population size of 80 individuals, with uniform crossover, gray codification and 0.4% mutation rate are used in a set of 30 independent executions of the three evolutionary algorithms tested: standard panmictic EA, Nash-EAs with left-right partition and Nash-EAs with bottom-up partition. The stopping criterion is set as maximum number of fitness evaluations equal to $2 \cdot 10^5$.

Figure 5 shows the final average number of fitness evaluations versus the standard deviation of fitness evaluations of those executions of the three algorithms required to achieve the optimum (null) solution (all the three EAs were able to reach the optimum in all the runs in this first test case, as shown in table 1). Convergence curves of the evolution of the average, best and standard deviation values of the fitness function value are shown in figures 6, 7 and 8, respectively.

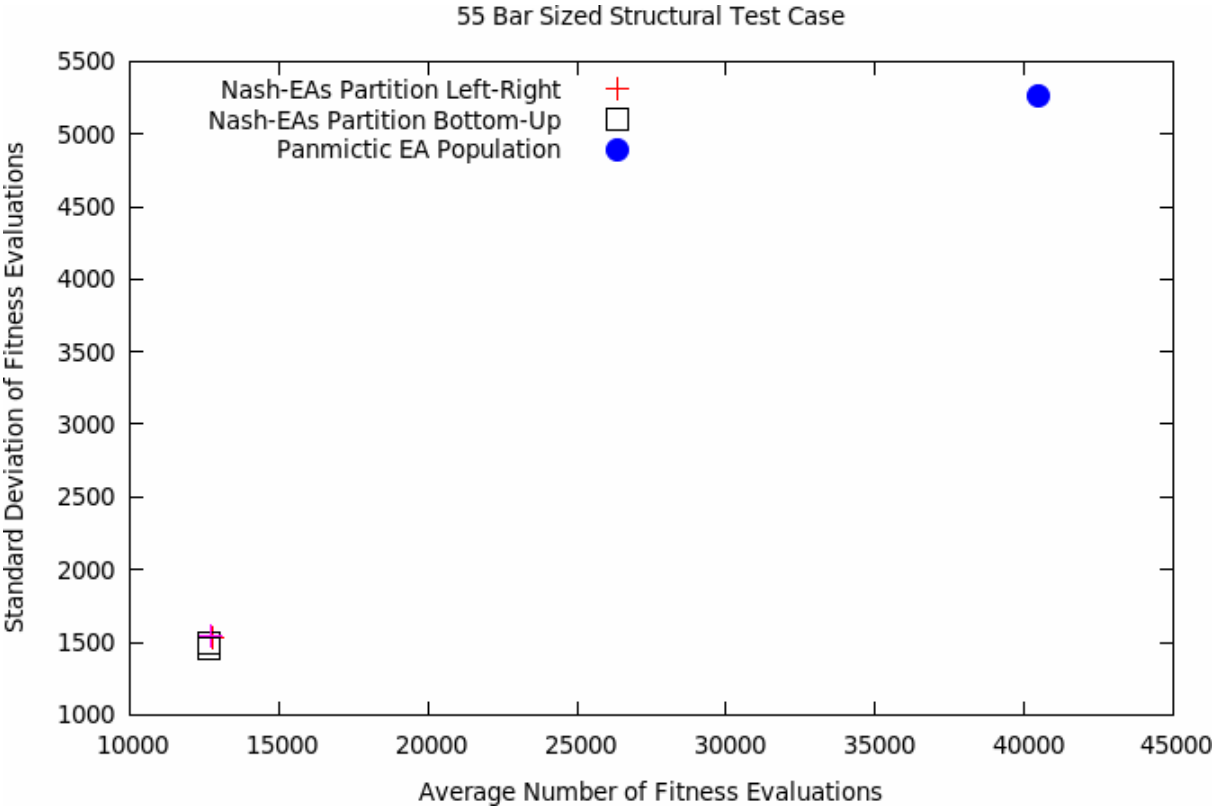


Figure 5: Panmictic EA and Nash-EAs Results - Fitness Evaluations required to obtain the optimum design out of 30 independent runs.

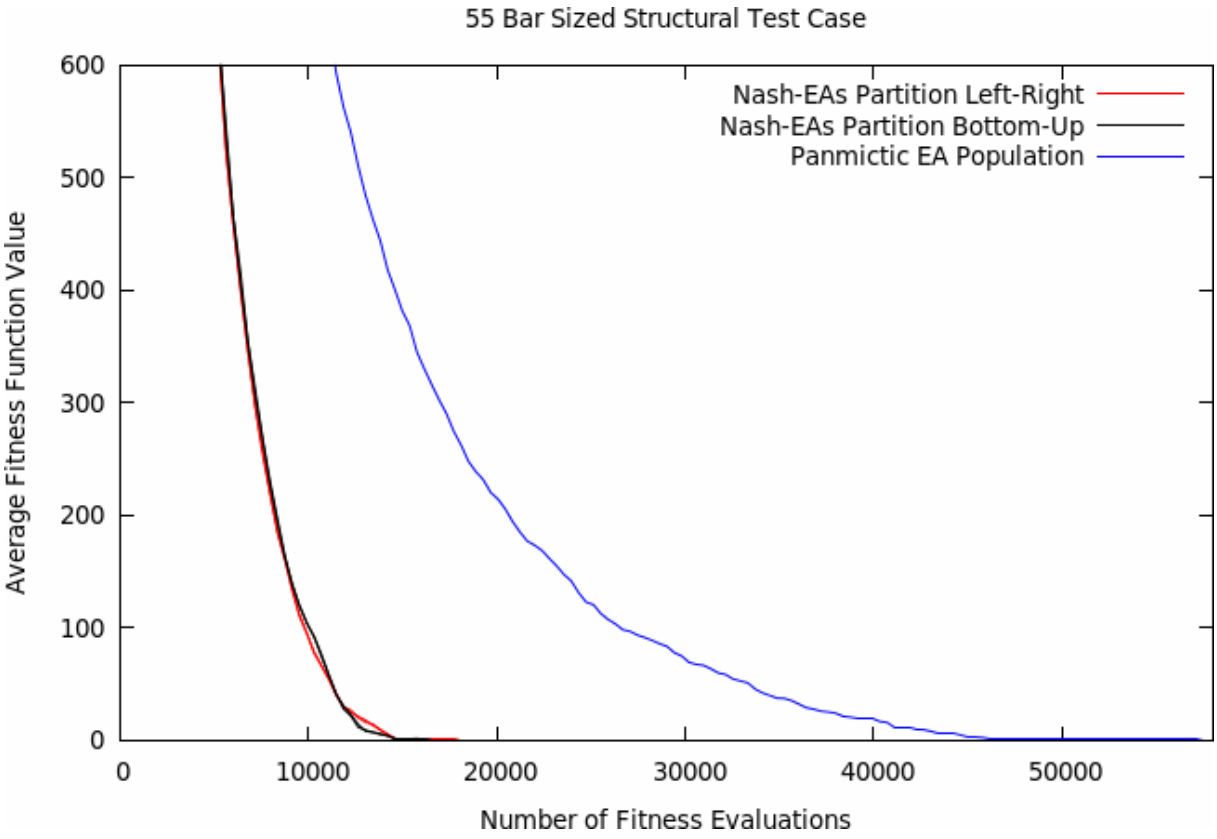


Figure 6: Average Convergence; Nash-EAs versus Panmictic EA; 55 bar sized test case

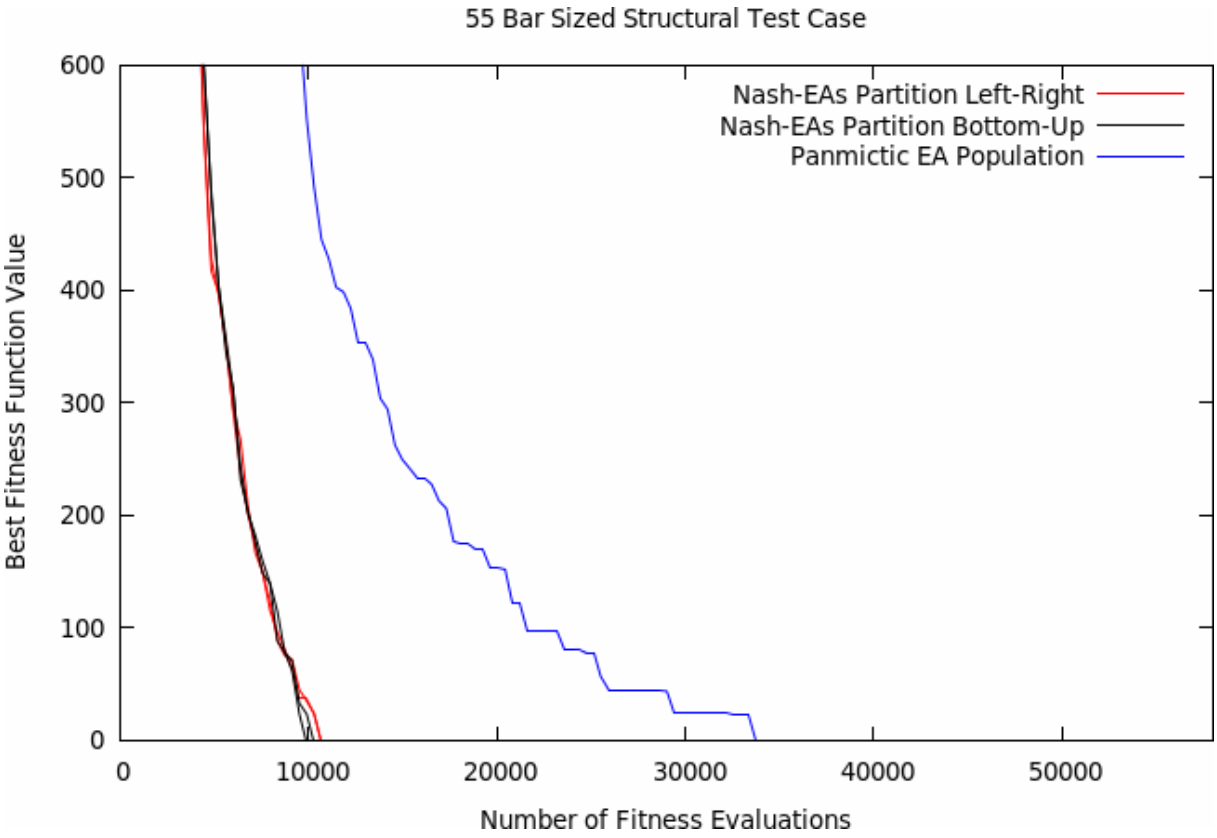


Figure 7: Best Convergence; Nash-EAs versus Panmictic EA; 55 bar sized test case

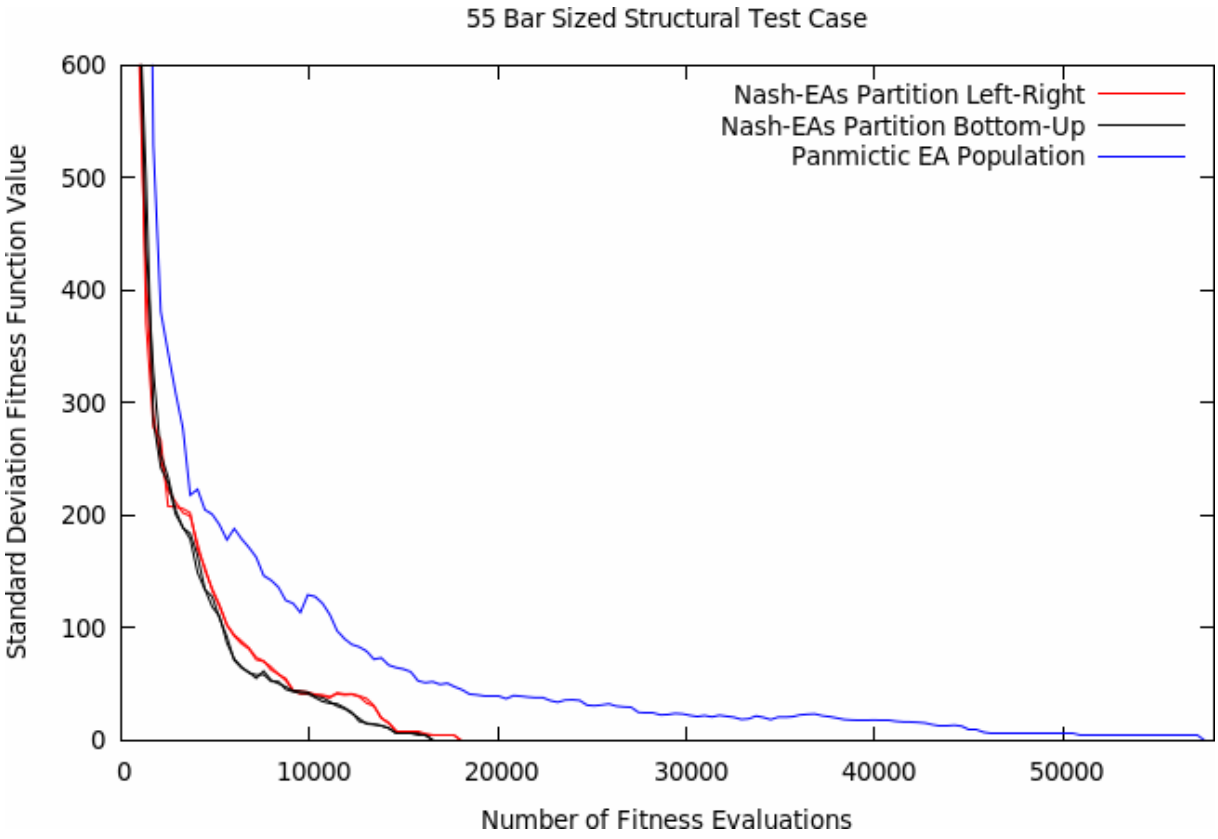


Figure 8: Standard Deviation Convergence; Nash-EAs versus Panmictic EA; 55 bar sized test case

Algorithm	Average Number Fitness Evaluations	Standard Deviation Fitness Evaluations	Best Value Fitness Evaluations	Worst Value Fitness Evaluations	Success Rate
Panmictic EA	40458	5258.8	33698	57488	**30 / 30**
Nash-EAs Left-Right Player 1	12781	1528.7	10688	18098	**30 / 30**
Nash-EAs Left-Right Player 2	12742	1536.8	10688	18098	**30 / 30**
Nash-EAs Bottom-Up Player 1	**12638**	1496.9	10298	**16538**	**30 / 30**
Nash-EAs Bottom-Up Player 2	12664	**1452.0**	**9908**	**16538**	**30 / 30**

Table 1: Final Results (over 30 independent runs, after a maximum number of $2 \cdot 10^5$ fitness evaluations), 55 bar sized test case.

5.2 Test Case 2: 105 bar sized frame structure

A population size of 150 individuals, with uniform crossover, gray codification and 0.4% mutation rate are used in a set of 30 independent executions of the three evolutionary algorithms tested: standard panmictic EA, Nash-EAs with left-right partition and Nash-EAs with bottom-up partition. The stopping criterion is set as maximum number of fitness evaluations equal to $2 \cdot 10^6$.

Table 2 shows the final average, standard deviation, best and worst number of fitness evaluations of those executions of the three algorithms required to achieve the optimum (null) solution, expressed as the success rate. The two Nash-EAs were able to reach the optimum in all the runs in this second test case, while the panmictic EA was completely unsuccessful (no execution was able to achieve the optimum. Convergence curves of the evolution of the average (figure 9 and zoomed figure 10), best (figure 11 and zoomed figure 12) and standard de-

viation values (figure 13 and zoomed figure 14) of the fitness function value are shown in the aforementioned figures, respectively.

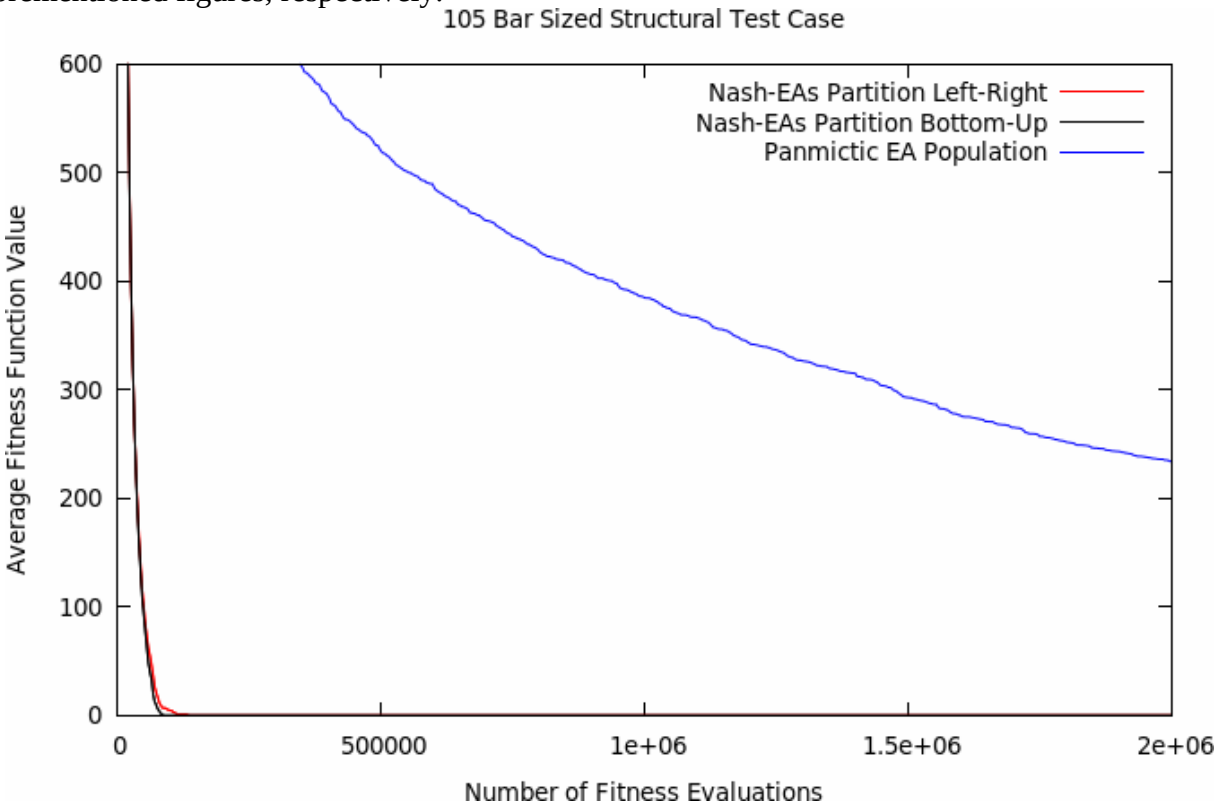


Figure 9: Average Convergence; Nash-EAs versus Panmictic EA; 105 bar sized test case

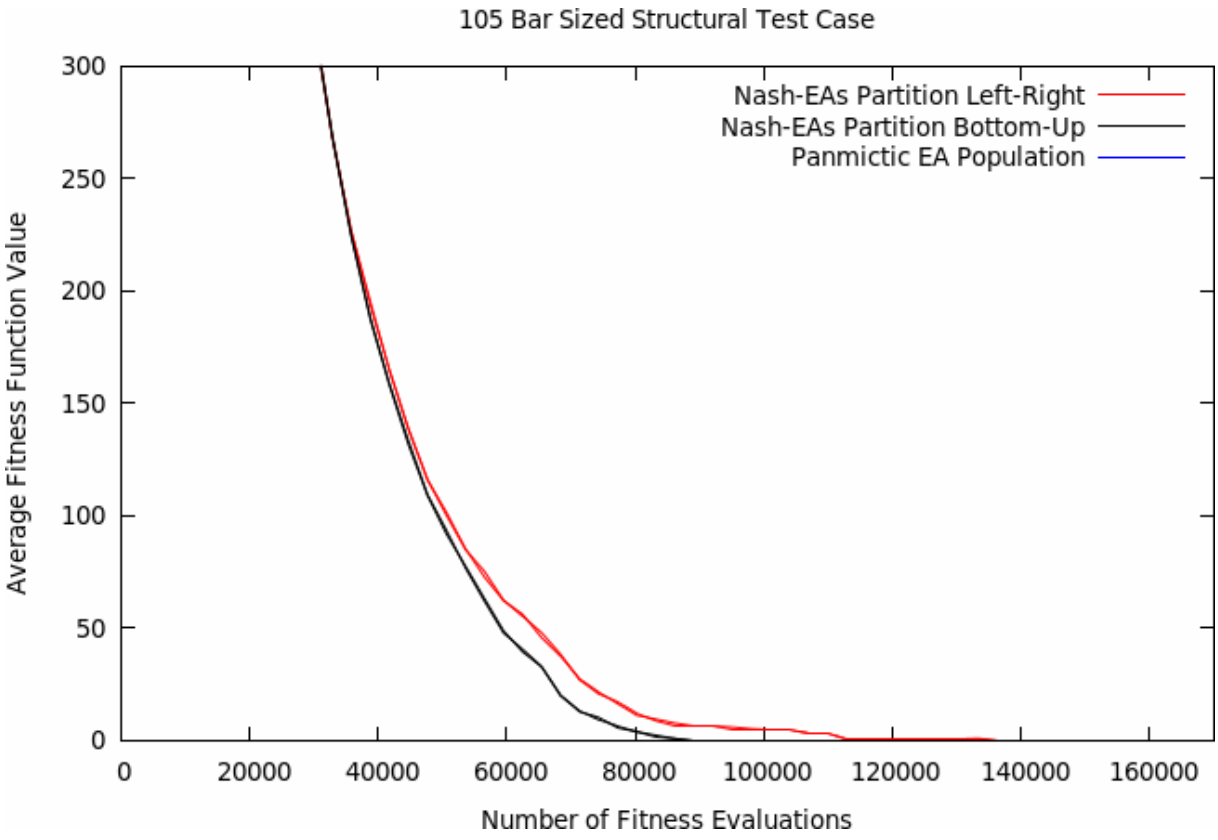


Figure 10: Average Convergence; Nash-EAs versus Panmictic EA; 105 bar sized test case (Zoomed figure 9)

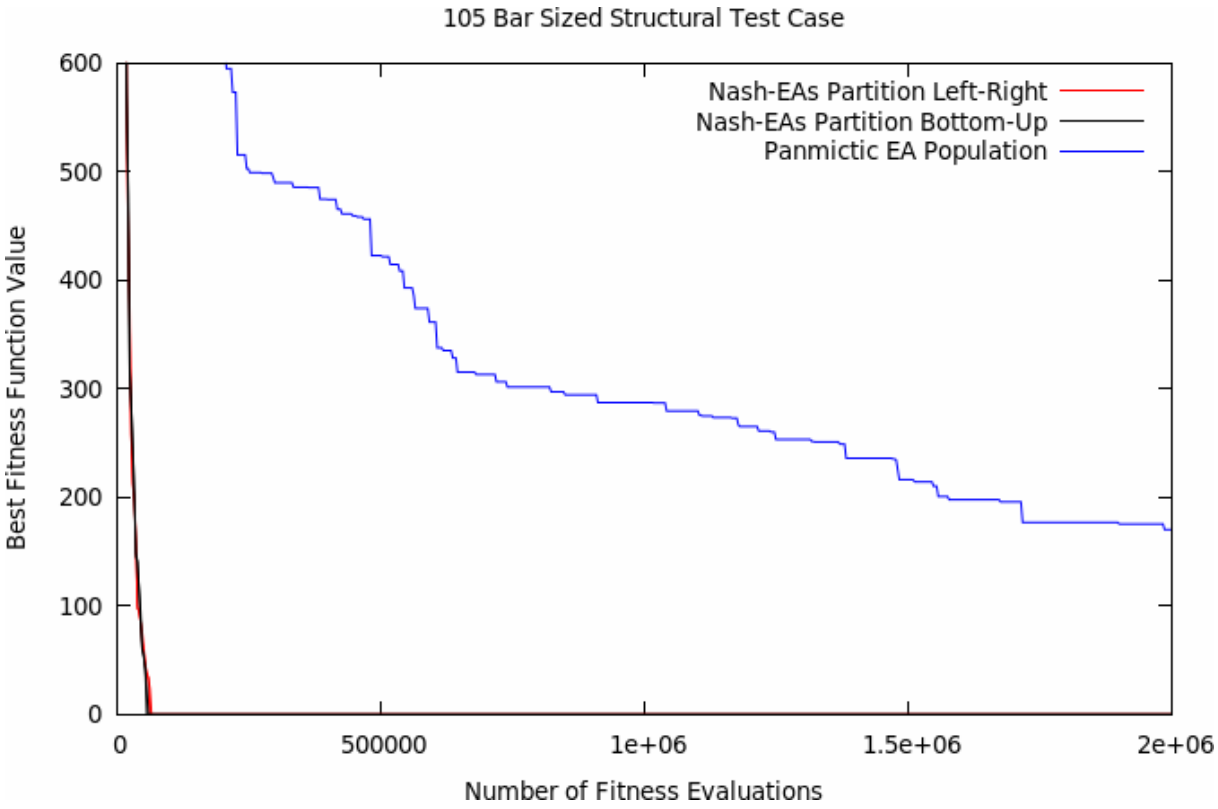


Figure 11: Best Convergence; Nash-EAs versus Panmictic EA; 105 bar sized test case

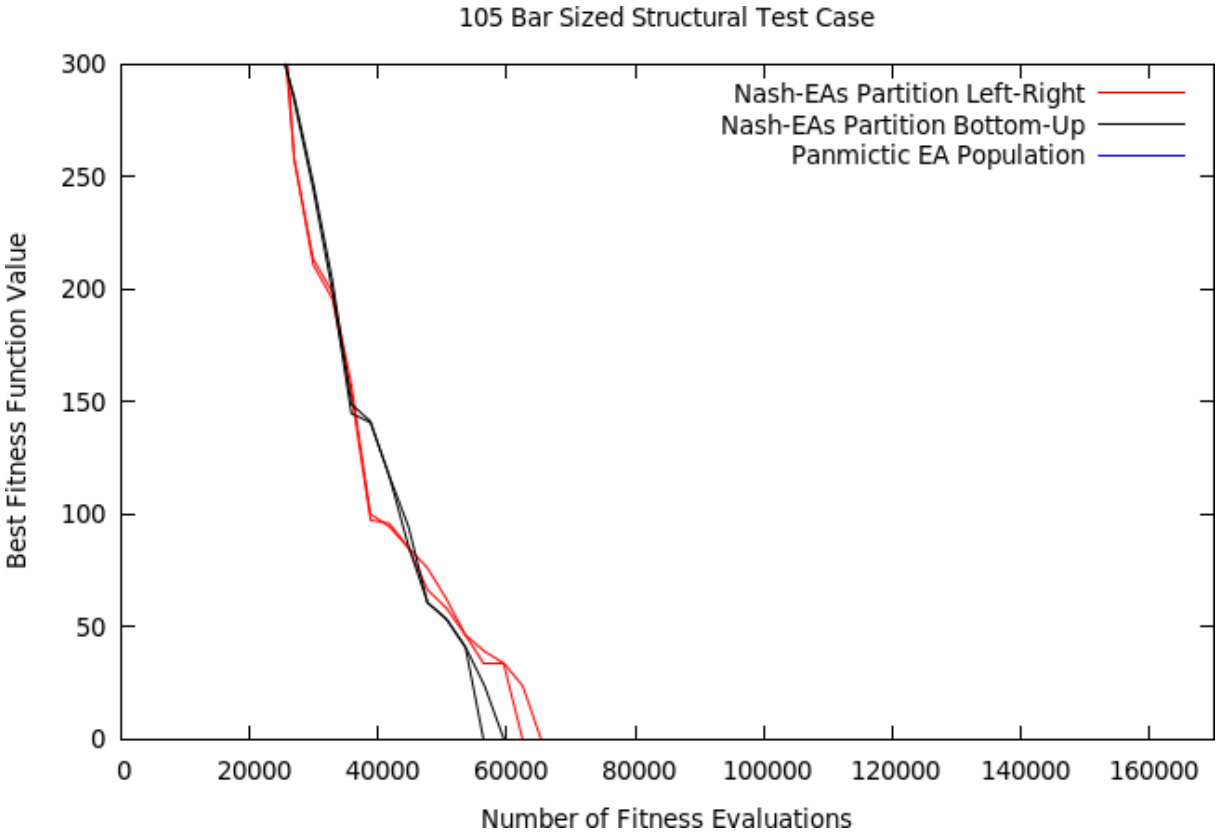


Figure 12: Best Convergence; Nash-EAs versus Panmictic EA; 105 bar sized test case (Zoomed figure 11)

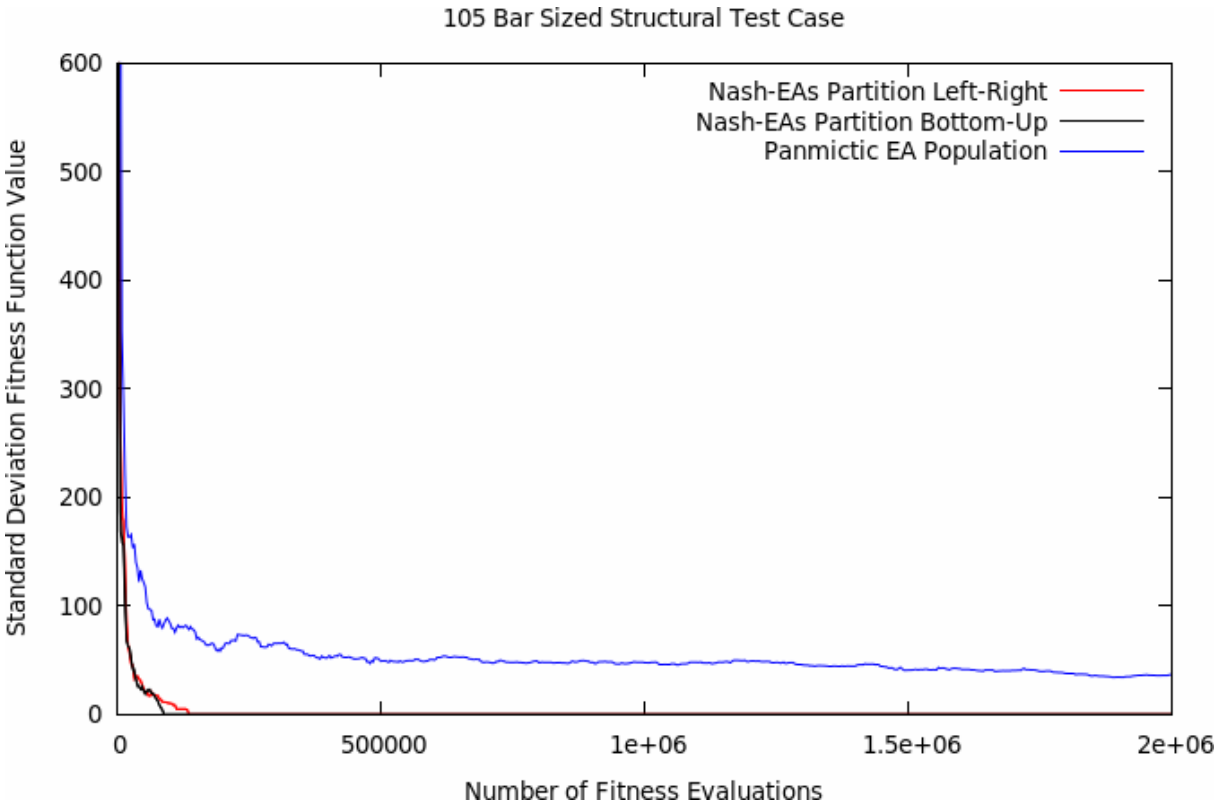


Figure 13: Standard Deviation Convergence; Nash-EAs versus Panmictic EA; 105 bar sized test case

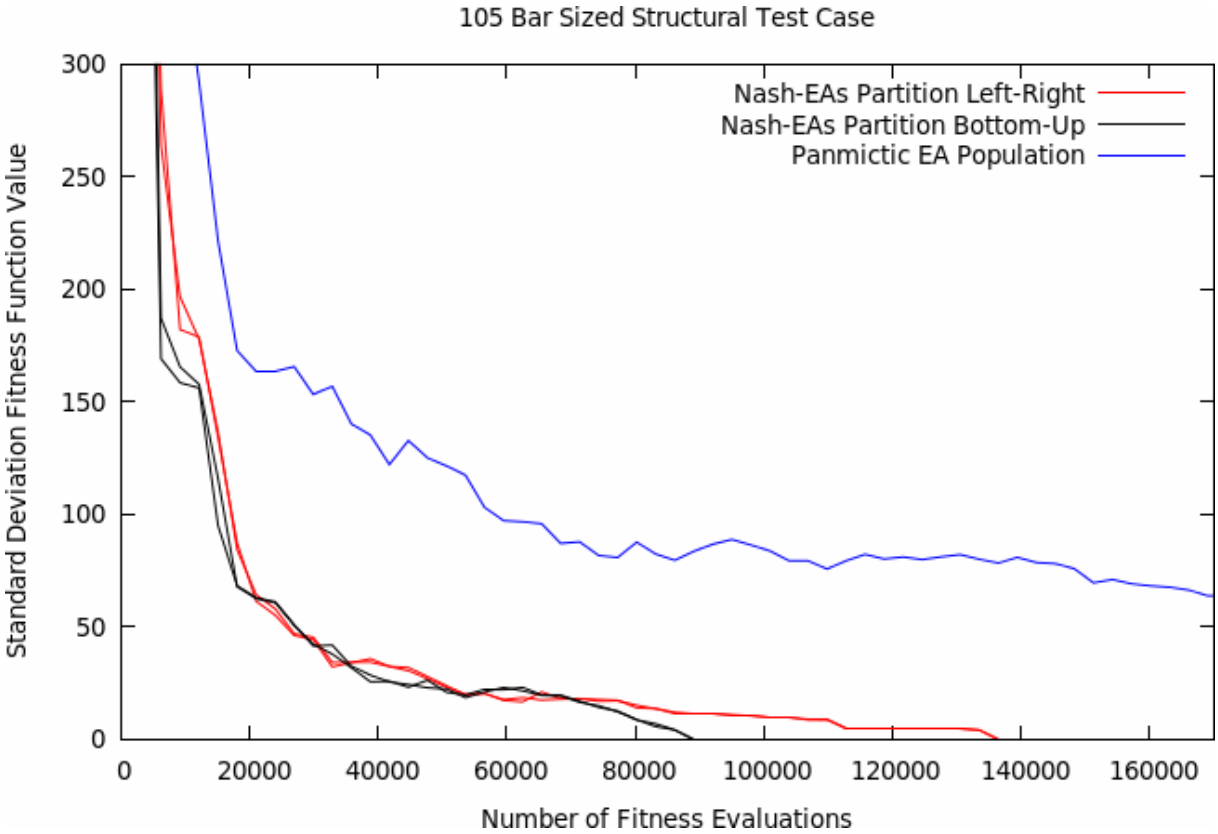


Figure 14: Standard Deviation Convergence; Nash-EAs versus Panmictic EA; 105 bar sized test case (Zoomed figure 13)

Algorithm	Average Number Fitness Evaluations	Standard Deviation Fitness Evaluations	Best Value Fitness Evaluations	Worst Value Fitness Evaluations	Success Rate
Panmictic EA	Not reached	Not available	Not reached	Not reached	0 / 30
Nash-EAs Left-Right Player 1	84559.3	16617.1	65418	136458	**30 / 30**
Nash-EAs Left-Right Player 2	84658.0	16546.9	62458	136458	**30 / 30**
Nash-EAs Bottom-Up Player 1	73114.0	**7472.6**	59498	**89098**	**30 / 30**
Nash-EAs Bottom-Up Player 2	**72916.7**	7555.6	**56538**	**89098**	**30 / 30**

Table 2: Final Results (over 30 independent runs, after a maximum number of $2{\cdot}10^6$ fitness evaluations), 105 bar sized test case.

5.3 Discussion

In the smaller size test case (first one), both panmictic and Nash-EAs were equally able to reach the optimum solution consistently in all the executions. Nevertheless, Nash-EAs were much faster than panmictic EA (see table 1 and figures 6 to 8). Between the Nash-EAs Left-Right domain decomposition and the Nasb-EAs Bottom-Up domain decomposition, results are slightly better in the latter.

In the bigger size test case (second one), only Nash-EAs were able to reach the optimum solution consistently in all the executions. On the contrary, the panmictic EA was completely unsuccessful. Also Nash-EAs were much faster than panmictic EA (see table 2 and figures 9 to 14). Between the Nash-EAs Left-Right domain decomposition and the Nasb-EAs Bottom-Up domain decomposition, results are slightly better in the latter, fact that is shared with the first test case.

CONCLUSIONS

Results obtained through a reconstruction inverse problem in structural frames have evidenced that the advantage of use and efficiency of Nash-EAs versus standard panmictic EAs is greater when higher size search spaces are involved. Here particularly, two search space sizes were compared: 55 and 105 bar sized problems, which correspond to discrete search spaces of 2^{220} and 2^{420}, respectively. Robustness of the search has remained unaffected in Nash-EAs, while seriously affected in panmictic EAs.

Acknowledgements

The first author gratefully acknowledges support given at the Mathematical Information Technology Department, University of Jyväskylä (Finland), during 2016.

REFERENCES

[1] N. Lagaros, M. Papadrakakis, *Engineering and Applied Sciences Optimization.* Computational Methods in Applied Sciences, **38**, Springer, 2015.

[2] D. Greiner, B. Galván, J. Periaux, N. Gauger, K. Giannakoglou, G. Winter, *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences.* Computational Methods in Applied Sciences, **36**, Springer, 2015.

[3] D. Greiner, J.M. Emperador, B. Galvan, G. Winter, J. Periaux, Optimum Structural Design using Bio-inspired Search Methods: a Survey and Applications. V. Becerra and M. Vasile eds. *Computational Intelligence in Aerospace Sciences*, Progress in Aerospace

Sciences, **244**, American Institute of Aeronautics and Astronautics AIAA, 373 - 414, 2014.

[4] D. Greiner, J.M. Emperador, G. Winter, Multiobjective Optimisation of Bar Structures by Pareto-GA. *European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS 2000*, Barcelona, España, 2000.

[5] J. Periaux, F. González, D.S.C. Lee, *Evolutionary Optimization and Game Strategies for Advanced Multi-Disciplinary Design*. Intelligent Systems, Control and Automation: Science and Engineering, **75**, Springer, 2015.

[6] M. Sefrioui, J. Periaux, Nash Genetic Algorithms: Examples and Applications. *Proceedings of the 2000 Congress on Evolutionary Computation CEC-2000*, IEEE Press, La Jolla Marriott Hotel La Jolla, California, USA, 509-516, 2000.

[7] J.F. Nash, Equilibrium points in N-person games. *Proceedings of the National Academy of Science*, **36**, 46–49, 1950.

[8] J.F. Nash, Non-cooperative games, *Annals of Mathematics*, **54(2)**, 286-295, 1951.

[9] J. Leskinen, H. Wang, J. Périaux, Increasing parallelism of evolutionary algorithms by Nash games in design inverse flow problems. *Engineering Computations*, **30(4)**, 581-600, 2013.

[10] D. Greiner, J. Periaux, J. M. Emperador, B. Galvan, G. Winter, A Study of Nash-Evolutionary Algorithms for Reconstruction Inverse Problems in Structural Engineering. D. Greiner et al. eds. *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences,* Computational Methods in Applied Sciences, **36**, Springer, 321-333, 2015.

[11] D. Greiner, G. Winter, J.M. Emperador, Single and Multi-objective Frame Optimization by Evolutionary Algorithms and the Auto-adaptive Rebirth Operator. *Computer Methods in Applied Mechanics and Engineering*, **37(35)**, 3711–3743, 2004.

[12] J. Periaux, D. Greiner, Efficient Parallel Nash Genetic Algorithm for Solving Inverse Problems in Structural Engineering. P. Neittaanmäki et al. eds. *Mathematical Modeling and Optimization of Complex Structures,* Computational Methods in Applied Sciences, **40**, Springer, 205-228, 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

OPTIMUM DESIGN OF REINFORCED CONCRETE FRAMES ACCORDING TO EC8 AND MC2010 WITH GENETIC ALGORITHMS

Panagiotis E. Mergos[1]

[1] City University London
Northampton Square, EC1V 0HB, London UK
e-mail: panagiotis.mergos.1@city.ac.uk

Keywords: Reinforced concrete, seismic, performance-based, Eurocodes, fib Model Code 2010, optimization, genetic-algorithms.

Abstract. *The need for cost-efficient seismic design in limited time has led to the development of automated structural optimization methodologies. Genetic Algorithms (GA) belong to the class of stochastic, nature-inspired heuristic optimization algorithms. GA can be easily implemented and applied to advanced structural problems since they don't require use of gradients of cost or constraints functions. Furthermore, they are able to identify global optima as opposed to local optimum solutions. Early efforts to optimise seismic design of concrete structures were based on traditional seismic design code approaches. The last two decades, performance- and deformation-based seismic design methodologies have emerged. These methodologies provide enhanced structural damage control for different earthquake intensities reducing both economic losses and human casualties. Lately, a fully-fledged performance- and deformation-based design methodology has been incorporated in the fib Model Code for Concrete Structures MC2010, which is meant to serve as a guidance document for future design codes of concrete structures. To the best of author's knowledge, there exists no study investigating optimum seismic design solutions in accordance with MC2010. The aim of this study is to develop a GA-based seismic design optimization framework for reinforced concrete frames in accordance with the provisions of both Eurocode 8 (EC8) and MC2010. Application of this framework to reinforced concrete frames is conducted and comparisons of the optimum solutions obtained by the two seismic design guidelines are made. The advantages and disadvantages of the two seismic design methodologies, in the context of structural optimization, are highlighted and discussed.*

1 INTRODUCTION

The need for improved control of structural damage for different levels of seismic action has led to the development of performance-based seismic design. Performance-based seismic design is a transparent and direct design framework that requires a set of performance levels to be met for different levels of seismic hazard. Performance levels are related to the level of structural damage of the structure, which in turn is directly related to structural member deformations and/or inter-story drifts.

The new *fib* Model Code 2010 (MC2010) includes a fully-fledged performance-based seismic design and assessment methodology for various levels of seismic hazard (*fib* 2010, Fardis 2013). MC2010 will serve as a basis for future codes for concrete structures.

In MC2010, each performance limit state corresponds to a specific physical condition of the structure and it is expressed in terms of deformation limits of the structural members providing direct control of allowable seismic damage. The levels of seismic hazard are identified by their annual probability of being exceeded. Seismic actions are specified in terms of acceleration time-histories of the ground motion components. The reference method for determining seismic demands is the most rigorous inelastic response history analysis with step-by-step integration of the equation of motion in the time domain (Fardis 2013).

Extensive research has been conducted over the past decades on optimum seismic design of structures (Fragiadakis and Lagaros 2011). However, only a small part of this research has been dedicated to reinforced concrete structures. This can be partially attributed to the complex nature and detailing of reinforced concrete structures that increases significantly the number of design variables (Sarma and Adeli 1998).

The number of research studies on optimization of performance and deformation-based seismic design of reinforced concrete structures is limited. Ganzerli *et al*. (2000) were the first, to the best of the author's knowledge, to consider seismic optimization with performance-based constraints. Chan and Zou (2004) examined optimum seismic design of reinforced concrete frames by employing optimality criteria approach. Lagaros and Papadrakakis (2007) compared the provisions of EC8 for seismic analysis of 3D reinforced concrete structures with a performance-based seismic design methodology in the framework of multi-objective optimization. Fragiadakis and Papadrakakis (2008) presented a performance-based optimum seismic design methodology for reinforced concrete frames based on nonlinear time history analyses. Inter-story drifts were used as performance criteria. Gencturk (2013) investigated performance-based seismic design optimization of reinforced concrete and reinforced engineered cementitious composites (ECC) frames, by using Taboo Search optimization algorithm.

It can be concluded from the above, that no study has been conducted so far on optimization of reinforced concrete frames in accordance with MC2010 seismic design provisions. To address this gap, this study presents optimum seismic design solutions of reinforced concrete frames obtained by MC2010 and compares them with optimum designs following EC8 (CEN 2004) guidelines. To serve this goal, a general computational optimization framework for reinforced concrete frames is developed that makes use of a genetic algorithm able to track global optima of complex problems with discrete design variables.

2 OPTIMIZATION OF REINFORCED CONCRETE FRAMES

2.1 Optimization framework

In optimization problem formulations, the goal is to minimize an objective function $C(\boldsymbol{x})$ subject to m number of constraints $g_j(\boldsymbol{x}) \leq 0$ (j=1 to m). A design solution is represented by the

design vector $\boldsymbol{x}$, which contains n number of independent design variables x_i (i=1 to n). In deterministic structural seismic design optimization, the objective function $C(\boldsymbol{x})$ is typically the initial cost of the structure. Constraints g_j are either related to engineering demand parameters (*EDP*) (e.g. forces, displacements, rotations, drifts, etc.) or to detailing rules set by design codes and construction practice. To realistically represent construction practice, design variables x_i typically take values from discrete sets of values $\boldsymbol{D}_i=(d_{i1}, d_{i2}, \ldots, d_{iki})$, where d_{ip} (p=1 to k_i) is the p-th possible discrete value of design variable x_i and k_i is the number of allowable discrete values of x_i. For reinforced concrete structures, design variables are generally related to concrete section dimensions and steel reinforcement.

2.2 Genetic algorithm (GA)

Genetic Algorithm (GA) (Holland 1975) belongs to the class of stochastic, nature-inspired heuristic algorithms. It is based on Darwin's theory of natural selection and evolution. Genetic algorithm can be easily implemented and applied to advanced optimization problems since it doesn't require use of gradients of cost or constraints functions. Furthermore, it is able to identify global optima as opposed to local optimum solutions.

The GA iteratively modifies populations (generations) of individuals in order to evolve toward an optimum solution. An individual $\boldsymbol{x}$ (genome) represents a candidate solution to the optimization problem. The values of the design variables x_i (i=1 to n) forming each individual are called genes. Furthermore, the objective function of each individual is known as fitness function. The best fitness value of a generation is the smallest fitness value of all individuals of the generation. In order to create the next population, the genetic algorithm selects certain individuals in the current population (parents) and uses them to create individuals in the next generation (children).

In this study, the mixed integer GA as implemented in MATLAB-R2015a (MathWorks 2015) is employed. This algorithm can handle both continuous and discrete design variables. The algorithm is able to account for nonlinear constraints by using the penalty function approach. The genetic algorithm in this study is terminated when one of the following stopping criteria is met:

i) Number of generations exceeds a pre-specified maximum number of generations.

ii) The mean relative variation of the best fitness value does not exceed a pre-specified tolerance over a pre-specified number of generations.

2.3 Design variables

As shown in Fig. 1, design variables are separated in column and beam section design variable sub-vectors. Assembly of these sub-vectors forms the design variables vector $\boldsymbol{x}$.

Column design variable sub-vectors are the heights $\boldsymbol{h_c}$ and widths $\boldsymbol{b_c}$ of the column sections, the diameters $\boldsymbol{d_{bc}}$ and numbers of main bars per side $\boldsymbol{n_c}$, assumed the same for all column section sides for simplicity, the diameters $\boldsymbol{d_{bwc}}$, spacings $\boldsymbol{s_c}$ and numbers of legs $\boldsymbol{n_{wc}}$ of transverse reinforcement assumed again the same in both column section directions for simplification purposes.

Beam design variable sub-vectors are the heights $\boldsymbol{h_b}$ and widths $\boldsymbol{b_b}$ of the beam sections, the diameters $\boldsymbol{d_{bt}}$ and numbers of main bars of the top sides $\boldsymbol{n_{tb}}$, the diameters $\boldsymbol{d_{bb}}$ and numbers of main bars of the bottom sides $\boldsymbol{n_{bb}}$, the diameters $\boldsymbol{d_{bwb}}$, spacings $\boldsymbol{s_b}$ and numbers of legs $\boldsymbol{n_{wb}}$ of transverse reinforcement parallel to beam section heights.

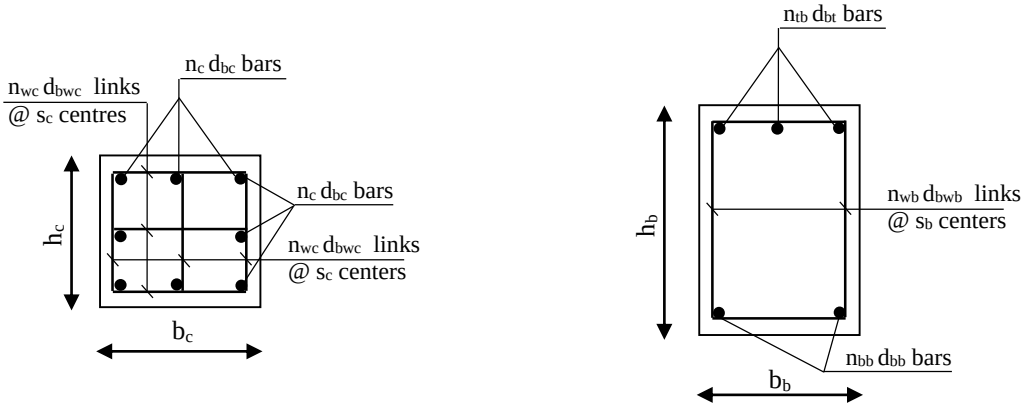


Figure 1: Design variables: a) column sections, b) beam sections

2.4 Objective function

In this study, the objective function $C(\boldsymbol{x})$ is the initial or construction cost of the reinforced concrete frames. The construction cost consists of the cost of the materials and the cost of the formwork of beam and column members. The following unit costs are assumed for concrete: 100Euros/m^3, steel: 1Euro/kg and formwork: 15Euros/m^2.

2.5 Design constraints

In seismic design of reinforced concrete frames, constraints $g_j(\boldsymbol{x})$ are either related to engineering demand parameters (*EDP*) (e.g. forces, displacements, rotations, drifts, etc.) or to detailing rules set by design codes and construction practice. In the first case, an *EDP* must remain below a limit value EDP_{cap}. This type of constraints can be written in the following normalized form

$$EDP \leq EDP_{cap} \rightarrow \frac{EDP}{EDP_{cap}} - 1 \leq 0 \tag{1}$$

Regarding detailing requirements, the constraints can be expressed in terms of structural design parameters *SDP*. It is noted that a *SDP* can be a design variable itself (e.g. column height, main bar diameter) or a simple function of the design variables like the volumetric ratios of steel reinforcement.

In some cases, it is required that a *SDP* remains lower than or equal to a maximum value SDP_{max}. This category of constraints is written in the following general form:

$$SDP \leq SDP_{max} \rightarrow \frac{SDP}{SDP_{max}} - 1 \leq 0 \tag{2}$$

In other cases, it is required that a *SDP* is greater than or equal to a minimum value SDP_{min}. The latter family of constraints is expressed in the normalized form shown below:

$$SDP \geq SDP_{min} \rightarrow \frac{SDP_{min}}{SDP} - 1 \leq 0 \tag{3}$$

3 OPTIMUM DESIGN ACCORDING TO EUROCODE 2

Prior to designing for seismic actions, RC frames must be designed to resist dead and live loads. Eurocode 2 (EC2 – CEN 2004) provisions are applied in this study for designing against static loads. EC2 provisions consist of a number of detailing rules and a number of requirements related to *EDPs*.

Regarding detailing rules, design constraints of minimum volumetric ratio of longitudinal reinforcement, minimum diameter of longitudinal and transverse reinforcement, minimum distance between two longitudinal steel bars and minimum volumetric ratio of transverse reinforcement are expressed in the general form of Eq. (3). On the other hand, constraints of the maximum volumetric ratio of longitudinal reinforcement, maximum spacing of shear reinforcement and maximum distance of unrestrained next to restrained main bars of columns are written in the form of Eq. (2).

For the ULS, *EDPs* are member forces (moments and shear forces) derived by linear elastic analysis for the following load combination, where G_k represents the characteristic value of the permanent action and Q_k stands for the characteristic value of the variable action.

$$S_d = 1.35G_k + 1.50Q_k \tag{4}$$

EDPs constraints are written in the general form of Eq. (1), where capacities R_d are derived by using characteristic material strengths divided by partial safety factors equal to γ_c=1.50 for concrete and γ_s=1.15 for reinforcing steel. For bending moments of column members, moment capacities are calculated for the axial load demand of the load combination under examination.

For beam deflections, the limiting span to depth ratio approach is used herein ensuring that deflections are limited to span/250. Moreover, crack control is achieved by limiting maximum bar size or spacing.

4 OPTIMUM SEISMIC DESIGN ACCORDING TO EUROCODE 8

Seismic design according to EC8 can be performed either without provisions for energy dissipation and ductility (Ductility Class Low – DCL) or with provisions for energy dissipation and ductility (Ductility Classes Medium and High – DCM and DCH). DCM and DCH differ in the levels of lateral strength and allowable inelastic response. DCH allows for further reductions in seismic forces, but requires more demanding prescriptive rules for increasing ductility capacities than DCM.

For DCL, all seismic *EDPs* are calculated from the seismic load combination shown below, where design seismic actions E_d are calculated by the design response spectrum that is derived from the elastic response spectrum reduced by the behaviour factor q. ψ_2 is the quasi-permanent load combination coefficient of the variable action. Reference analysis method of EC8 is the modal response spectrum analysis. However, for regular buildings with unimportant higher modes the linear elastic lateral force method can also be applied.

$$S_{Ed} = G_k + \psi_2 \cdot Q_k + E_d \tag{5}$$

For DCM and DCH, first the dissipative zones of structural members (typically located at the ends) are designed in bending under the seismic design load combination. Next, capacity design principles are forced to ensure ductile structural response. In particular, column sections

are designed in bending following the strong column – weak beam capacity rule to prevent soft storey failure mechanisms. Moreover, capacity design in shear is applied to beam and column members and joints to preclude brittle shear failures.

In addition to the above, RC frames are checked for a 'frequent' earthquake with 10% probability of exceedance in 10 years (10/10) to satisfy the Damage Limitation (DL) limit state. Checks verify that interstorey drifts developed for the 'frequent' earthquake are less than limit values depending on the type of non-structural elements (e.g. 1% for non-structural elements that don't interfere with structural response).

P-delta (2nd order) effects are considered at the i storey level with calculating ratio θ^i from Eq. (6). In this equation, N_{tot}^i and V_{tot}^i are the total vertical and shear load at the storey respectively, $\Delta\delta^i$ is interstorey drift and H^i is storey height. It is required that θ^i never exceeds 0.2. Furthermore, if θ^i exceeds 0.1 then 2nd order effects are taken into account by multiplying 1st order effects by the magnification factor $1/(1-\theta^i)$.

$$\theta^i = \frac{N_{tot}^i \cdot \Delta\delta^i}{V_{tot}^i \cdot H^i} \tag{6}$$

All previous requirements are regarded as *EDPs* constraints and are included in the optimization problem in the general form of Eq. (1). The *EDPs* are member bending moments and shear forces, interstorey drifts and θ^i ratios. Apart from *EDPs* constraints and EC2 detailing rules, DCM and DCH necessitate additional or stricter detailing rules in the critical regions to accommodate local ductility demands.

The additional column constraints of minimum cross-section sides, minimum volumetric ratio of longitudinal reinforcement, minimum diameter of transverse reinforcement, minimum number of bars per side, and minimum confinement of transverse reinforcement in critical regions are expressed in the general form of Eq. (3). The same holds for the additional beam constraints in critical regions such as minimum volumetric ratio of longitudinal reinforcement, minimum longitudinal bar diameter for DCH, minimum bottom reinforcement at the supports and minimum longitudinal bar diameters crossing interior or exterior joints.

On the other hand, the more demanding column constraints in critical regions for maximum spacing between restrained main bars and spacing of transverse reinforcement are formulated in accordance with Eq. (2). The same holds for the beam constraints of maximum longitudinal reinforcement volumetric ratio and spacing of transverse reinforcement in the locations of the critical regions.

5 OPTIMUM SEISMIC DESIGN ACCORDING TO FIB MODEL CODE 2010

fib MC2010 adopts a fully-fledged performance-based seismic design methodology (Fardis 2013). The code employs deformation limits, which are directly related to seismic damage, in order to verify 4 district Limit States. The Operational (OP) and Immediate Use (IU) Limit States are related to serviceability of structures, whilst the Life Safety (LS) and Collapse Prevention (CP) are related to loss of lives and structural collapse (Ultimate Limit States ULS). Limit States are checked for different levels of Seismic Hazard. Deformation limits controlling Limit States and corresponding levels of Seismic Hazard recommended by *fib* MC2010 for ordinary structures are listed in Table 1 (Fardis 2013).

The verification of Limit States entails comparisons of chord rotation demands θ_{Ed} at member ends with yield chord rotations θ_y at the same locations for the OP Limit State and twice θ_y

for the IU Limit State. Furthermore, the two ULS are checked by comparing the plastic part of chord rotation demands at member ends θ^{pl}_{Ed} with characteristic values (lower 5% percentile) of the cyclic ultimate plastic hinge rotation capacities θ^{pl}_{uk} divided by a factor of γ^*_R=1.35 for the LS Limit State and with θ^{pl}_{uk} without safety factor for the CP Limit State.

Limit State	Seismic Hazard	Deformation Limit
Operational (OP)	Frequent with 70% probability of exceedance in 50 years (70/50)	Mean value of θ_y
Immediate Use (IU)	Occasional with 40% probability of exceedance in 50 years (40/50)	Mean value of θ_y may be exceeded by a factor of 2.0
Life Safety (LS)	Rare with 10% probability of exceedance in 50 years (10/50)	Safety factor γ^*_R of 1.35 against $\theta^{pl}_{u,k}$
Collapse Prevention (CP)	Very rare with 2% probability of exceedance in 50 years (2/50)	$\theta^{pl}_{u,k}$ capacity may be reached (γ^*_R =1)

Table 1: Limit States, Seismic Hazard levels and Deformation Limits recommended by *fib* MC2010 for ordinary structures.

It is recommended (Fardis 2013) that for beams and rectangular columns with ribbed bars yield chord rotation θ_y is taken from the following equation, where φ_y is end section yield curvature, L_s the shear span of the member on the side of the end section, z is lever arm of end section, a_{scr} is a coefficient equal to 1 if shear cracking precedes flexural yielding or equal to 0 if not, h is end section height, d_{bl} and f_{yl} diameter and yield strength of longitudinal reinforcement (MPa) and f_c member concrete strength in MPa.

$$\theta_y = \frac{\varphi_y(Ls+a_{scr}\cdot z)}{3} + 0.0014\cdot\left(1+\frac{1.5h}{L_s}\right) + \frac{\varphi_y d_{bL} f_{yl}}{8\sqrt{f_c}} \tag{7}$$

Furthermore, characteristic ultimate plastic hinge rotation capacity $\theta^{pl}_{u,k}$ is derived by the respective mean value θ^{pl}_{um} divided by safety factor γ_{Rd}. When $\theta^{pl}_{u,m}$ is calculated by the following empirical relationship γ_{Rd} can be taken equal to 1.75.

$$\theta^{pl}_{um} = 0.0143\cdot 0.25^{v}\cdot f_c^{0.2}\cdot\left(\frac{\max(0.01;\omega_2)}{\max(0.01;\omega_1)}\right)^{0.3}\cdot\left(\min\left(9;\frac{L_s}{h}\right)\right)^{0.35}\cdot 25^{\left(\frac{a\rho_w f_{yw}}{f_c}\right)} \tag{8}$$

In Eq. (8), ω_1 and ω_2 are mechanical ratios of reinforcement in tension and compression zone respectively, v is normalized axial load ratio, a is confinement effectiveness factor and ρ_w and f_{yw} are volumetric ratio and yield strength of transverse reinforcement. It is noted that Eq. (8) is recommended for rectangular beams and columns with ductile steel reinforcement and without diagonal reinforcement.

In addition to chord rotation checks, brittle shear failures are checked in terms of internal shear force demands V_{Ed} and design shear force capacities V_{Rd}. V_{Rd} outside plastic hinge regions is calculated as for static loadings. Inside plastic hinge regions, *fib* MC2010 specifies a strut inclination of 45º when plastic rotation θ^{pl} exceeds $2\cdot\theta_y$ and 21.8º for elastic response (θ^{pl}=0). Interpolation is allowed for intermediate values of θ^{pl}.

The reference analysis method of *fib* MC2010 is nonlinear response history analysis with step-by-step integration of motion equations in the time domain. The finite element model applied should use realistic estimates of the effective elastic stiffness of concrete members EI_{eff}. It is recommended in MC2010 that EI_{eff} of concrete members is taken by the following relationship, where M_y represents member end section yield moment and the other parameters have been defined previously.

$$EI_{eff} = \frac{M_y L_s}{3\theta_y} \tag{9}$$

Lumped plasticity finite elements with bilinear moment-rotation hysteretic models and realistic rules for stiffness degradation during unloading and reloading may be employed to model inelastic response of reinforced concrete members.

It is worth noting that when conducting nonlinear analysis both types of seismic demands (i.e. deformations and forces) are obtained directly by the analytical solution without additional considerations for brittle modes of failure (i.e. capacity design principles).

It is also important to clarify that no additional prescriptive rules, like detailing rules set by EC8 for DCM and DCH, need to be applied when designing in accordance with MC2010 apart from the detailing rules required for designing against static loads.

In MC2010, seismic actions are represented by acceleration time-histories of the ground motions. At least seven ground motions are required to use average response values. All acceleration time histories should be scaled such that their elastic response spectrum is not lower than 90% of the target response spectrum for periods ranging between $0.2 \cdot T$ to T, where T is the fundamental period of the structure. As it will be shown later in this study, this requirement set by MC2010 can be very onerous and may lead to important increases in the structural cost. It is reminded that EC8 specifies that the mean spectrum of the set of ground motions and not all spectra shouldn't be less than 90% of the target response spectrum in the same range of periods. It is also noted that prior to designing, T is not known and cannot be estimated with accuracy because it depends on steel reinforcement which affects members' yield moments M_y and consequently effective elastic stiffness EI_{eff} as defined in Eq. (9).

6 OPTIMUM SEISMIC DESIGN CASE STUDY

In this section, application of the optimum seismic design methodologies described previously to a simple portal RC plane frame is presented. The frame is of ordinary importance, located in a region of high seismicity and rests on soil class B according to the classification of EC8. The frame is designed for 0.16g, 0.24g and 0.36g peak ground acceleration values for the 10/50 seismic hazard level in order to examine the influence of the level of seismicity on the optimum seismic design solutions. The elastic (target) response spectrum with 5% damping of EC8 determined for these specifications and 0.24g peak ground acceleration is shown in Fig. 2.

Peak ground accelerations for the other seismic hazard levels of MC2010 objectives are calculated by multiplying by the importance factor γ_I given by the following equation proposed in EC8-Part 1, where P_L is the target probability of exceedance in 50 years and P_{LR} is the reference probability of exceedance in 50 years (=10%).

$$\gamma_I = \left(\frac{P_L}{P_{LR}}\right)^{-1/3} \tag{10}$$

The portal frame is designed following the provisions of EC8 for all three ductility classes (i.e. DCL, DCM and DCH) and in accordance with MC2010. In the latter case and in order to evaluate the influence of ground motions selection two different cases are examined. In the first case, designated as THA, the frames are designed for a set of 7 scaled ground motion records satisfying EC8-Part 1 recommendations as described in the previous section. In the second case, designated as THB, the frames are designed for a set of 7 scaled ground motion records satisfying following MC2010 specifications.

Figure 2a presents the scaled and mean elastic spectra with 5% damping of the set of 7 ground motions selected and scaled following EC8 provisions. Selection and scaling was achieved by employing computer program REXEL (Iervolino *et al*. 2009). Because the fundamental period of the structures is unknown prior to their design it was decided to match the mean and target spectrum for periods between 0.1s and 4s in order to capture most possible solutions. The selected ground motion records are all recorded on soil type B and have magnitude M_w>5.5. It is evident that the mean spectrum follows very closely the target spectrum.

No computer tools exist for selecting record sets according to MC2010 guidelines. To serve this goal, in this study, a simplified procedure is applied. All records of the European Strong Motion Database (Ambraseys *et al*. 2004) on soil type B with M_w>5.5 are scaled so that their scaled 5% damping spectra are not less than 90% of the target spectrum in periods range between 0.1s and 4s. The scaled spectra are later ranked in accordance with their "goodness-of-fit" to the target spectrum as quantified by the normalized root-mean-square-error (Katsanos and Sextos 2013). The first 7 ground motions comprise the set of records used herein. Figure 2b presents the scaled and mean elastic spectra with 5% damping of the set of 7 ground motions selected and scaled following MC2010. It can be seen that the mean spectrum importantly exceeds the target spectrum leading to serious overestimation of seismic demands. Obviously, more advanced selection methods can produce closer mean and target spectra. This set is used herein simply to illustrate the importance of ground motion selection on the optimum design cost of RC frames.

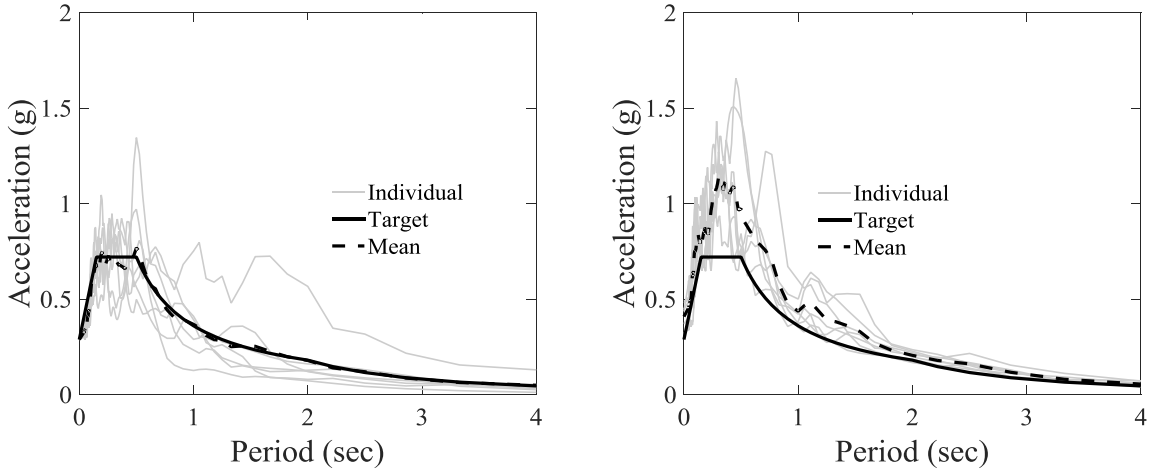


Figure 2: Elastic spectra with 5% damping for ground motion sets selected and scaled in accordance with a) EC8, b) MC2010

For the optimum designs, it is assumed that section dimensions $\boldsymbol{h_c}$, $\boldsymbol{b_c}$, $\boldsymbol{h_b}$, $\boldsymbol{b_b}$ take values from the following discrete set: (0.25m; 0.30m; 0.40m; … ; 1.0m). Furthermore, longitudinal bars $\boldsymbol{d_{bc}}$, $\boldsymbol{d_{bb}}$, and $\boldsymbol{d_{bt}}$ are defined in the following discrete values set: (12mm; 16mm; 20mm; 25mm). Transversal bars $\boldsymbol{d_{bwc}}$, and $\boldsymbol{d_{bwb}}$ take values from: (8mm; 10mm; 12mm). Transverse reinforcement spacing $\boldsymbol{s_c}$ and/or $\boldsymbol{s_b}$ may take the following values: (0.1m; 0.15m; 0.20m; 0.25m; 0.30m).

Finally, numbers of main bars n_c, n_{tb}, n_{bb} and legs of shear reinforcement n_{wc} and n_{wb} may take any integer value greater than one.

The portal reinforced concrete frame optimally designed herein is shown in Figure 3. The span of the frame is 4m and the height 3m. Concrete C25/30 and reinforcing steel B500C in accordance with EC2 specifications are used. Concrete cover is assumed to be 30mm. Vertical symmetric concentrated loads are applied at the column locations equal to 120.0kN for the permanent and 80.0kN for the live loading. Storey mass for the seismic combination is 29.4t.

The frame consists of two columns C1 and C2 and one beam B1. Due to symmetry, it is assumed that C1 and C2 have exactly the same sections and reinforcement. Concrete members are represented by three sections. Two at the ends to represent critical regions and one to represent the rest part of the element. Due to construction reasons, it is assumed that all sections have the same longitudinal reinforcement. Furthermore, it is assumed because of symmetry that member end sections have the same transverse steel reinforcement. Thus, the same sections are assumed for member ends. Intermediate and end sections differ in the spacing of transverse reinforcement in order to account for the additional requirements in the member critical regions. In total, 18 (8 for columns and 10 for the beam) independent design variables are used in this problem.

The results presented in the following were obtained by running GA with populations of 75 individuals. Iterations were terminated when the mean relative variation of the best fitness value was negligible for 100 generations. MATLAB-R2015a default options were used for GA operations. Furthermore, a significant number of different GA runs for each design solution were conducted and the minimum cost obtained is reported herein.

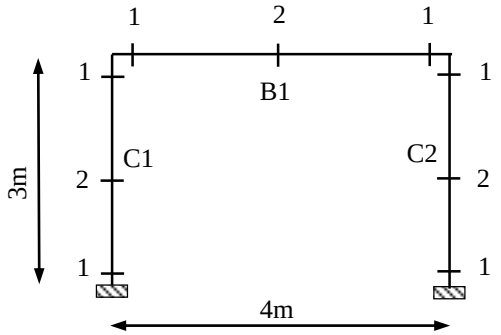


Figure 3: Example portal frame

Figure 4 presents optimization histories of the designs obtained by MC2010 methodology for the THA ground motion set and the three design peak ground accelerations. It can be seen that, as expected, optimum cost increases as design accelerations increase.

Figure 5 compares optimum costs obtained by all seismic design methodologies for the three design peak ground accelerations of the 10/50 seismic hazard level. It can be seen that in all cases costs increase as design accelerations increase. Designs according to EC8 DCL and DCM yield similar costs for all design PGA values. On the other, DCH yields significantly increased costs. This occurred because of the enhanced detailing rules of this ductility class and the discrete design variable sets assumed in this study. This observation shows the influence of detailing requirements on the final costs of reinforced concrete structures.

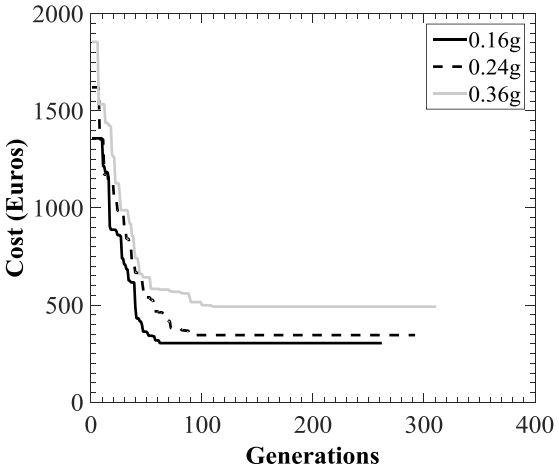


Figure 4: Optimization histories of designs obtained by MC2010 methodology for THA ground motion set and three different design PGAs for the 10/50 seismic hazard level

Furthermore, the direct comparison of optimum costs obtained by MC2010 methodology for the THA and THB ground motion sets shows the importance of the applied accelerograms set. For 0.16g PGA both solutions yield same optimum costs. This is because the design in this case is controlled by minimum detailing requirements. However, for larger seismicity levels the increase in cost by selecting a ground motion data set in accordance with MC2010 provisions is very important.

It is also evident that designs obtained by the MC2010 for both ground motion sets (THA and THB) drive to significantly reduced design costs for the low 0.16g and moderate 0.24g design accelerations. The MC2010 design with THA motion set yields slightly smaller cost than the EC8 designs according to DCL and DCM approaches for 0.36g. However, the same design methodology with the THB motion set drives to significantly greater design costs than all EC8 designs obtained for 0.36g.

Figure 6 presents MC2010 checks of rotation demand constraints for the THA ground motion set and for all Limit States performed for the optimum solutions obtained by all design methodologies for 0.36g. Column sections locations are defined by the column member number (e.g. C1) and a letter designating the location of the section in the member (i.e. B=bottom and T=top). Similarly, beam sections are defined by the beam member number (e.g. B1) and a letter designating the location of the section in the member (i.e. L=left and R=right). Limit States are written with the acronyms shown in Table 1.

It can be concluded that all design solutions perform rather well. DCM and DCH designs do not satisfy beam rotation constraints for the OP Limit State. It is recalled that EC8 does not have any provisions for the OP Limit State. Furthermore, it can be seen that the MC2010 design for THA motion set marginally satisfies beam and column rotation constraints at the OP Limit State and column rotation constraints at the CP Limit State. This shows that these where the controlling (active) constraints of this design. It is also evident that MC2010 design for THB motion satisfies all constraints with a high level of conservatism.

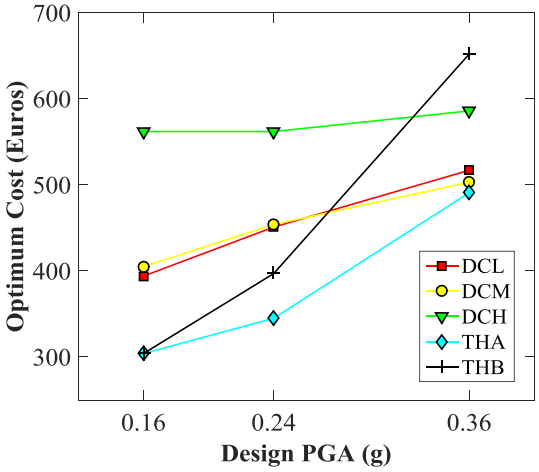


Figure 5: Optimum costs obtained by different design methodologies and design PGAs for the 10/50 seismic hazard level.

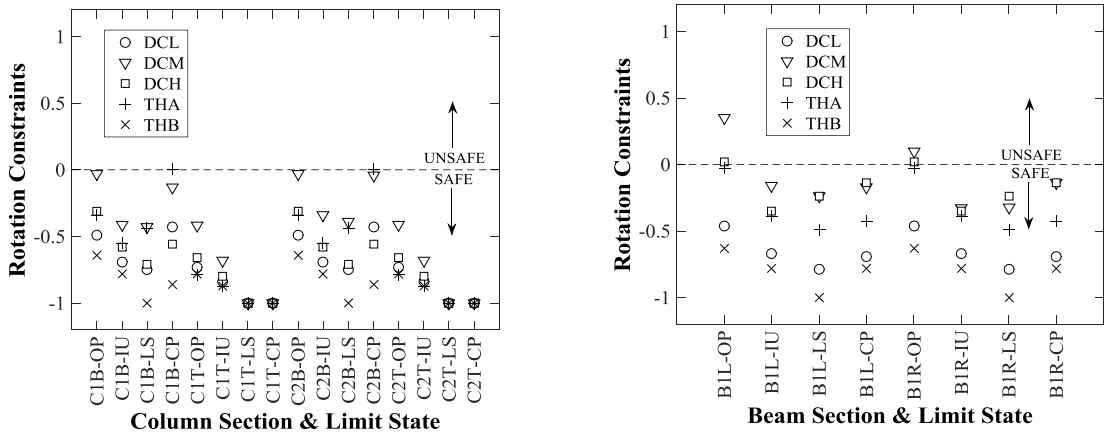


Figure 6: MC2010 rotation constraints of beam and column section optimum solutions obtained by different design methodologies for 0.36g 10/50 seismic hazard PGA.

7 CONCLUSIONS

- The new *fib* Model Code 2010 (MC2010) includes a fully-fledged performance-based seismic design and assessment methodology for various levels of seismic hazard. MC2010 will serve as a basis for future codes for concrete structures.
- The reference analysis method of *fib* MC2010 is nonlinear response history analysis with step-by-step integration of motion equations in the time domain. Due to the complexity of this approach, genetic algorithms are well suited to this design methodology since they don't require use of gradients of cost or constraints functions. Furthermore, they are able to identify global optima as opposed to local optimum solutions.
- This study presents optimum seismic design solutions of a portal reinforced concrete frame obtained by MC2010 and compares them with optimum designs following EC8 (CEN 2004) guidelines for all three ductility classes. The frame is designed for three different peak ground accelerations (0.16g, 0.24g and 0.36g) for the 10/50 seismic hazard level.

- It is found that the construction cost increases with the level of peak ground acceleration for all seismic design methodologies.
- The MC2010 seismic design methodology may lead to serious cost savings for low seismicity levels. However, for high design peak ground accelerations MC2010 and EC8 yield similar optimum costs.
- The selection of the ground motion records set plays an important role on the optimum design costs obtained by the MC2010 approach. If the MC2010 prescriptions are used the design costs increase importantly with respect to the selection of a set of ground motions following EC8 guidelines.
- EC8 optimum design solutions do not satisfy rotation demand constraints set by MC2010 for the OP limit state. This can be attributed to the fact that no explicit provisions exist in EC8 to address this limit state.

REFERENCES

[1] *fib, Model Code 2010*. *Bulletins Nos. 65/66*, Federation Internationale du Beton, Lausanne, 2012.

[2] M.N. Fardis, Performance- and displacemenent-based seismic design and assessment of concrete structures in *fib* Model Code 2010. *Structural Concrete*, **14**, 215-229, 2013.

[3] M. Fragiadakis, N.D. Lagaros, An overview to structural seismic design optimization frameworks. *Computers and Structures*, **89**, 1155-1165, 2011.

[4] K.C. Sarma, H. Adeli, Cost optimization of concrete structures. *Journal of Structural Engineering ASCE*, **124**, 570-578, 1998.

[5] S. Ganzerli, C.P. Pantelides, L.D. Reaveley, Performance-based design using structural optimization. *Earthquake Engineering and Structural Dynamics*, **29**, 1677-1690, 2000.

[6] C.M. Chan, X.K. Zou, Elastic and inelastic drift performance optimization for reinforced concrete buildings under earthquake loads. *Earthquake Engineering and Structural Dynamics*, **33**, 929–950, 2004.

[7] N.D. Lagaros, M. Papadrakakis, Seismic design of RC structures: A critical assessment in the framework of multi-objective optimization. *Earthquake Engineering and Structural Dynamics*, **36**, 1623-1639, 2007.

[8] M. Fragiadakis, M. Papadrakakis, Performance-based optimum seismic design of reinforced concrete structures. *Earthquake Engineering and Structural Dynamic*, **37**, 825-844, 2008.

[9] B. Gencturk, Life-cycle cost assessment of RC and ECC frames using structural optimization. *Earthquake Engineering and Structural Dynamics*, **42**, 61-79, 2013.

[10] CEN, *Eurocode 8: Design of structures for earthquake resistance. Part 1: General rules, seismic actions and rules for buildings*. European Standard EN 1998-1, Brussels, Belgium, 2004.

[11] J. Holland, *Adaptation in natural and artificial systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.

[12] MathWorks, *MATLAB R2015a – Global Optimization Toolbox*. The MathWorks Inc., Natick, MA, USA, 2000.

[13] CEN, *Eurocode 2: Design of concrete structures. Part 1-1: General rules and rules for buildings*. European Standard EN 1992-1-1, Brussels, Belgium, 2004.

[14] I. Iervolino, C. Galasso, E. Cosenza, REXEL: computer aided record selection for code-based seismic structural analysis. *Bulletin of Earthquake Engineering*, 8, 339-362, 2009.

[15] N.N. Ambraseys, J. Douglas, D. Rinaldis, *et al*., *Dissemination of European Strong Motion Data. Vol. 2*, Engineering and Physical Sciences Research Council, UK, 2004.

[16] E.I. Katsanos, A.G. Sextos, An integrated software environment for structure-specific earthquake ground motion selection. *Advances in Engineering Software*, 58, 70-85, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CALIBRATION OF THE NUMERICAL MODEL OF A STAND IN DRAGÃO STADIUM BASED ON GENETIC ALGORITMS

Jorge Leite[1], **Diogo Ribeiro***[2], **Hugo Marques**[3], **Rui Calçada**[4]

[1] Polytechnic of Porto, School of Engineering
Rua Dr. António Bernardino de Almeida, 431 | 4249-015 Porto
1130177@isep.ipp.pt

[2] CONSTRUCT-LESE
Polytechnic of Porto, School of Engineering
Rua Dr. António Bernardino de Almeida, 431 | 4249-015 Porto
drr@isep.ipp.pt

[3] GEG
Gabinete de Estruturas e Geotecnia
Rua Justino Teixeira, Centro de Campanhã, Bloco A, Piso 3, Esc.307 | 4300-273 Porto
hmarques@geg.pt

[4] CONSTRUCT-LESE
University of Porto, Faculty of Engineering
Rua Dr. Roberto Frias, s/n | 4200-465 Porto
ruiabc@fe.up.pt

Keywords: stand, seating deck units, numerical modelling, dynamic test, model calibration, genetic algorithm.

Abstract. *This paper is focused on the experimental calibration of the numerical model of a stand in Dragão stadium based in genetic algorithms. A finite element numerical model of a group of seating deck units including the connections between them and the connections with the stands, was developed. Also a dynamic test was performed in the South stand in order to characterize the local dynamic properties of a group of seating deck units, particularly its natural frequencies, modal configurations and damping coefficients. The experimental calibration of the numerical model was performed using an iterative method based on a genetic algorithm. The stability of a significant number of parameters, considering different initial populations, proved the robustness of the adopted algorithm in the scope of the optimization of the numerical model. Also the calibration results demonstrate a very good agreement between numerical and experimental modal responses and a significant improvement of the numerical model before calibration.*

1 INTRODUCTION

In recent years, the organization of sport events of international scale, especially in European countries, has launched new challenges in terms of rehabilitation and reconstruction of the existing football stadiums and in some cases, the construction of new stadiums.

The need to reduce the obstacles in front of the public forced the stadiums to be built using larger span and slender structural elements. This situation took the structures to be more susceptible to resonance phenomena, especially due to the proximity between the frequencies associated with the synchronized movements of the public with the natural frequencies of vibration of the structure [1].

The new functional requirements of sports facilities, especially those associated with the safety and comfort of spectators, have led to a growing interest of researchers for the study of dynamic effects induced by the public in this type of structures [1]. These studies usually involve performing dynamic tests and the development of numerical finite element models to support decision and typically calibrated based on experimental data [1].

In this context it should be noted the experimental studies carried out by Littler et al. [2] at Twickenham Stadium (United Kingdom), by Pavic and Reynolds [3] in the stadium of Bradford (United Kingdom) and Cigada et al. [4] at Giuseppe Meazza Stadium (Italy), and involving measuring the vertical acceleration of the stands, with and without spectators, using a permanent monitoring system. The results of these studies have revealed the importance of the presence of spectators for the dynamic properties of the stands, in particular the values of vibration frequencies and damping coefficients.

Most experimental studies have mainly focused on the global analysis of the dynamic behavior of the stands, neglecting the characterization of local vibration modes associated with seating deck units. Of the few studies identified it is important to point the work of Marques et al. [5] and Marovic et al. [6], which despite focusing on the evaluation of the dynamic response values of the seating deck units during sport events, they did not include the identification of its local modal parameters.

From the works developed by Lima Avila and Doz [7], Saudi et al. [8], Marques [5] and others, experimental modal information is used for calibration of numerical finite element models of the stands. The calibrated models have proved to be especially useful in the study of reinforcement interventions [8], in the design of vibration control systems [5], identification of structural damage [7] and for simulation of load scenarios different from those normally acting with the structure in normal operation.

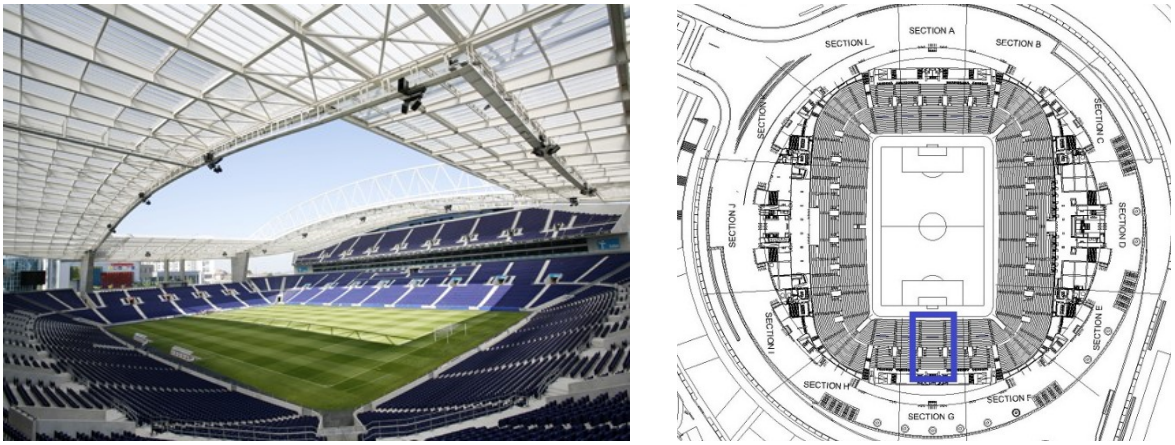
This article focuses on the experimental calibration of a numerical model of seating deck units of Dragão Stadium. For this purpose, it is developed a numerical finite element model that aims to study the dynamic behavior of a seating deck unit and considers the influence of its connection with neighboring seating deck units. It is also performed the experimental characterization of a series of seating deck units of the south stand of Dragão stadium, in order to identify its local dynamic properties, particularly natural frequencies, mode shapes and damping coefficients. Finally, automatic calibration of the numerical model was performed using an iterative method based on genetic algorithms, which uses a computational interface between three programs (Autodesk Robot, Excel and Matlab), allowing the inclusion of a conventional structural calculation program in an optimization flow through specific API routines.

2 DRAGÃO STADIUM

Dragão stadium is located in the city of Porto (Portugal) and was built on the occasion of the European Football Championship in 2004. The stadium has a capacity of 50092 spectators and

comprises four stands: south, west, north and east (Figure **1**a). The north and south stands consist of a single level, while the east and west stands consist of two levels, one lower and one upper.

Each stand is divided into structurally independent bodies separated by joints. The structure of each body of the stand is formed by a set of frames, spaced by 8.1 m, at the interior perimeter, and 10.5 m on the perimeter of the periphery. The seating deck units are supported by the raker beams that integrate these frames.



a) b)

Figure 1: Dragão stadium: a) global view; b) stands plan.

The seating deck units of the lower stands are T-shaped prefabricated elements in reinforced concrete (Figure **2**). The seating deck units are formed by a horizontal plate of thickness equal to 0.10 m and width equal to 0.80 m, which supports on a vertical rib with 0.15 m thickness and variable height between 0.48 m and 0.63 m, depending on the position of the seating deck unit in the stand frames.

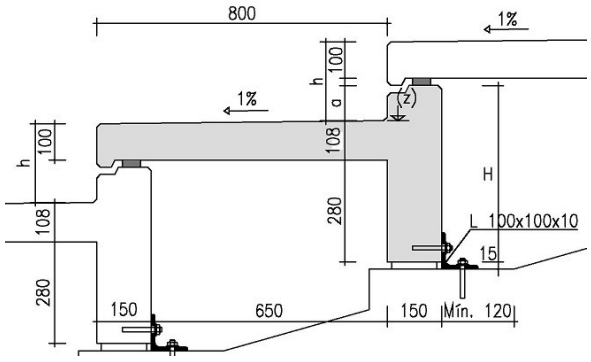


Figure 2: Seating deck units of Dragão stadium (lower stands).

In radial direction, the connection between the seating deck units is guaranteed, in an alternating solution, by rigid connections by means of 3 metal connectors type M20, and flexible connections through EPDM rubber pads. Towards its perimeter each seating deck unit is separated from neighbors by joints.

Each seating deck unit is supported in raker beams by means of a clamping device consisting of a steel angle L100×100×10 which is connected to the vertical seating deck unit rib and to the beam frame by means of two steel connectors type M12. Between the base of the rib of the seating deck unit and the stand beam there is a neoprene pad with 15 mm thick.

3 NUMERICAL MODEL

3.1 Description

The numerical finite element model of the seating deck units of the stands of Dragão stadium was developed using the commercial software Autodesk Robot Structural Analysis [9] (Figure 3a).

The numerical model comprises a set of 10 consecutive seating deck units, located on the south stand inside the area identified in Figure 1b, and aims the characterization of the local dynamic behavior of seat deck unit of row 14 considering the influence of its connection with neighboring seating deck units.

The horizontal plates and ribs of seating deck units in addition to the steel angle of the supports, were modeled by shell finite elements, while the remaining elements, in particular the steel connectors and the elastic supports between seating deck units and between the seating deck units and the stand frame, were modeled using beam finite elements. In Figure 3b is illustrated in detail the modeling of the links between seating deck units and between the seating deck units and the stand frame.

In the numerical model all the structural elements that constitute the seating deck units, were reproduced, in accordance with the design information. For simplicity all seating deck units were modeled with a span equal to 8.22 m corresponding to the span of the seating deck unit of row 14.

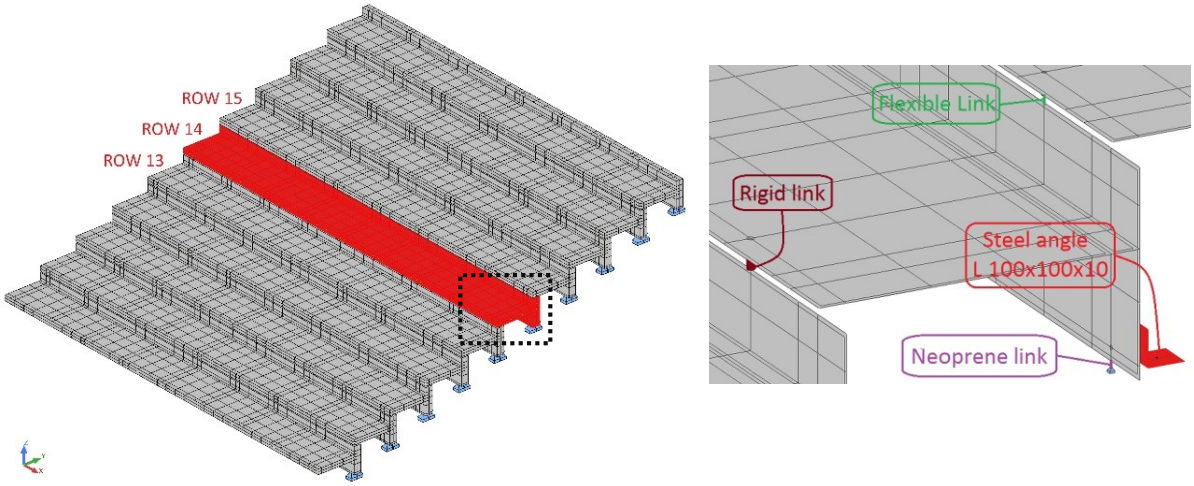


Figure 3: FE numerical model: a) global view, b) detail of modeling of the links between seating deck units and between the seating deck units and the stand

3.2 Geometrical and mechanical properties

Table 1 presents the most relevant geometric and mechanical parameters adopted in the numerical model of seating deck units, including its name, the value adopted and respective units. In addition, the upper and lower limits that will be used later in the calibration phase of the numerical model are also indicated.

Table 1: Geometrical and mechanical parameters of the numerical model of the seating deck units.

Parameters			Unit	Adopted value	Limits	
					Lower	Upper
E_c	Elasticity modulus of concrete		GPa	32.2	28.0	36.0
ρ_c	Concrete density		kN/m^3	25	--	--
E_{M20}	Elasticity modulus of steel in M20 connectors between seating deck units		GPa	210	150	250
$E_{S,0}$	Elasticity modulus of steel connectors between seating deck units and stand beams	Other rows	GPa	210	150	250
$E_{S,13L}$		Row 13 (left)				
$E_{S,13R}$		Row 13 (right)				
$E_{S,14L}$		Row 14 (left)				
$E_{S,14R}$		Row 14 (right)				
$E_{S,15L}$		Row 15 (left)				
$E_{S,15R}$		Row 15 (right)				
E_{EPDM}	Elasticity modulus of EPDM rubber pads between seating deck units		MPa	50	0	150
$E_{N,0}$	Elasticity modulus of neoprene layer between seating deck units and stand beams	Other rows	MPa	50	0	150
$E_{N,13L}$		Row 13 (left)				
$E_{N,13R}$		Row 13 (right)				
$E_{N,14L}$		Row 14 (left)				
$E_{N,14R}$		Row 14 (right)				
$E_{N,15L}$		Row 15 (left)				
$E_{N,15R}$		Row 15 (right)				

The value adopted for the elasticity modulus of concrete was defined on the basis of the results of a concrete ultrasonic test, since there was no design information about its mechanical properties. The results of this test allowed to estimate an average value of the dynamic elasticity modulus of concrete equal to 32.2 GPa with a coefficient of variation of 3.0%. The variation in this parameter limits were, however, extended to meet the fact that the ultrasonic test has been performed on a limited number of seating deck units due to accessibility constraints for the remaining seating deck units.

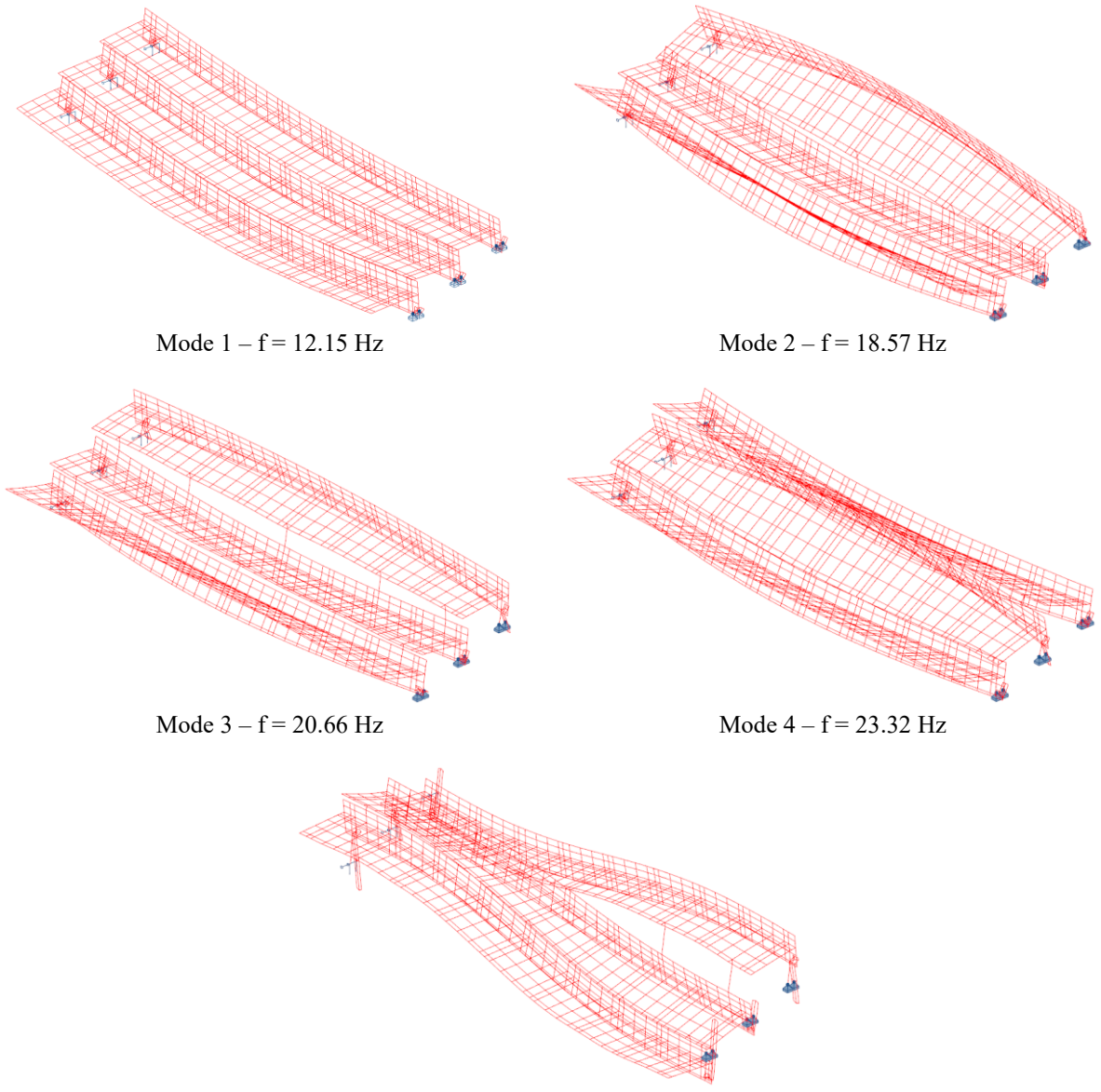
The definition of variation limits of the elasticity modulus of steel of the connectors between seating deck units and between the seating deck units and the stand beams allowed to meet some singularities observed *in situ*, in particular the use of a larger number of connectors than those provided in the design stage, the use of connectors of different diameter of the specified in design, and the loss of connection stiffness due to poor filling of the holes with a chemical component.

The definition of the limits of variation of the elasticity modulus of EPDM rubber pads and neoprene, in the connection between seating deck units and between the seating deck units and the stand beams, respectively, also allowed to meet situations observed *in situ*, namely, the deterioration or inexistence of supports, and the use of elastic supports in different materials of the specified design.

In the case of the elasticity modulus of steel of the connectors between the seating deck units and the stand beams, and the elasticity modulus of neoprene, distinct characteristics are defined for the supports located to the left and right on seating deck units 13, 14 and 15, and other seating deck units.

3.3 Numerical modal parameters

Figure 4 shows the values of the vibration frequencies of the main local modes of the seating deck units and the corresponding mode shapes obtained from the numerical model developed based on the adopted values of parameters listed in Table 1. Modes 1 and 5 essentially involve bending movements of the seating deck units. Modes 2, 3 and 4 mainly involve torsional movements of the seating deck units. In the mode shapes were only represented the seating deck units of rows 13, 14 and 15.



Mode 1 – f = 12.15 Hz

Mode 2 – f = 18.57 Hz

Mode 3 – f = 20.66 Hz

Mode 4 – f = 23.32 Hz

Mode 5 – f = 30.71 Hz

Figure 4: Numerical modal parameters.

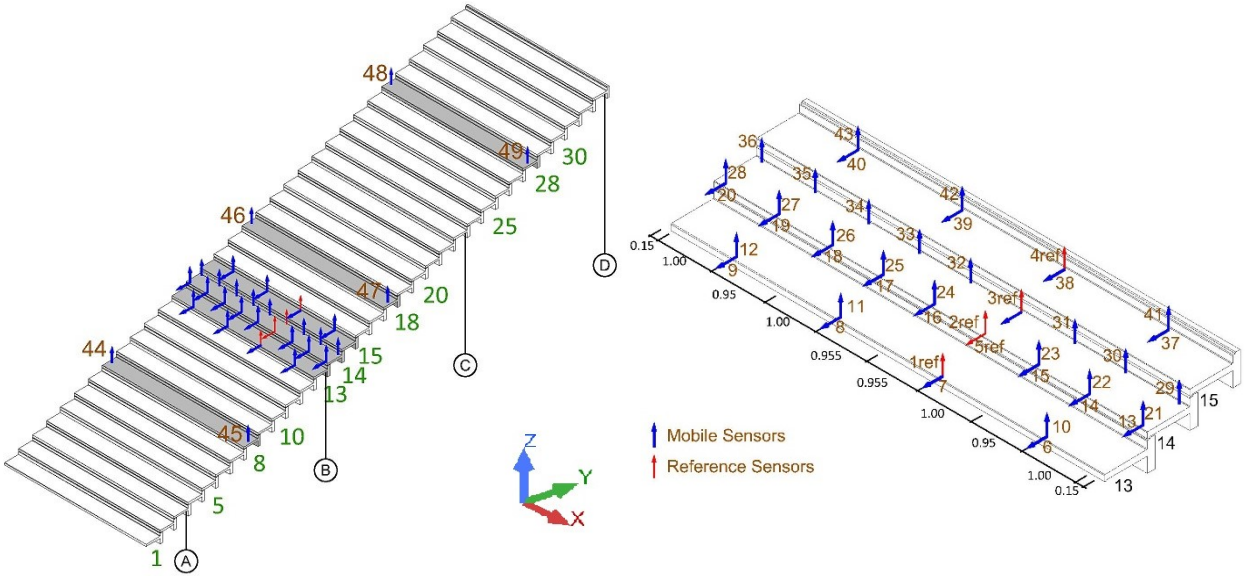
4 DYNAMIC TEST

4.1 Description

The dynamic test aimed to identify the modal properties of the seating deck units, especially its natural frequencies and local vibration modes and the respective damping coefficients.

The test was conducted using a technique with fixed reference points and mobile measurement points, involving the use of 16 high sensitivity piezoelectric accelerometers, PCB 393B12 model. The accelerations were measured in the vertical (z) and radial (y) directions in a total of 49 measurement points located in the seating deck units of rows 8, 13, 14, 15, 18 and 28 of south stand.

Figure 5a illustrates the position of the accelerometers in the seating deck units. The ends of the seating deck units of rows 8, 14, 18 and 28 were instrumented in the vertical direction, in order to identify eventually global movements associated with the stand frames. In turn, the seating deck units of rows 13, 14 and 15 were instrumented in the vertical and radial directions in order to characterize the local movements of the seating deck units.



a) b)

c) d)

Figure 5: Dynamic test: a) measurement points (global view); b) measurement points of rows 13, 14 and 15; c) accelerometers; d) external excitation.

The reference transducers were located on the seating deck units of rows 13, 14 and 15 in positions 1, 2, 3, 4 and 5 (Figure 5b). Data acquisition was performed using NI cDAQ-9172 system, using four NI 9234 modules for IEPE type accelerometers. The time series were acquired with a duration of 5 minutes, with a sampling frequency of 2048 Hz, further decimated

at a frequency equal to 256 Hz. The connection of accelerometers to seating deck units was performed by means of metallic plates and angles bonded to the concrete surface (Figure 5c).

The test took place under an external excitation ensured by the action of a group of individuals who performed jumps and walked randomly over time (Figure 5d).

4.2 Modal parameters identification

The identification of modal parameters was performed by the application of the Enhanced version of Frequency Domain Decomposition method (EFDD) using the commercial software ARTeMIS [10].

Figure 6 show the curves of the average and normalized singular values of spectral density matrices of all experimental setups, obtained by the EFDD method. Five local vibration modes associated with the seating deck units in correspondence with the 5 peaks indicated in the curve of the first singular value, were identified.

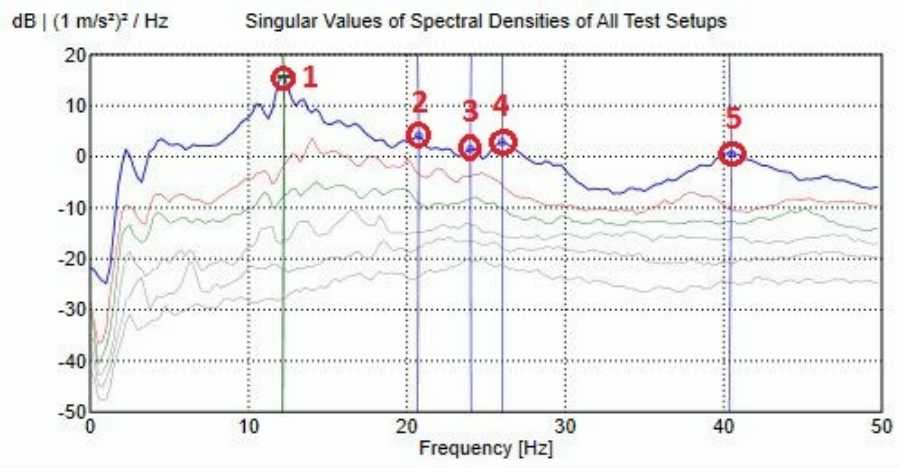
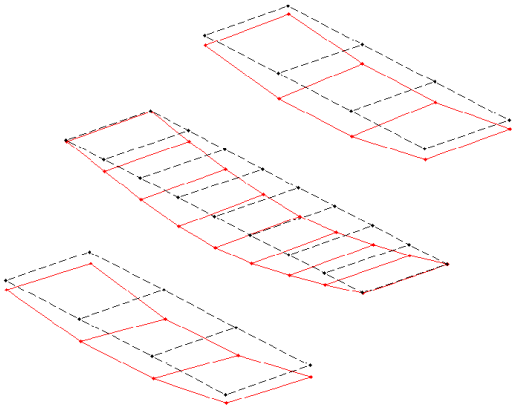
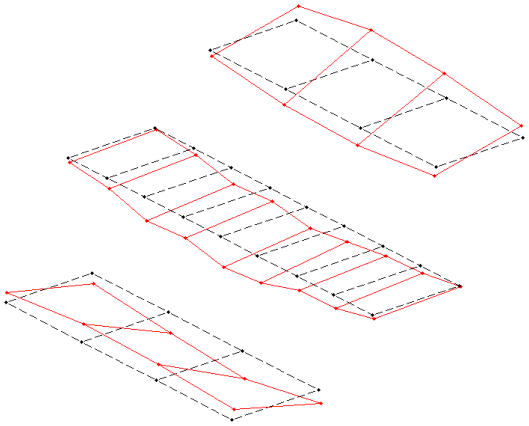


Figure 6: EFDD method: average and normalized singular values of spectral density matrices.

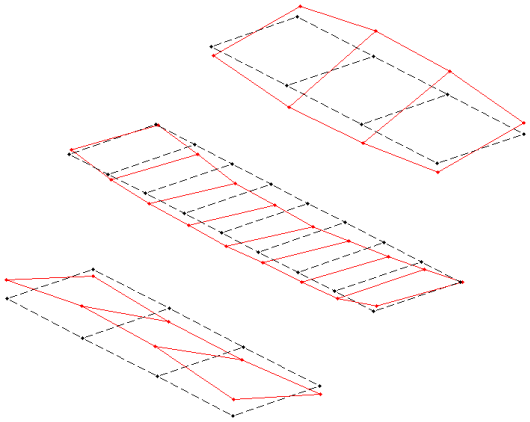
In Figure 7 are shown the mean values of natural frequencies and corresponding local modes of vibration of the seating deck units. The analysis of the modal configurations allows to identify movements associated with bending and twisting of the seating deck units with very good definition. In these modes of vibration the movements of the stand frames are negligible. Modes 2 and 3 are distinguished by the fact that the movement of the seating deck unit of row 14 involves transverse and vertical bending, respectively. The values of the damping coefficients vary between 1.11% and 2.53%.



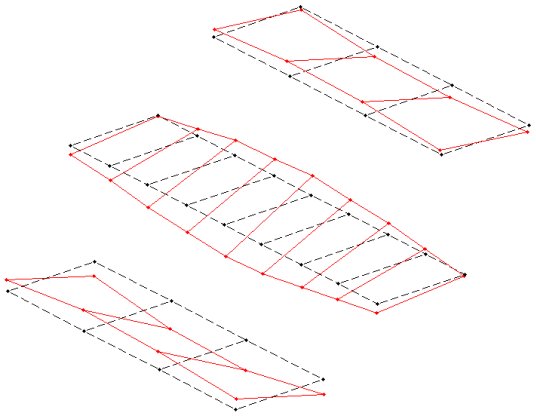
Mode 1 – f = 12.15 Hz
ξ = 2.53 %



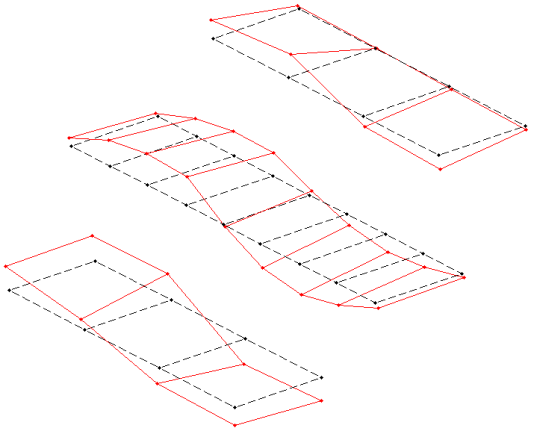
Mode 2 – f = 20.68 Hz
ξ = 1.36 %



Mode 3 – f = 24.06 Hz
ξ = 1.26 %



Mode 4 – f = 26.01 Hz
ξ = 1.26 %



Mode 5 – f = 40.35 Hz
ξ = 1.11 %

Figure 7: Experimental modal parameters.

5 CALIBRATION

The calibration of the numerical model of the seating deck units was based on the results of the dynamic test and involved performing a sensitivity analysis and an optimization based on genetic algorithms.

5.1 Methodology

Figure 8 presents a flowchart illustrating the iterative process of the numerical model calibration based on a genetic algorithm and involving the use of three softwares: Autodesk Robot [9], Excel [11] and Matlab [12]. A detailed explanation of the proposed methodology is presented in reference [13].

In the present work some improvements to the tool developed by Ribeiro et al. [13] which allowed to improve its versatility and computational efficiency, were introduced, in particular: i) the possibility of performing the numerical modeling of the structure in a conventional automatic structural calculation software, in this case Autodesk Robot; ii) the implementation of API routines from Robot software, in Excel environment, that enable an efficient interconnection of the numerical model to the optimization algorithm, namely for the extraction of numerical modal parameters, from Robot to Matlab, and for the introduction of new set of values of numerical parameters, from Matlab to Robot, and iii) the application of genetic algorithms based on existing routines of Matlab software.

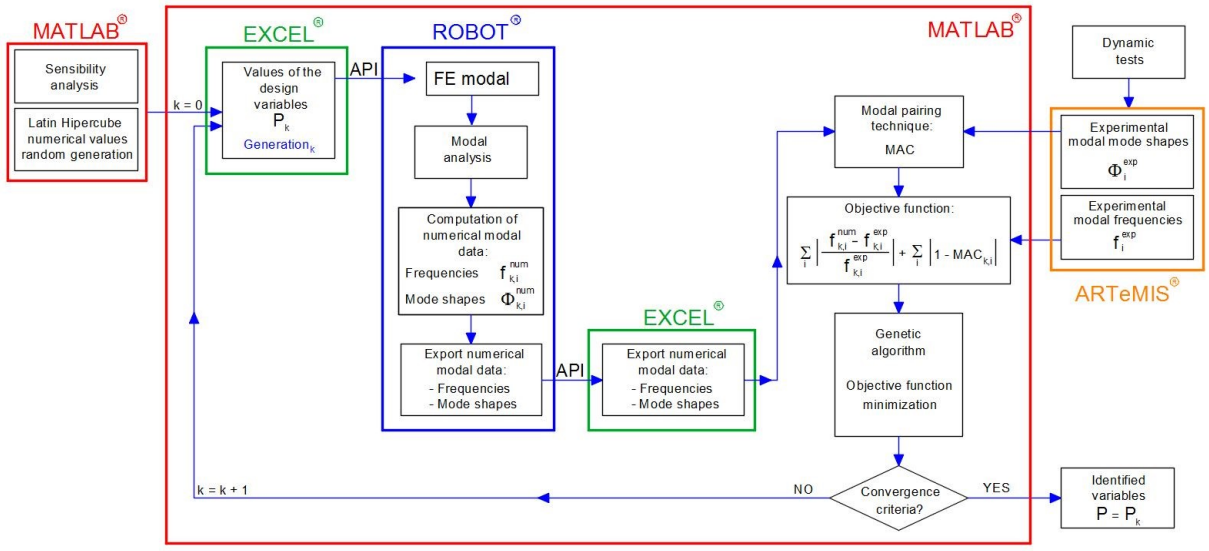


Figure 8: Flowchart of the calibration methodology of the numerical model.

5.2 Sensitivity analysis

The sensitivity analysis aims identifying the parameters that most influence the responses of seating deck units, particularly the natural frequencies and vibration modes, and which will be later included in the optimization phase.

Figure 9 shows the results of the sensitivity analysis using a Spearman correlation coefficients matrix [14]. The sensitivity analysis was performed using a stochastic sampling technique based on 1000 samples generated by Latin Hypercube method. The correlation coefficients situated in the range [-0.20; +0.20] were excluded from the graphical representation.

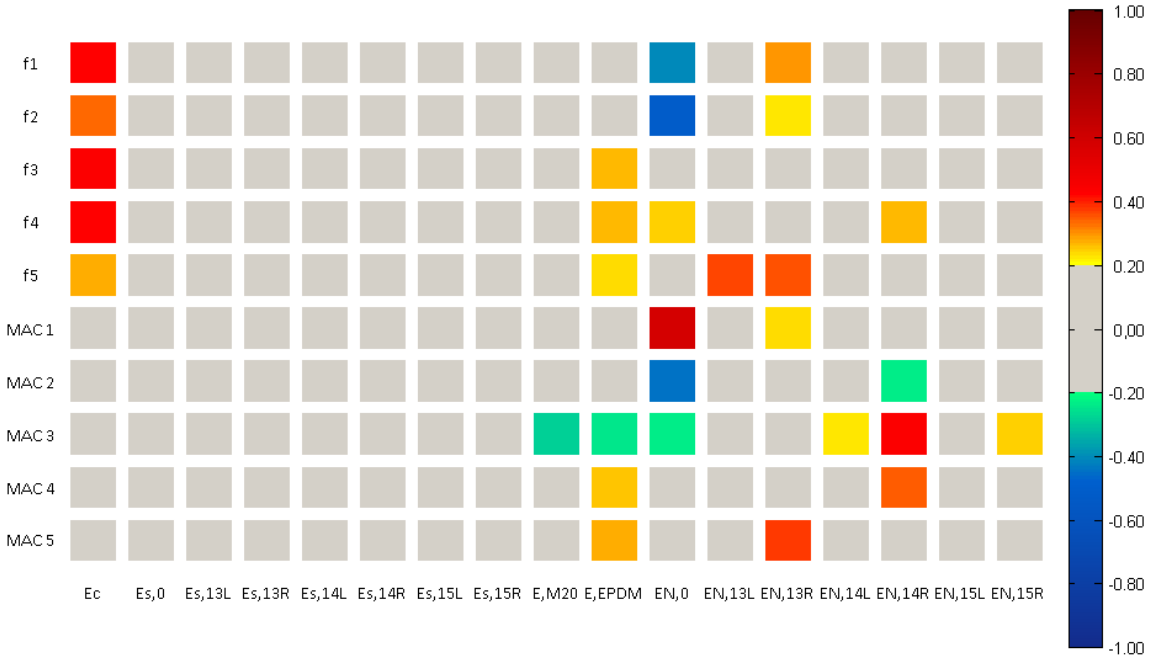


Figure 9: Spearman correlation matrix.

The correlation matrix shows that the elasticity modulus of concrete (E_c), elasticity modulus of steel in M20 connectors between seating deck units (E_{M20}), elasticity modulus of EPDM rubber pads between seating deck units (E_{EPDM}) and elasticity modulus of neoprene (E_N) of almost all the seating rows, are the parameters that most influence the modal responses.

5.3 Optimization

The optimization phase aimed to obtain the values of the numerical parameters that minimize the differences between the numerical and experimental modal parameters, and involves the definition of an objective function and the application of an optimization technique based on a genetic algorithm.

The objective function (f) comprises two terms, one related with the residuals of the vibration frequencies and other related with the residuals of MAC values:

$$f = a \sum_{i=1}^{5} \frac{\left| f_i^{exp} - f_i^{num} \right|}{f_i^{exp}} + b \sum_{i=1}^{5} \left| MAC\left(\varnothing_i^{exp}, \varnothing_i^{num} \right) - 1 \right| \quad (1)$$

where f_i^{exp} and f_i^{num} are the experimental and numerical frequencies for mode i, ϕ_i^{exp} and ϕ_i^{num} are the vectors containing the experimental and numerical modal information regarding mode i, and a and b are weighting factors of the terms of the objective function assumed, in this situation, equal to 1.0.

The optimization model involves 9 numerical parameters and 10 modal results. The genetic algorithm was based on an initial population of 30 individuals and 100 generations, in a total of 3000 individuals. The initial population was randomly generated by Latin Hypercube method. In this algorithm was defined a number of elites equal to 1, a replacement rate equal to 5% and a crossing rate of 50%.

In Figure 10 are presented the ratios of the values of each numerical parameter in relation to the limits given in Table 1 for the independent optimization runs GA1 to GA4. A ratio of 0% means that parameter coincides with the lower limit and a ratio of 100% means that coincides with the upper limit. The global parameters and parameters related to the interfaces between

the seating deck units are shown in Figure 10a, with the correspondent values indicated in brackets. The parameters of the interfaces of the seating deck units to stand beams are shown in Figure 10b.

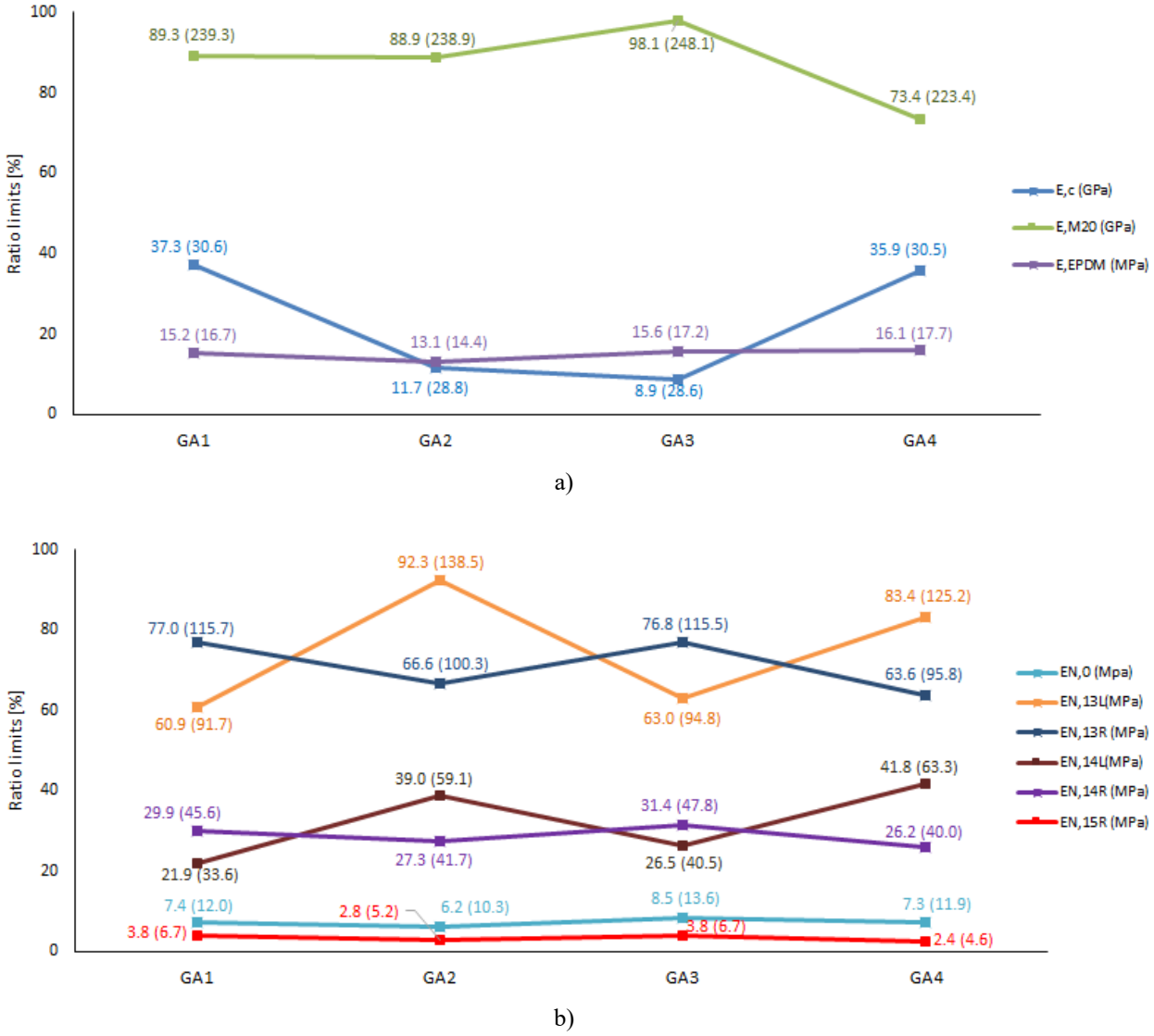


Figure 10: Values of numerical parameters obtained for optimization runs GA1 to GA4: a) global parameters and parameters of the interfaces between the seating deck units; b) parameters of the interfaces between seating deck units and stand beams.

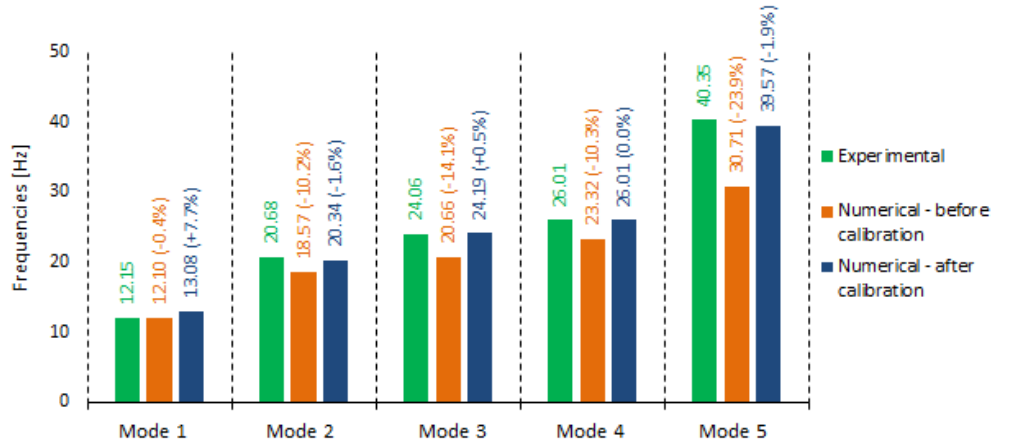
The results help identifying as that the most sensitive parameter, the elasticity modulus of neoprene layer between seating deck units and stand beams of the other rows (E_N), which is the one exhibiting the lower variations, close to 3%. The elasticity modulus of EPDM rubber pads between seating deck units (E_{EPDM}) also has lower variations, in the order of 3%, possibly due to the fact that it has correlation with several modal responses. For the elasticity modulus of concrete (E_c) and the elasticity modulus of steel in M20 connectors between seating deck units (E_{M20}), which are less sensitive to the responses, estimates have higher variations and near 25%.

Regarding elasticity modulus of neoprene layer between seating deck unit and stand beams of row 13, at left (E_{N13L}) and right sides (E_{N13R}), the estimates show variations with opposite trend, that is, the increased stiffness of the left support is normally associated with a decreased

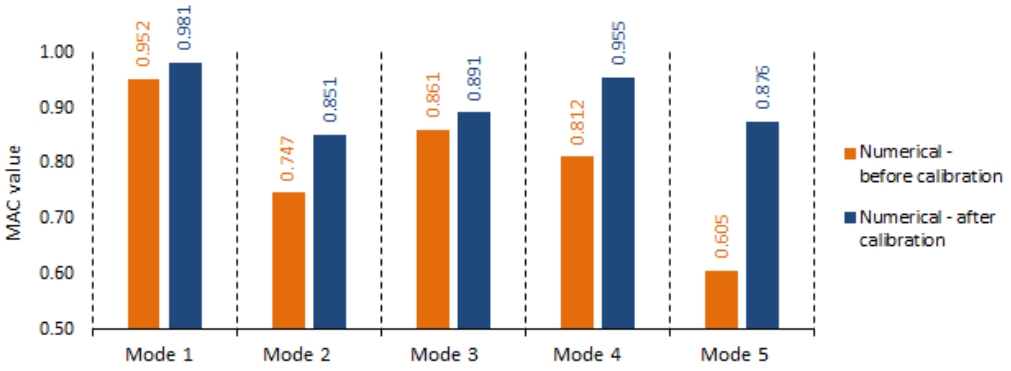
stiffness of the right support and vice-versa. The same applies to the elasticity modulus of neoprene layer between seating deck unit and stand beams of row 14, at left (E_{N14L}) and right sides (E_{N14R}). This should be related to the fact that there are different combinations of these sets of parameters that lead to the same solution in terms of the optimization problem.

Figure 11a presents the experimental and numerical values of vibration frequencies before and after calibration, indicating the values of errors of numerical and experimental vibration frequencies, with reference to the values of the experimental frequencies. The numerical results after calibration refer to the optimization run GA3, which was the case that led to the lowest residue of the objective function. Figure 11b shows the MAC values before and after calibration.

The average error of the frequencies decreased from 11.8% before calibration to 2.4% after calibration. The average value of the MAC parameter increased from 0.795 before calibration to 0.911 after calibration.



a)



b)

Figure 11: Correlation analysis of experimental and numerical modal parameters: a) frequencies of vibration; b) MAC values.

6 CONCLUSIONS

This article focused on the experimental calibration of a numerical model of seating deck units of Dragão stadium.

The three dimensional finite element numerical model includes a series of 10 consecutive seating deck units. In the modeling special attention was given to the link between the seating deck units as well as the connection of the seating deck units to the stand beams.

The dynamic test carried out on a series of seating deck units of the south stand of the stadium allowed the identification of five local vibration modes, mainly involving coupled bending and torsion movements of the seating deck units, with frequencies ranged between 12.15 Hz and 40.36 Hz.

The results of the optimization of the numerical model of the seating deck units demonstrated a very good approximation with experimental results and a significant improvement in relation to the numerical model before calibration. On the other hand, the genetic algorithm allowed to obtain sufficiently stable estimates of a significant number of parameters, considering different initial populations, proving its efficiency and robustness.

The analysis of the values of numerical parameters after calibration allowed to established that: i) the optimal value of the elasticity modulus of steel in M20 connectors between seating deck units approached the upper limit, reflecting, possibly, a more effective link between the seating deck units in relation to the planned design; ii) the values of elasticity modulus of neoprene layer between seating deck units and stand beams have very different values depending on the seat deck unit and position on the seat deck unit, which corroborates the evidences of a visual inspection carried out *in situ*, which in some situations found the absence of supports, and in other situations the existence of supports performed with materials and thicknesses different from those specified in design; iii) the optimal value of the elasticity modulus of concrete approached its lower bound, standing in the range between 28.6 GPa and 30.6 GPa.

As future developments the authors intend to carry out the validation of the numerical model of seating deck units under the action of controlled movements of public based on dynamic analysis that include public-structure interaction and taking into account eventual nonlinearities of the dynamic system. For this purpose, it is planned to carry out a dynamic test under public action with the measurement of the forces applied by individuals and the responses in terms of accelerations of seating deck units.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Porto Estádio for all the information provided about Dragão stadium and for all the support and facilities given during the experimental campaign.

REFERENCES

[1] SCOSS, Thirteenth Report of SCOSS - The Standing Committee on Structural Safety, Strutural Safety 2000-01, London, UK, 2001.

[2] T. Ji, J. Littler, B. Ellis, The response of grandstands to dynamic crowd loads, *Proceedings of the Institution of Civil Engineers: Structures and Buildings,* **140**, 355–365, 2000.

[3] P. Reynolds and A. Pavic, Modal testing of a sports stadium, in *20th International Modal Analysis Conference*, February,2002.

[4] A. Cigada, A. Caprioli, M. Redaelli, Numerical Modeling and Experimental Modal Analysis of a Concrete Grand-stand Structure to Structural Health Monitoring Purposes, in *Proceedings of the international conference (IMAC XXVI)*, Orlando, 2007.

[5] H. Marques, A. Arêde, R. Delgado, Vibration monitoring of a grandstand in Dragon Stadium, in *EVACES - Experimental Vibration Analysis for Civil Engineering Structures*, 2007.

[6] P. Marovic, M. Galic, M. Bertolino, Experimental testing of grandstand RC girders of the Spaladium Arena in Split, in *2009 Montanuniversität Leoben/Austria*, September 23-26, 2009.

[7] G. Lima, S. Avila, G. Doz, Numerical Dynamic Analysis of the New Brasilia National Stadium, in *11th International Conference on Vibration Problems*, September, 2013.

[8] G. Saudi, P. Reynolds, M. Zaki, H. Hodhod, Finite-Element Model Tuning of Global Modes of a Grandstand Structure Using Ambient Vibration Testing, *Journal of Performance of Constructed Facilities*, **23**, 467–479, 2003.

[9] Autodesk, *Autodesk Robot Structural Analysis 2015 - Getting Started Guide,* San Francisco, USA, 2014.

[10] SVS, *ARTeMIS Extractor Pro 2011 - Academic licence, Release 5.4*, Aalborg, Denmark, 2011.

[11] Microsoft, *Microsoft Office Excel 2013 user guide*, Redmond, USA, 2013.

[12] MathWorks, *MathWork (MATLAB) - Getting strated guide*. MathWorks, Natick, Massachusetts, USA, 2016.

[13] D. Ribeiro, R. Calçada, R. Delgado, M. Brehm, V. Zabel, Finite element model updating of a bowstring-arch railway bridge based on experimental modal parameters, *Engineering Structures*, **40**, 413–435, 2012.

[14] M. Brehm, V. Zabel, C. Bucher, An automatic mode pairing strategy using an enhanced modal assurance criterion based on modal strain energies, *Journal of Sound and Vibration*, **329**, 5375–5392, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DESIGN OF ISOTROPIC MICROSTRUCTURES VIA A TWO-SCALE APPROACH

Alexis Faure[1], Georgios Michailidis[1], Rafael Estevez[1], Guillaume Parry[1] and Grégoire Allaire[2]

[1]SIMaP, INP Grenoble
1130 rue de la Piscine, Domaine universitaire BP75, 38402 Saint Martin d'Hères, France
e-mail: {alexis.faure,georgios.michailidis,Rafael.Estevez,guillaume.parry}@simap.grenoble-inp.fr

[2] Centre de Mathématiques Appliquées (CMAP), École Polytechnique
CNRS UMR 7641, Université Paris-Saclay, 91128 Palaiseau, France
e-mail: gregoire.allaire@polytechnique.fr

Keywords: Materials, homogenization, topology optimization, level-set.

Abstract. *Architectured materials are promising to reach extreme properties and ultimately address issues related to lightweight or non conventional properties for bulk materials (eg. high specific rigidity, extremal conductivity or auxetism (negative Poisson's ratio)) [1]. A very efficient way to obtain optimal forms is via inverse homogenization, i.e. using shape and topology optimization techniques in order to achieve target material properties [2].*

A great number of publications has been devoted to the design of isotropic materials with extreme properties. Isotropy is usually prescribed via a combination of symmetric planes and penalization techniques, which are quite delicate to handle in an optimization framework.

In this work, we present an approach for the design of isotropic multi-materials with extremal conductivity via laminate geometries, consisting in anisotropic phases [3]. More specifically, we design composites with extremal conductivity using rank-1 laminates, composed by two orthotropic phases along parallel layers. The second phase is obtained by a 90-degree rotation of the first one, while their volume fractions are explicitly chosen so that the laminate is isotropic. The orthotropic phases are considered to have their own periodic micro-structure, composed by multiple phases. By optimally distributing the different phases in a periodic cell, we can achieve the Hashin-Shtrikman bounds for the isotropic laminate. We present examples in two dimensions using the level-set method for shape and topology optimization [4].

1 INTRODUCTION

Designing materials with prescribed or extremal properties is of great interest in material science. Combining the different attributes of materials in nature, e.g. by designing properly their microstructure, one can create novel materials with mechanical properties that are not met by any material in nature alone (extreme conductivity, auxetism, etc.). This class of materials is frequently reffered to as *architectured* or *hybrid materials* [5, 6].

Among the different methodologies proposed for their design [7], the geometric approach has recently started attracting again the interest of reseachers and manufacturers. This is mainly due to the recent advances in additive manufacturing, which have rendered possible the realization of such materials, broadening the perspectives to fill the holes in the material property maps [7]. In this approach, the microstructure of the hybrid material, composed by multiple phases or one phase and void, is designed by specifying the exact configuration of its consituents in a periodic unit cell.

Beyond analytical solutions, based mainly on the experience and intuition of researchers [1], numerical methods for the design of architectured materials have been presented long ago. Using the theory of periodic homogenization [8, 9], it amounts to formulate an optimization problem where the homogenized coefficients appear and the design variables are chosen to describe the geometry and connectivity of the microstructure's shape [10, 11, 2, 12].

Among the different possibilities, the design of isotropic materials is of particular interest [13, 2, 14]. For such materials, explicit bounds for their effective homogenized properties have been derived and are used for validating the efficiency of the numerical methods to provide morphologies with extremal properties. However, ensuring numerically the isotropy of the microstructure is a computationally tedious task and often leads to complicated geometries. It is of interest to search for methods able to produce simpler geometries and requiring less numerical effort for imposing the isotropy conditions.

In this work, we propose to design isotropic architectured materials with extremal conductivity, using a two-scale approach [3, 15]. The isotropic material at the macroscale (scale 1) disposes a microstructure (scale ϵ), composed of periodic rank-1 laminates. The two materials forming the laminate are assumed orthotropic and to have their own microstructure (scale ϵ^2). More specifically, the second material is obtained from the first via a 90° rotation, while their volume fraction at the scale ϵ is explicitly defined so that the global material is isotropic. The microstructure of the orthotropic materials is composed by one or mutliple phases, optimally distributed in a periodic unit cell such that the global laminate achieves the Hashin-Shtrikman bounds. Using the level-set method for Topology Optimization (T.O.) [4, 16], we present two-dimensional examples of microstructures achieving extremal conductivity.

2 SETTING OF THE PROBLEM

2.1 Two-scale approach

We search to design macroscopically isotropic materials with extreme conductivity coefficients. As aforementioned, the macroscopic material (see Figure 1,right) is assumed to be composed of periodic rank-1 laminates (see Figure 1,middle), with period $\epsilon << 1$. The laminate is formed by arranging two orthotropic phases in proportions f and $1-f$ correspondingly. The conductivity tensor of the first phase, in proportion f (see Figure 1,middle in orange colour) reads

$$A^0 = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}, \quad \alpha, \beta > 0, \tag{1}$$

while the one of the second phase, in proportion $1-f$ (see Figure 1,middle in blue colour), denoted A^{90}, is derived by rotating A^0 by 90°, i.e.

$$A^{90} = \begin{pmatrix} \beta & 0 \\ 0 & \alpha \end{pmatrix} = \Pi^t A^0 \Pi, \quad \text{where} \quad \Pi = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{2}$$

The effective (homogenized) conductivity tensor A^* of the material at the macroscale reads

$$A^* = \begin{pmatrix} A^*_{11} & 0 \\ 0 & A^*_{22} \end{pmatrix}, \tag{3}$$

where its coefficients A^*_{11},A^*_{22} are analytically expressed as functions of f, α, β using classical lamination formulas [17, 18], as

$$A^*(f,\alpha,\beta) = \begin{pmatrix} \left(\frac{f}{\alpha} + \frac{1-f}{\beta}\right)^{-1} & 0 \\ 0 & f\beta + (1-f)\alpha \end{pmatrix}. \tag{4}$$

One can then easily verify that choosing the proportion of mixture f equal to

$$f = f(\alpha,\beta) = \frac{\sqrt{\alpha\beta} - \alpha}{\beta - \alpha}, \tag{5}$$

the homogenized tensor A^* becomes isotropic ($A^*_{11} = A^*_{22}$). In the particular case where $\alpha = \beta$, the laminate is isotropic independently of f. In this case, one can choose f as the limit of (5) when $\alpha \to \beta$, that is $f = 0.5$.

Remark 2.1. *The above method is generalized in three dimensions using rank-2 laminates constructed as in [3].*

Until now, we have considered the coefficients of the conductivity tensor A^0 to be constants. Assuming that the orthotropic material, used as constituent for the rank-1 laminate, disposes its own periodic microstructure at a scale $\epsilon^2 << \epsilon$ (see Figure 1,left), we search to optimally distribute one or multiple phases in a periodic unit cell Y in order to achieve extremal conductivity for the homogenized global conductivity matrix $A^*(f,\alpha,\beta)$. Therefore, the conductivity tensor A^0 and the global conductivity tensor A^* now read

$$A^0 = \begin{pmatrix} A^0_{11}(\Omega) & 0 \\ 0 & A^0_{22}(\Omega) \end{pmatrix}, \quad A^0_{11}(\Omega), A^0_{22}(\Omega) > 0,$$

and

$$A^*(f, A^0_{11}(\Omega), A^0_{11}(\Omega)) \equiv A^*(f,\Omega).$$

2.2 Optimization problem

For the sake of simplicity, we present the case where the unit cell Y is filled by a single phase, occupying a domain $\Omega \subset Y$. The extension to multiple phases has been thoroughly presented in [19, 20].

Furthermore, we assume that the scale $\epsilon^2 << \epsilon << 1$ remains sufficiently large compared to the atomistic scale, in order for the framework of continuum mechanics to remain valid for

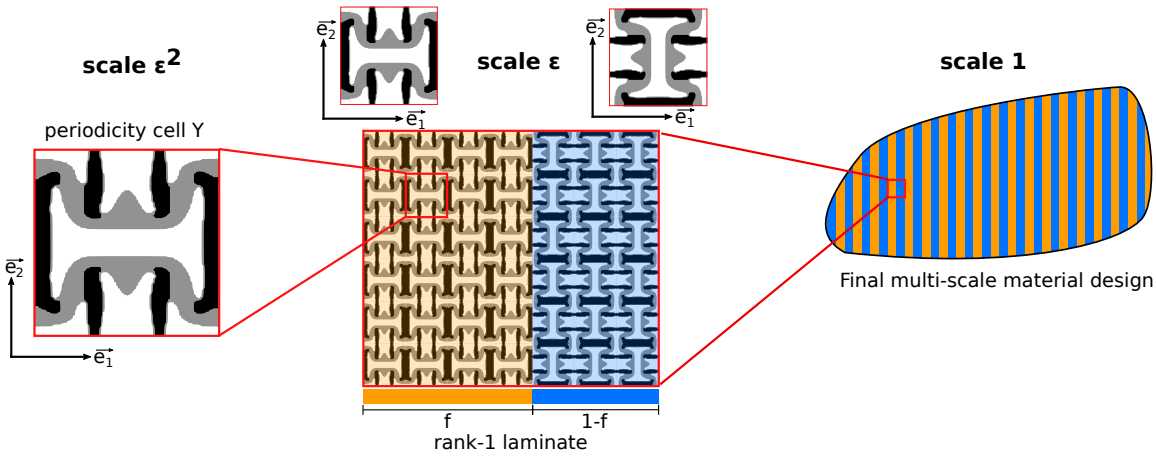


Figure 1: Two-scale approach for designing isotropic homogenized media.

our analysis. Then, the homogenized coefficients of the tensor $A^0(\Omega)$ are calculated using the formula

$$A^0_{ii} = \int_Y A(y)\,(e_i + \nabla_y w_i(y))\,(e_i + \nabla_y w_i(y))\,dy, \tag{6}$$

where w_i $(i = 1, 2)$ are solutions of the cell problems (see section 7.2.2 in [18])

$$\begin{cases} -\mathrm{div}_y\left(A(y)(e_i + \nabla_y w_i(y))\right) = 0 & \text{in Y} \\ y \to w_i(y) & \text{Y-periodic,} \end{cases} \tag{7}$$

e_i are vectors of the canonical Euclidean basis and $A(y)$ denotes the conductivity tensor of the consituent phases at $y \in Y$.

In order to design isotropic materials with extreme conductivity coefficients $A^*(f, \Omega)$, for a certain volume fraction of the consituent material, one can solve the optimization problem

$$\begin{aligned} \min_{\Omega} \quad & J(\Omega) = \pm tr(A^*(f,\Omega)) = \pm\left(A^*_{11}(f,\Omega) + A^*_{22}(f,\Omega)\right) \\ \text{s.t. :} \quad & A^*_{11}(f,\Omega) = A^*_{22}(f,\Omega) \\ & V(\Omega) = \alpha|Y|, \ \alpha \in (0,1), \end{aligned} \tag{8}$$

where the "+" and "-" signs correspond to design for minimal and maximal conductivity correspondingly.

3 TOPOLOGY OPTIMIZATION FRAMEWORK

In the framework of structural optimization, Topology Optimization refers to a form-finding method where the design variables are chosen to describe the shape and the connectivity of the structure. In general, a T.O. method is characterized by two main choices:

- a method to describe the shape and
- a method to evolve the shape during the optimization process.

Among the variety of T.O. developed during the last decades [21], we favor a geometric approach using the level-set method for the shape description, coupled with a shape sensitivity analysis for the shape advection [4]. The basic characteristics of the method are briefly presented in the following of this Section.

3.1 Level-set method

The level-set method, developped by Osher and Sethian [22], uses an implicit representation of an evolving front as the zero level-set of an auxiliary function ϕ. More precisely, assuming that the domain Ω of interest is a subset of a big working domain D, the level-set representation of Ω can be defined as (see Figure 2)

$$\begin{cases} \phi(x) = 0 & \leftrightarrow x \in \partial\Omega \cap D, \\ \phi(x) < 0 & \leftrightarrow x \in \Omega, \\ \phi(x) > 0 & \leftrightarrow x \in \left(D \setminus \overline{\Omega}\right), \end{cases} \tag{9}$$

A great benefit of the level-set representation for T.O. consists in performing topological

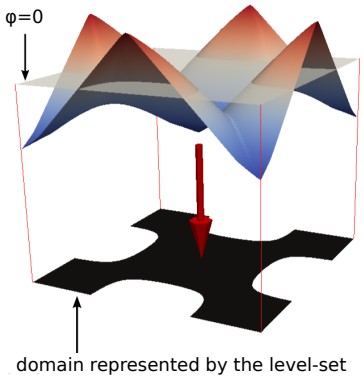


Figure 2: Level-set function representing a two-dimensional domain (in black).

changes with great ease and robustness, contrary to other methods using a parametrization for the shape description. Moreover, multiple phases can be described in a natural way [19, 23]. Defining n level-set functions in the same design domain and combining their values, one can describe up to $m = 2^n$ different materials (see Figure 3). The advection of a front (shape bound-

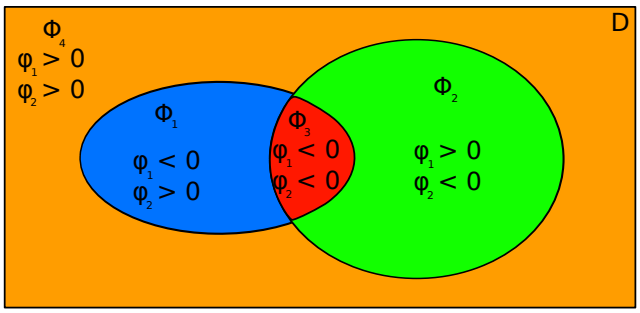


Figure 3: Multiphase representation in the level-set framework.

ary) with a velocity $V(x)$, normal to the shape boundary, is described in the level-set framework by introducing a pseudo-time, $t \in \mathbb{R}^+$, and solving the well-known Hamilton-Jacobi transport equation:

$$\frac{\partial \phi}{\partial t} + V(x)|\nabla \phi| = 0, \tag{10}$$

using an explicit second order upwind scheme [24], [25].

3.2 Shape sensitivity approach

In T.O. the velocity field $V(x)$ used for the shape evolution is chosen so that the objective function gets iteratively reduced. To employ a gradient-based optimization method, one needs to compute a derivative for the functionals involved in the optimization problem with respect to changes of the optimization parameters. However, no such parameters are a priori defined in the level-set framework. In [4], the authors proposed to derive such a descent direction via a classical shape sensitivity analysis. In this method, which dates back to Hadamard, a notion of **shape derivative**, i.e. a derivative of a functional with respect to variations of the shape in a direction $\theta(x) = V(x)n(x)$, can be defined as follows [18, 26].

Starting from a domain Ω, one considers perturbations by a smooth enough vector field $\theta(x)$, such that the new domain, Ω_θ, is described by (see Figure 4):

$$\Omega_\theta = \big(Id + \theta\big)\Omega.$$

Then, the shape derivative $J'(\Omega)(\theta)$ of the functional $J(\Omega)$ in a direction $\theta(x)$ is obtained

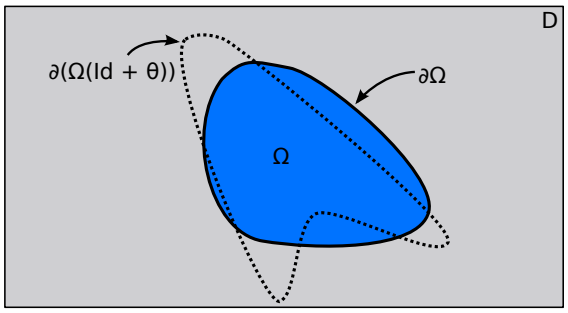


Figure 4: Perturbation of the domain Ω via a vector field $\theta(x)$.

through an asymptotic expansion formula of the type:

$$J\big((Id+\theta)\Omega\big) = J(\Omega) + J'(\Omega)(\theta) + o(\theta), \quad \text{with} \quad \lim_{\theta\to 0} \frac{|o(\theta)|}{\|\theta\|} = 0\ .$$

Once calculated, a descent direction can be found by advecting the shape in the direction $\theta = -tJ'(\Omega),\ t > 0$. For the new shape $\Omega_t = (\,\mathrm{Id} + t\theta)\,\Omega$, if $V \neq 0$, one can formally write:

$$J\left(\Omega_t\right) = J\left(\Omega\right) - t(J'(\Omega))^2 + \mathcal{O}(t^2) < J\left(\Omega\right),$$

which guarantees a descent direction for small positive t.

3.3 SHAPE DERIVATIVE

In this section, we compute a shape derivative for the objective function $J(\Omega)$ defined in (8). The isotropy constraint is treated by choosing f to satisfy the relation (5).

The shape derivative of $J(\Omega)$ reads

$$\begin{aligned} J'(\Omega)(\theta) &= \sum_{i=1}^{2} \pm \frac{\partial A_{ii}^*}{\partial \Omega}(\theta) \\ &= \sum_{i=1}^{2} \left[\sum_{j=1}^{2} \pm \frac{\partial A_{ii}^*}{\partial A_{jj}^0} \frac{\partial A_{jj}^0}{\partial \Omega}(\theta) \pm \frac{\partial A_{ii}^*}{\partial f} \frac{\partial f}{\partial A_{jj}^0} \frac{\partial A_{jj}^0}{\partial \Omega}(\theta) \right], \end{aligned} \tag{11}$$

where

$$\frac{\partial A^*_{11}}{\partial A^0_{11}} = \frac{(A^0_{22})^2 f}{(A^0_{22}f + (1-f)A^0_{11})^2} \quad , \quad \frac{\partial A^*_{11}}{\partial A^0_{22}} = \frac{(A^0_{11})^2 (1-f)}{(A^0_{22}f + (1-f)A^0_{11})^2}$$

$$\frac{\partial A^*_{22}}{\partial A^0_{11}} = (1-f) \quad , \quad \frac{\partial A^*_{22}}{\partial A^0_{22}} = f$$

$$\frac{\partial A^*_{11}}{\partial f} = \frac{-A^0_{11}A^0_{22}(A^0_{22} - A^0_{11})}{(A^0_{22}f + (1-f)A^0_{11})^2} \quad , \quad \frac{\partial A^*_{22}}{\partial f} = A^0_{22} - A^0_{11}$$

$$\frac{\partial f}{\partial A^0_{11}} = \frac{A^0_{22}(\sqrt{A^0_{11}} - \sqrt{A^0_{22}})^2}{2\sqrt{A^0_{11}A^0_{22}}(A^0_{22} - A^0_{11})^2} \quad , \quad \frac{\partial f}{\partial A^0_{22}} = \frac{-A^0_{11}(\sqrt{A^0_{11}} - \sqrt{A^0_{22}})^2}{2\sqrt{A^0_{11}A^0_{22}}(A^0_{22} - A^0_{11})^2}$$

and the shape derivative of the homogenized coefficients $\frac{\partial A^0_{jj}}{\partial \Omega}(\theta)$ are provided by the expression (see [27, 12])

$$\frac{\partial A^0_{jj}}{\partial \Omega}(\theta) = \int_{\partial\Omega} \theta(x) \cdot n(x) \left[A(x) \left(e_j + \nabla_y w_j(x) \right) \left(e_j + \nabla_y w_j(x) \right) \right] dx.$$

4 NUMERICAL RESULTS

In order to test the efficiency of the proposed methodology and algorithm, we compare our results with the Hashin-Shtrikman (HS) bounds [28, 29] for extremal conductivity coefficients of isotropic composites, using multiple isotropic phases. The lower (λ^l_{HS}) and upper (λ^u_{HS}) HS bounds read

$$\lambda^l_{HS} = -\min_i(\lambda_i) + \left[\sum_i \frac{V_i}{\lambda_i + \min_j(\lambda_j)} \right]^{-1} \tag{12}$$

and

$$\lambda^u_{HS} = -\max_i(\lambda_i) + \left[\sum_i \frac{V_i}{\lambda_i + \max_j(\lambda_j)} \right]^{-1} , \tag{13}$$

where V_i, λ_i the volume fraction and the thermal conductivity of the ith phase, respectively.

For all our numerical results, we discretize both the level-set advection equation (10) and the cell problem (7) on a regular grid. For the cell problem Q1 finite elements have been used. The optimization problem (8) is solved using an SLP method.

4.1 Bi-material composites for minimal conductivity

In the first example, we design composites with minimal conductivity using two isotropic phases 1 and 2, with conductivity $\lambda_1 = 1$ and $\lambda_2 = 10$ respectively. Material 1 (in black colour) occupies domain Ω, while material 2 (in white colour) occupies the complementary of the periodic unit cell $Y \setminus \Omega$. The optimization problem reads

$$\begin{aligned} \min_{\Omega} \quad & J(\Omega) = tr(A^*(f,\Omega)) = (A^*_{11}(f,\Omega) + A^*_{22}(f,\Omega)) \\ \text{s.t. :} \quad & A^*_{11}(f,\Omega) = A^*_{22}(f,\Omega) \\ & V(\Omega) = 0.5|Y|. \end{aligned} \tag{14}$$

The HS lower bound for this case equals $\lambda^l_{HS} = 2.385$.

In Figure 5 we see two optimized shapes (right column) for two different initializations (left column). For the first result (upper row), the homogenized conductivity coefficients A^*_{ii} equal 2.588, while the volume fraction of laminate A^0 equals $f = 0.5$. For the second result (lower row), the homogenized conductivity coefficients A^*_{ii} equal 2.433 and the volume fraction of laminate A^0 equals $f = 0.506$.

As expected, in both results the less conductive material (in black) "encapsulates" the highly conductive phase (in white), leading to a composite with effective conductivity close to the HS bounds.

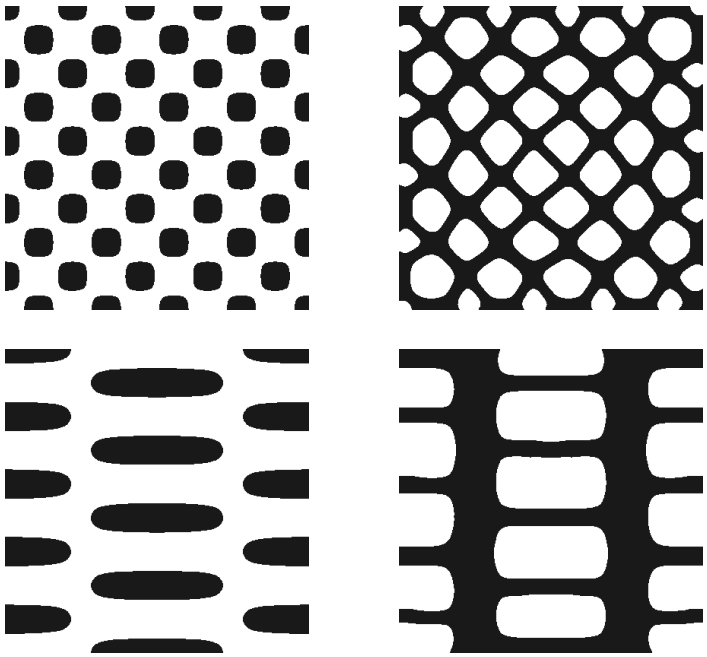


Figure 5: Bi-material design for minimal conductivity; initializations (left) and resulting optimal shapes (right).

4.2 Three-phase composites for maximal conductivity

In the second example, we design composites with maximal conductivity using three isotropic phases 1, 2 and 3, with conductivity $\lambda_1 = 10$, $\lambda_2 = 2$ and $\lambda_3 = 1$. Materials 1 and 2, in red and white colour respectively, occupy domains Ω_1 and Ω_2, while material 3 (in blue colour) occupies the complementary of the periodic unit cell. The optimization problem reads

$$\begin{aligned} \min_{\Omega_i} \quad & J(\Omega_i) = -tr(A^*(f,\Omega_i)) = -\left(A^*_{11}(f,\Omega_i) + A^*_{22}(f,\Omega_i)\right) \\ \text{s.t. :} \quad & A^*_{11}(f,\Omega_i) = A^*_{22}(f,\Omega_i) \\ & V(\Omega_i) = V_i,\ i = 1,2,3. \end{aligned} \tag{15}$$

In Figure 6, we show different optimized shapes for various combinations of volume fractions for the three constitutive phases. The values of the homogenized coefficients appear in Table 1 and are compared both with the upper HS bounds λ^u_{HS}, as well as with results obtained by Zhou et al. in [29].

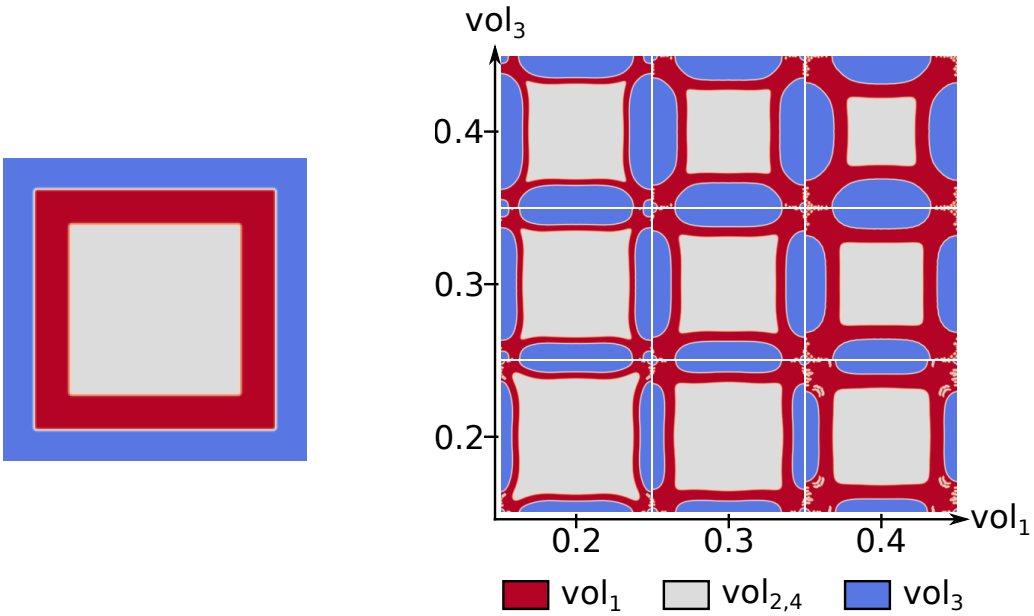


Figure 6: Initialization (left) and optimized shapes (right) for different volume fractions in the unit periodic cell Y.

Case	V_1	V_3	λ_1	λ_2	λ_3	λ^u_{HS}	λ_{Zhou}	A^*_{ii}
a	0.2	0.2	10	2	1	2.791	—	2.803
b	0.2	0.3	10	2	1	2.668	—	2.671
c	0.2	0.4	10	2	1	2.548	2.501	2.554
d	0.3	0.2	10	2	1	3.360	—	3.367
e	0.3	0.3	10	2	1	3.226	—	3.241
f	0.3	0.4	10	2	1	3.095	—	3.110
g	0.4	0.2	10	2	1	3.983	3.903	4.017
h	0.4	0.3	10	2	1	3.836	—	3.857
i	0.4	0.4	10	2	1	3.693	—	3.716

Table 1: Results of the optimized shapes of Figure 6.

Remark 4.1. *We shall note that the overpass of the HS bounds that appears in Table 1 is purely numerical. In fact, using the multi-material setting in [19], the volume of each phase is calculated approximatively, using the same smooth interpolation functions as for the mechanical properties. Reducing the interpolation width, the results converge to the HS bounds.*

As expected, starting from different initializations, one may obtain different optimized shapes. In Figure 7(right), we show the optimized shape for the case (c) in Table 1, starting from the initialization shown in Figure 7(left). The value of the homogenized coefficients is similar to Table 1(c) and equals 2.580. Similarly, starting from different initializations for the case (g) in Table 1, the optimized shapes are shown in Figure 8, where a repetition of the same pattern is observed. For all the results of this section, a cubic symmetry is observed, which leads to a volume fraction of $f = 0.5$ for the laminate A^0 at scale ϵ.

5 CONCLUSION

A novel method, at least to our knowledge, for the design of isotropic composites using T.O. has been proposed, based on a two-scale approach. The method has been applied for the

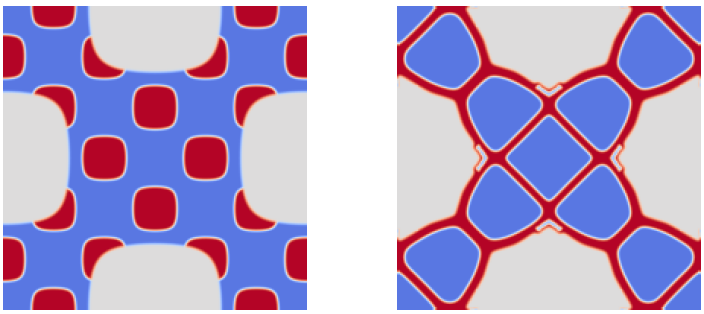


Figure 7: Initialization (left) and optimized shapes (right) for the case (c) in Table 1, starting from a different initialization.

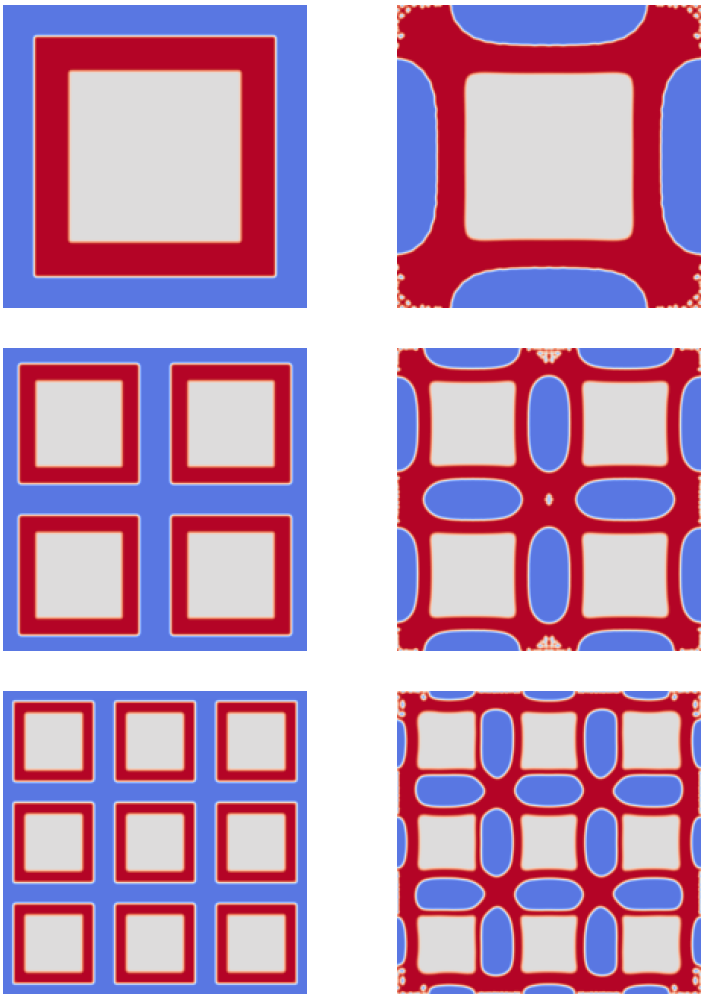


Figure 8: Initializations (left) and optimized shapes (right) for the case (g) in Table 1, starting from a different initializations.

design of architectured multi-materials with extremal conductivity, using the level-set method for Topology Optimization. The obtained results are very close to the HS bounds. Future work shall include three-dimensional results as well as the elasticity framework, aiming to obtain extremal micro-structures with simpler geometry.

REFERENCES

[1] G. Jefferson, T. Parthasarathy, and R. Kerans, "Tailorable thermal expansion hybrid structures," *International Journal of Solids and Structures*, vol. 46, no. 11, pp. 2372–2387, 2009.

[2] O. Sigmund, "Tailoring materials with prescribed elastic properties," *Mechanics of Materials*, vol. 20, no. 4, pp. 351–368, 1995.

[3] M. Avellaneda and G. Milton, "Optimal bounds on the effective bulk modulus of polycrystals," *SIAM Journal on Applied Mathematics*, vol. 49, no. 3, pp. 824–837, 1989.

[4] G. Allaire, F. Jouve, and A.-M. Toader, "Structural optimization using sensitivity analysis and a level-set method," *Journal of computational physics*, vol. 194, no. 1, pp. 363–393, 2004.

[5] M. Ashby, "Designing architectured materials," *Scripta Materialia*, vol. 68, no. 1, pp. 4–7, 2013.

[6] M. Ashby and Y. Brechet, "Designing hybrid materials," *Acta materialia*, vol. 51, no. 19, pp. 5801–5821, 2003.

[7] M. Ashby, "Hybrids to fill holes in material property space," *Philosophical Magazine*, vol. 85, no. 26-27, pp. 3235–3257, 2005.

[8] F. Murat and L. Tartar, "Calcul des variations et homogénéisation," *Les méthodes de lhomogénéisation: théorie et applications en physique*, no. 57, pp. 319–369, 1985.

[9] L. Tartar, *The general theory of homogenization: a personalized introduction*, vol. 7. Springer, 2009.

[10] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve, "Shape optimization by the homogenization method," *Numerische Mathematik*, vol. 76, no. 1, pp. 27–68, 1997.

[11] M. Bendsøe and N. Kikuchi, "Generating optimal topologies in structural design using a homogenization method," *Computer methods in applied mechanics and engineering*, vol. 71, no. 2, pp. 197–224, 1988.

[12] M. Wang and X. Wang, "A level-set based variational method for design and optimization of heterogeneous objects," *Computer-Aided Design*, vol. 37, no. 3, pp. 321–337, 2005.

[13] E. Andreassen, B. Lazarov, and O. Sigmund, "Design of manufacturable 3d extremal elastic microstructure," *Mechanics of Materials*, vol. 69, no. 1, pp. 1–10, 2014.

[14] O. Sigmund, "A new class of extremal composites," *Journal of the Mechanics and Physics of Solids*, vol. 48, no. 2, pp. 397–428, 2000.

[15] G. Milton, *The theory of composites*, vol. 6. Cambridge University Press, 2002.

[16] M. Wang, X. Wang, and D. Guo, "A level set method for structural topology optimization," *Computer methods in applied mechanics and engineering*, vol. 192, no. 1, pp. 227–246, 2003.

[17] G. Allaire, *Shape optimization by the homogenization method*, vol. 146 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2002.

[18] G. Allaire, *Conception optimale de structures*, vol. 58 of *Mathématiques & Applications (Berlin)*. Springer-Verlag, Berlin, 2007.

[19] G. Allaire, C. Dapogny, G. Delgado, and G. Michailidis, "Mutli-phase structural optimization via a level-set method," *ESAIM: Control, Optimisation and Calculus of Variations, 20, pp 576-611. doi:10.1051/cocv/2013076*, 2014.

[20] N. Vermaak, G. Michailidis, G. Parry, R. Estevez, Y. Brechet, and G. Allaire, "Material interface effects on the topology optimization of multi-phase thermoelastic structures using a level set method," *(submitted in SMO: Structural and Multi-disciplinary Optimization)*, 2013.

[21] M. Bendsoe and O. Sigmund, *Topology optimization: theory, methods and applications*. Springer, 2004.

[22] S. Osher and J. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.

[23] M. Wang and X. Wang, "Color level sets: a multi-phase method for structural topology optimization with multiple materials," *Computer Methods in Applied Mechanics and Engineering*, vol. 193, no. 6, pp. 469–496, 2004.

[24] S. Osher and R. Fedkiw, *Level set methods and dynamic implicit surfaces*, vol. 153 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2003.

[25] J. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge university press, 1999.

[26] J. Simon and F. Murat, "Sur le contrôle par un domaine géométrique," *Publication 76015 du Laboratoire d'Analyse Numérique de l'Université Paris VI*, no. 76015, p. 222 pages, 1976.

[27] G. Michailidis, *Manufacturing Constraints and Multi-Phase Shape and Topology Optimization via a Level-Set Method*. PhD thesis, Ecole Polytechnique X, 2014, available at: http://pastel.archives-ouvertes.fr/pastel-00937306.

[28] A. Cherkaev, "Bounds for effective properties of multimaterial two-dimensional conducting composites," *Mechanics of Materials*, vol. 41, no. 4, pp. 411–433, 2009.

[29] S. Zhou and Q. Li, "Computational design of multi-phase microstructural materials for extremal conductivity," *Computational Materials Science*, vol. 43, no. 3, pp. 549–564, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SHAPE OPIMIZATION OF GRAPHENE SHEETS FOR MAXIMUM FUNDAMENTAL FREQUENCY

Jin-Xing Shi [1] **and Masatoshi Shimoda** [2]

[1] Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi 468-8511, Japan
shi@toyota-ti.ac.jp

[2] Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi 468-8511, Japan
shimoda@toyota-ti.ac.jp

Keywords: Graphene sheets, Frame structure, Frequency, Shape optimization,.

Abstract. *Graphene sheets (GSs) have been proposed to be a base material for nanoelectro-mechanical systems (NEMS) because of their excellent mechanical and electrical behaviors. In this work, we carry out shape optimization of GSs to enhance their dynamic behaviors. According to the Tersoff-Brenner force field theory and a link between molecular mechanics and solid mechanics of C-C bond, we model GSs as continuum frame structures at first. Then, we optimize the shape of the atomistic finite element models based on a developed free-form optimization method for frame structures. In this optimization process, we use the fundamental frequency as objective function and maximize it under a volume constraint and considering repeated eigenvalue problem. We assume each equivalent continuum beam to vary in the off-axis direction to the centroidal axis, and derive the shape gradient functions for determining the optimal design velocity field. According to the derived optimal design velocity field, the shape optimization of GSs can be carried out without shape parametrization. The numerical results show that, the fundamental frequency of GSs can be significantly enhanced after shape optimization, which would be helpful for applying GSs in NEMS.*

1 INTRODUCTION

Vibrational properties of graphene sheets (GSs) deserve to be studied, because they may contribute to the application of GSs as NEMS resonators. Up to now, various continuum mechanical models have been proposed to study the vibration modes of GSs, such as frame structure model, shell/plate model, and beam model [1-4]. Shape optimization is a procedure to improve or enhance the performance of the structure by changing design parameters. Recently, the shape of GSs can be controlled by an external electric field [5] or chemically modifying the adherence of GSs on metal [6]. Hence, shape optimization of GSs can make an effective role to improve their mechanical behaviors.

In our previous work, we adopted the molecular mechanics (MM) method to construct the continuum frame work and carried out the shape optimum design of GSs in terms of the compliance minimization problem based on a free-form optimization method for frame structures [7]. In the present work, we extend this optimization method to the fundamental frequency maximization problem of GSs, and make shape optimum design of GSs.

We arrange this study as following. In section 2, we introduce the MM method for constructing the frame structures of GSs. In section 3, we develop a free-form shape optimization method for frame structures and introduce the shape optimization process for shape design of GSs. In section 4, using the developed shape optimization process, we carry out two examples to obtain the optimal shapes of GSs. At last, we remark conclusions in section 5.

2 SHAPE OPTIMIZATION OF GRAPHENE SHEETS

2.1 Governing equation

We introduce the MM method for modeling frame structure of GSs at first. In MM method, the forces between two individual atoms are depicted as continuum beam elements. We assume an equivalent C-C beam with a circular cross-section of diameter d and initial length 1.42 Å. According to the Tersoff-Brenner force field theory [8] and a link between molecular and solid mechanics of C-C bond, we derive the Young's modulus E_b, shear modulus G_b and diameter d as [9]:

$$E_b = \frac{k_l^2 l}{4\pi k_\theta},\ G_b = \frac{k_l^2 k_\tau l}{8\pi k_\theta^2},\ d = 4\sqrt{\frac{k_\theta}{k_l}} \tag{1}$$

where k_l = 93800 kcal/mol/nm^2, k_θ = 126 kcal/mol/rad^2, and k_τ = 40 kcal/mol/rad^2 are obtained by fitting to the structural and vibrational frequency data on small molecular fragments [10].

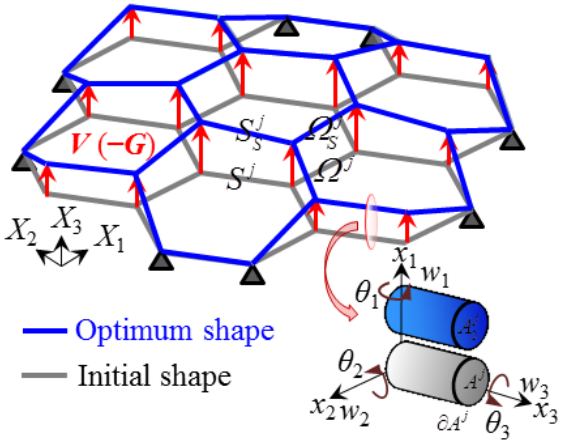


Figure 1: Shape variation of graphene sheets.

As shown in Fig. 2, a frame structure of GS with a bounded domain $\Omega \subset \mathbb{R}^3$ consists of Timoshenko beams Ω^j, $j = 1, 2, ..., N$, where $\mathbb{R}$ is a set of positive real numbers and N is the number of beams. The notations (x_1, x_2, x_3) and (X_1, X_2, X_3) indicate the local coordinate system in terms of a Timoshenko beam and the global coordinate system, respectively.

$$\Omega^j = \left\{ \left(x_1^j, x_2^j, x_3^j \right) \in \mathbb{R}^3 \middle| \left(x_1^j, x_2^j \right) \in A^j \subset \mathbb{R}^2, x_3^j \in S^j \subset \mathbb{R} \right\},\ \Gamma^j = \partial A^j \times S^j,\ \Omega^j = A^j \times S^j \quad (2)$$

where S^j, Γ^j and Ω^j express the centroidal axis, circumference surface and whole domain of member *j*, respectively. A^j and ∂A^j are the cross section and its circumference of member *j*, respectively. The subscript *j* shall be omitted to avoid the complexity of expression in the sequel. $\boldsymbol{w} = \{w_i\}_{i=1,2,3}$ and $\boldsymbol{\theta} = \{\theta_i\}_{i=1,2,3}$ are the displacement vector and rotation vector in the x_1, x_2, x_3 directions of the local coordinate system, respectively. Then, the weak form governing equation of the eigenfrequency in terms of $(\boldsymbol{w}^{(r)}, \boldsymbol{\theta}^{(r)})$ can be expressed as

$$a\left(\left(\boldsymbol{w}^{(r)}, \boldsymbol{\theta}^{(r)} \right), \left(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}} \right) \right) = \lambda^{(r)} b\left(\left(\boldsymbol{w}^{(r)}, \boldsymbol{\theta}^{(r)} \right), \left(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}} \right) \right),\ \forall \left(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}} \right) \in U, \left(\boldsymbol{w}^{(r)}, \boldsymbol{\theta}^{(r)} \right) \in U \quad (3)$$

where $(\cdot)^{(r)}$ denotes the eigenvector of *r* th natural frequency mode and $\lambda^{(r)}$ denotes its eigenvalue (the square of the *r* th natural frequency). The notation $(\bar{\cdot})$ expresses a variation, and *U* is the admissible function space, in which the given constraint conditions of $(\boldsymbol{w}, \boldsymbol{\theta})$ are satisfied. In the frame structure of GSs as shown in Fig. 2, due to the domain variation $\boldsymbol{V}$ (design velocity field) in the out-of-plane direction, the initial domain Ω^j and a centroidal axis S^j of member *j* become Ω_s^j and S_s^j, respectively. The subscript *s* expresses the iteration history of the domain variation. Moreover, the bilinear form $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are the rigidity item and inertia item, respectively.

2.2 Fundamental frequency maximization problem

The free-form optimization method for frame structures is adopted to maximize the natural frequency of GSs, and is formulated as

$$\text{Given} \quad \Omega \quad (4)$$

$$\text{Find} \quad \boldsymbol{V} \quad (5)$$

$$\text{that minimizes} \quad -\lambda^{(1)} \quad (6)$$

$$\text{subject to} \quad \text{Eq. (3) and} \quad M\left(= \sum_{j=1}^{N} \int_{S^j} A dS \right) \le \hat{M} \quad (7)$$

where *M* and $\hat{M}$ are the volume of a GS and its constraint value, respectively.

Letting $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}})$ and Λ denote the Lagrange multipliers for the governing equation and volume constraint, respectively, the Lagrange functional *L* associated with the fundamental frequency maximization problem can be expressed as

$$\begin{aligned} L\left(\Omega, \left(\boldsymbol{w}^{(1)}, \boldsymbol{\theta}^{(1)} \right), \left(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}} \right), \Lambda \right) = & -\lambda^{(1)} + \lambda^{(1)} b\left(\left(\boldsymbol{w}^{(1)}, \boldsymbol{\theta}^{(1)} \right), \left(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}} \right) \right) \\ & - a\left(\left(\boldsymbol{w}^{(1)}, \boldsymbol{\theta}^{(1)} \right), \left(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\theta}} \right) \right) + \Lambda\left(M - \hat{M} \right) \end{aligned} \quad (8)$$

The material derivative $\dot{L}$ of the Lagrange functional is derived as

$$\begin{aligned}\dot{L} = {} & \dot{\lambda}^{(1)}\left\{b\left(\left(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}\right),\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right)-1\right\}+\lambda^{(1)}b\left(\left(\boldsymbol{w}^{(1)\prime},\boldsymbol{\theta}^{(1)\prime}\right),\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right)\\ & +\lambda^{(1)}b\left(\left(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}\right),\left(\bar{\boldsymbol{w}}',\bar{\boldsymbol{\theta}}'\right)\right)-a((\boldsymbol{w}^{(1)\prime},\boldsymbol{\theta}^{(1)\prime}),(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}))-a((\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}),(\bar{\boldsymbol{w}}',\bar{\boldsymbol{\theta}}'))\\ & +\dot{\Lambda}\left(M-\hat{M}\right)+\langle G\boldsymbol{n},\boldsymbol{V}\rangle,\ \boldsymbol{V}\in C_{\Theta}\end{aligned} \tag{9}$$

where $G\boldsymbol{n}$ ($\equiv \boldsymbol{G}$) expresses the shape gradient function (i.e., sensitivity function), which is a coefficient function in terms of $\boldsymbol{V}$. $\boldsymbol{n}$ is the outward unit normal vector on the circumference surface Γ or a unit normal vector on the centroidal axis S. The notations $(\cdot)'$ and $\dot{(\cdot)}$ are the shape derivative and the material derivative with respect to the domain variation, respectively.

The optimum conditions of the Lagrange functional L with respect to $(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)})$, $(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}})$, and Λ are shown as

$$a\left(\left(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}\right),\left(\bar{\boldsymbol{w}}',\bar{\boldsymbol{\theta}}'\right)\right)=\lambda^{(1)}b\left(\left(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}\right),\left(\bar{\boldsymbol{w}}',\bar{\boldsymbol{\theta}}'\right)\right),\ \ \forall\left(\bar{\boldsymbol{w}}',\bar{\boldsymbol{\theta}}'\right)\in U \tag{10}$$

$$a\left(\left(\boldsymbol{w}^{(1)\prime},\boldsymbol{\theta}^{(1)\prime}\right),\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right)=\lambda^{(1)}b\left(\left(\boldsymbol{w}^{(1)\prime},\boldsymbol{\theta}^{(1)\prime}\right),\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right),\ \ \forall\left(\boldsymbol{w}^{(1)\prime},\boldsymbol{\theta}^{(1)\prime}\right)\in U \tag{11}$$

$$b\left(\left(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}\right),\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right)=1 \tag{12}$$

$$\dot{\Lambda}\left(M-\hat{M}\right)=0\ \ M-\hat{M}\le 0\ \ \Lambda\ge 0 \tag{13}$$

When the optimality conditions are satisfied, considering the self-adjoint relationship $(\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)})=(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}})$, we get

$$\dot{L}=\langle G\boldsymbol{n},\boldsymbol{V}\rangle=\sum_{j=1}^{N}\int_{S^j}\left\{G_1\boldsymbol{V}\cdot\boldsymbol{n}_1+G_2\boldsymbol{V}\cdot\boldsymbol{n}_2+G_0\boldsymbol{V}\cdot\boldsymbol{n}\right\}dS \tag{14}$$

where the shape gradient functions G_1, G_2, and G_0 are applied in the free-form optimization method for frame structures of GSs to determine the optimal design velocity field $\boldsymbol{V}$.

2.3 Free-form optimization method and repeated eigenvalue problem

The free-form optimization method for frame structures was proposed by Shimoda [11]. In this method, the negative shape gradient function $-\boldsymbol{G}(=-G\boldsymbol{n})$ is applied as a distributed force in the off-axis direction to the centroidal axis of a fictitious-elastic frame structure under a Robin condition (spring constant α). This makes it possible both to reduce the objective functional and to maintain smoothness, i.e., mesh regularity, simultaneously. The optimal shape variation, or the optimal design velocity field $\boldsymbol{V}$ is determined as the displacement field in this pseudo-elastic frame analysis, and the obtained $\boldsymbol{V}$ is used to update the shape. This analysis is called velocity analysis. The governing equation of the velocity analysis is shown as

$$a\left((\boldsymbol{V},\boldsymbol{\theta}),\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right)+\alpha\left\langle(\boldsymbol{V}\cdot\boldsymbol{n})\boldsymbol{n},\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right\rangle=-\left\langle G\boldsymbol{n},\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\right\rangle,\ \forall\left(\bar{\boldsymbol{w}},\bar{\boldsymbol{\theta}}\right)\in C_{\Theta},\ (\boldsymbol{V},\boldsymbol{\theta})\in C_{\Theta} \tag{15}$$

$$C_{\Theta}=\left\{(V_1,V_2,V_3,\theta_1,\theta_2,\theta_3)\in\left(H^1(S)\right)^6\,\middle|\,\text{satisfy Dirichlet condition for shape variation}\right\} \tag{16}$$

In design problems where convexity is assured, this relationship definitely reduces the Lagrange functional in the process of updating the shape of GSs using the design velocity field

$\boldsymbol{V}$ determined from Eq. (15). The shape optimization process for frame structure of GSs is constructed by repeating the vibration analysis, calculation of the shape gradient functions, velocity analysis, and shape updating. The vibration analysis and velocity analysis are carried out using a standard commercial FEM code.

We also consider the repeated eigenvalue problem in this study. When the repeated eigenvalue problem happens, we change the objective and constraint functions to the summation forms as shown in the first term of the right side of Eq. (17). The notation r (≥ 2) denotes the multiplicity of the repeated eigenvalues. The occurrence of repeated eigenvalue is judged with a tolerance δ, and we set $\delta = \{_{-0}^{+0.02\lambda^{(1)}}$ in the present work.

$$\begin{aligned} &L((\boldsymbol{w}^{(1)},\boldsymbol{\theta}^{(1)}),\cdots,(\boldsymbol{w}^{(r)},\boldsymbol{\theta}^{(r)}),(\bar{\boldsymbol{w}}^{(1)},\bar{\boldsymbol{\theta}}^{(1)})\cdots,(\bar{\boldsymbol{w}}^{(r)},\bar{\boldsymbol{\theta}}^{(r)}),\Lambda) \\ &=\sum_{k=1}^{r}\left\{-\lambda^{(k)}+\lambda^{(k)}b((\boldsymbol{w}^{(k)},\boldsymbol{\theta}^{(k)}),(\bar{\boldsymbol{w}}^{(k)},\bar{\boldsymbol{\theta}}^{(k)}))-a((\boldsymbol{w}^{(k)},\boldsymbol{\theta}^{(k)}),(\bar{\boldsymbol{w}}^{(k)},\bar{\boldsymbol{\theta}}^{(k)}))\right\}+\Lambda(M-\hat{M}) \end{aligned} \tag{17}$$

In an analogous way, the shape gradient functions should be replaced as

$$G_1=\sum_{k=1}^{r}G_1^{(k)},\quad G_2=\sum_{k=1}^{r}G_2^{(k)} \tag{18}$$

3 OPTIMIZATION RESULTS OF GRAPHENE SHEETS

In order to evaluate the shape optimization process of GSs for enhancing their fundamental frequency, we carry out two numerical examples to optimize the shape of a graphene nanoribbon (GNR) and an irregular GS. The volume constraint is set to be $M \leq \hat{M} = 1.005 M_{\text{initial}}$, where M_{initial} is the initial volume of GSs. It should be noted that the constraint conditions are set as 1 (x_1 direction), 2 (x_2 direction), 3 (x_3 direction), 4 (θ_1 direction), 5 (θ_2 direction) and 6 (θ_3 direction) in the present work.

3.1 Shape optimization of a graphene nanoribbon

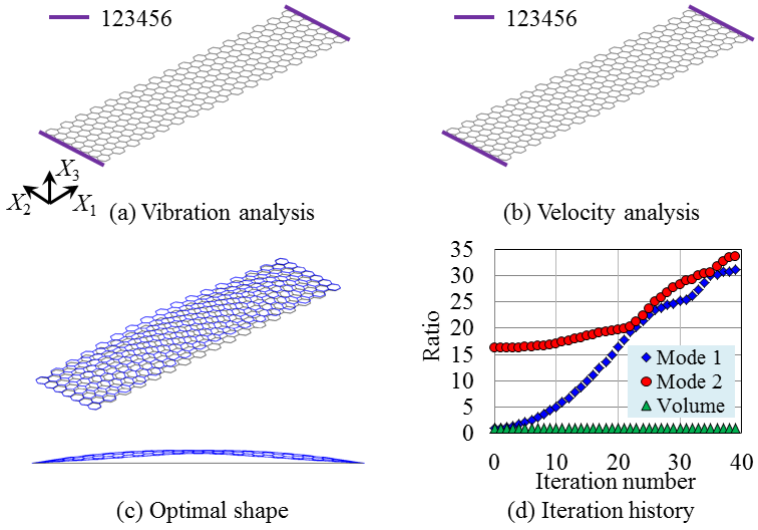


Figure 2: Shape optimization of a graphene nanoribbon.

In the first example, we design the shape a GNR shown in Fig. 2 with 510 carbon atoms and 724 equivalent C-C beams. As shown in Fig. 2 (a) and (b), two ends of the GNR are con-

strained in 123456 in both of the vibration analysis and the velocity analysis. Furthermore, to assure that all carbon atoms are varied in the X_3 direction, all of the remained nodes in velocity analysis are constrained in 12. We carry out shape optimization of the GNR by using the developed shape optimization process, and show the optimal shape and the iteration history in Fig. 2 (c) and (d), respectively. With respect to the optimal shape shown in Fig. 2 (c), the middle of the GNR becomes convex to enhance its fundamental frequency, and the optimal shape of the GNR is smooth, which is a feature of the free-form shape optimization method. In the iteration history shown in Fig. 2 (d), the fundamental frequency (mode 1) is enhanced smoothly while satisfying the volume constraint $M \le \hat{M} = 1.005 M_{\text{initial}}$, and the repeated eigenvalue problem occurs from the 22th iteration. This shape optimization process converges at the 39th iteration, and the fundamental natural frequency (mode 1) is enhanced to 31.15 times to the initial shape.

3.2 Shape optimization of an irregular graphene sheet

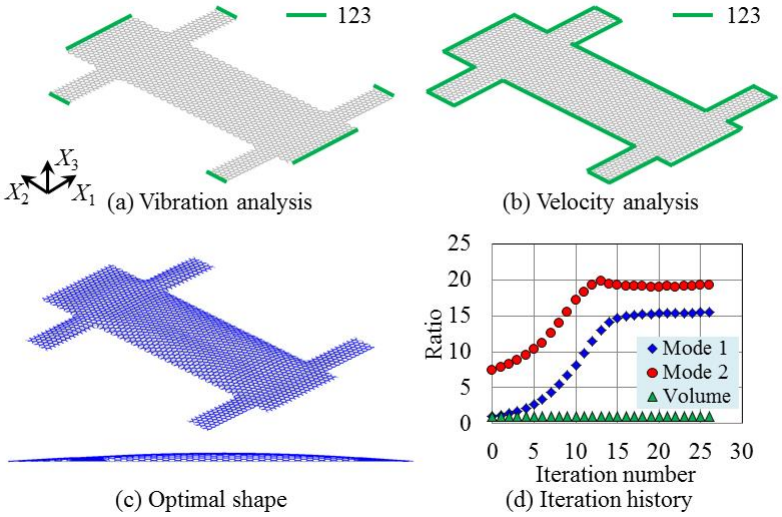


Figure 3: Shape optimization of an irregular graphene sheet.

We design an irregular GS shown in Fig. 3 containing 3194 carbon atoms and 4636 equivalent C-C beams in the example 2. We show the constraint condition of the vibration analysis and the velocity analysis in Fig. 3 (a) and (b), respectively. Using the proposed shape optimization process, the optimal shape of the GS is shown in Fig.3 (c). From the iteration history shown in Fig. 3 (d), the repeated eigenvalue problem does not occurs in this example, and the fundamental frequency of the GS is enhanced to 15.45 times to the initial shape in the 26th iteration, while the volume constraint is satisfied.

4 CONCLUSIONS

In this work, we developed a shape optimization system of GSs for the fundamental frequency maximization problem. At first, based on the MM method, GSs were modeled as frame structures composed of Timoshenko beams. Next, the fundamental frequency maximization problem of the frame structure of GSs was formulated. We adopted the free-form shape optimization method for frame structures to achieve the optimal shapes of GSs. This method has advantages that the optimal shapes of GSs could be determined smoothly and without requiring shape design parameterization. The objective of the shape optimization process was to maximize the fundamental frequency of GSs under the volume constrain. We also considered

the repeated eigenvalue problem in the present work. Two design examples were carried out to confirm the effectiveness of the developed shape optimization process, and the results showed that the obtained optimal shapes in both of the two design examples were smooth and the fundamental frequency of each example was enhanced significantly.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 15K21485 and grants in-aid from the Research Center of Smart and Tough Machines at Toyota Technological Institute.

REFERENCES

[1] A. Shakouri, T.Y. Ng, R.M. Lin, A new REBO potential based atomistic structural model for graphene sheets. *Nanotechnology*, **22**, 295711, 2011.

[2] J.B. Wang, X.Q. He, S. Kitipornchai, H.W. Zhang, Geometrical nonlinear free vibration of multi-layered graphene sheets. *Journal of Physics D: Applied Physics*, **44**, 135401 2011.

[3] J.X. Shi, Q.Q. Ni, X.W. Lei, T. Natsuki, Nonlocal vibration analysis of nanomechanical systems resonators using circular double-layer graphene sheets. *Applied Physics A*, **115**, 213-219, 2014.

[4] J.X. Shi, Q.Q. Ni, X.W. Lei, T. Natsuki, Nonlocal vibration of embedded double-layer graphene nanoribbons in in-phase and anti-phase modes. *Physica E: Low-dimensional Systems and Nanostructures*, **44**, 1136-1141, 2012.

[5] T. Georgiou, L. Britnell, P. Blake, R.V. Gorbachev, A. Gholinia, A.K. Geim, C. Casiraghi, K.S. Novoselov, Graphene bubbles with controllable curvature. *Applied Physics Letters*, **99**, 093103, 2011.

[6] J. Lu, A.H. Castro Neto, K.P. Loh, Transforming moiré blisters into geometric graphene nano-bubbles. *Nature Communications*, **3**, 823, 2012.

[7] J.X. Shi, M. Shimoda, Shape optimum design of graphene sheets. *Proceedings of 11th World Congress on Structural and Multidisciplinary Optimisation (WCSMO-11)*, Sydney, Australia, June 7-12, 2015.

[8] A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard III, W.M. Skiff, UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, **114**, 10024-10035, 1992.

[9] J.X. Shi, T. Natsuki, X.W. Lei, Q.Q. Ni, Equivalent Young's modulus and thickness of graphene sheets for the continuum mechanical models. *Applied Physics Letters*, **104**, 223101, 2014.

[10] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, **117**, 5179-5197, 1995.

[11] M. Shimoda, Y. Liu, T. Morimoto, Non-parametric free-form optimization method for frame structures. *Structural and Multidisciplinary Optimization*, **50**, 129-146, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOPOLOGY OPTIMIZATION WITH AN IMPLICIT FUNCTION AND PARAMETERIZED CUTTING SURFACE

Peter D. Dunning[1]

[1] University of Aberdeen
School of Engineering, University of Aberdeen, Aberdeen, AB24 3UE, Scotland
e-mail: peter.dunning@abdn.ac.uk

Keywords: Topology optimization, implicit functions, mathematical programming, design space reduction.

Abstract. *This paper introduces a new implicit function based method for topology optimization that can: obtain solutions with smooth boundaries, be solved using standard mathematical programming methods and reduce the number of design variables. Using implicit functions for topology optimization is attractive because the solutions have clearly defined, smooth boundaries. Most current methods use the zero level-set of the implicit function to define the boundary. The implicit function is then modified during optimization to move the boundary location and connectivity. The new approach proposed in this paper abandons the zero level-set idea and instead uses a fixed signed-distance implicit function. The definition of the boundary from the fixed implicit function is then modified during optimization. This is achieved by using a cutting surface and defining the boundary as the intersection of the cutting surface and signed-distance function. The cutting surface is parameterized and the parameters become the design variables during optimization. Thus, the optimization problem can be solved using mathematical programming and the number of parameters used to define the cutting surface is less than the number of elements in the analysis mesh. The new method is demonstrated using minimization of compliance, minimization of volume and complaint mechanism problems. The results show that the method can obtain good solutions to well-known problems with smooth, clearly defined boundaries and that this can be achieved using significantly fewer design variables compared with element-based methods.*

1 INTRODUCTION

The goal of structural topology optimization is to simultaneously optimize the size, shape and layout (or topology) of a structure. One benefit of this approach, compared with size or shape optimization, is that more efficient designs can be obtained because the optimizer is not constrained by a pre-determined layout.

Many topology optimization methods have emerged over the past 30 years, which have been successfully applied to many applications [1, 2]. There are two main categories of method: those using element-wise design variables, such as the Solid Isotropic Material with Penalization method [3] and Evolutionary Structural Optimization [4], and those using the zero-level set of an implicit functions to represent the boundary [5].

Topology optimization using an implicit function has several advantages over element-based methods, such as solutions with smooth boundaries and avoiding some numerical artifacts, such as checkerboard patterns [5]. Several methods have been developed to exploit these desirable qualities for topology optimization. Some methods use shape derivatives to move the implicit boundary by solving a PDE, such as the conventional level-set and phase-field methods [6, 7]. These approaches often converge slowly, and can require special techniques for solving constrained optimization problems [8]. An alternative approach is to parameterize the implicit function such that the problem can be solved using mathematical programming. This is attractive because advanced algorithms for efficiently solving constrained optimization problems can be employed. Some existing implicit function parameterization methods include, radial basis functions [9], geometric shapes [10], Fourier series [11] and finite element shape functions [12]. Also, projection methods offer similar features [13], as the solutions can have smooth boundaries and the problem can be solved using mathematical programming.

Another potential benefit of using parameterized implicit functions for topology optimization is that the number of design variables may be reduced, compared with methods using element-wise deign variables. This is because the design variables are not tied to the finite element discretization used for structural analysis. Reducing the number of variables reduces the problem size, which may result in faster convergence. Furthermore, as the number of design variables decreases optimizers that use full Hessians to speed up convergence, such as IPOPT, become more viable [14]. These advantages are particularly useful for large 3D problems, where a typical element-based method can use 10^6 (or more) design variables [15].

In this paper a new idea for using implicit functions for topology optimization is explored. The idea combines a signed-distance implicit function with a cutting surface. The cutting surface is parameterized and the parameters then become the design variables. The new method is introduced in Section 2, with numerical implementation issues discussed in Section 3, followed by examples in Section 4 and then conclusions.

2 CUTTING SURFACE PARAMETERIZATION

First, an initial design is chosen and mapped onto a design domain using an implicit signed-distance function, where the boundary is the zero level-set and a positive value indicates a point inside the structure, Figure 1. The structure is then defined by the scalar value of an implicit function:

$$\begin{cases} \phi(x) < 0 \ , \ x \notin \Omega \\ \phi(x) = 0 \ , \ x \in \Gamma \\ \phi(x) \geq 0 \ , \ x \in \Omega \end{cases} \tag{1}$$

where $\phi(x)$ is the implicit function, x is a point in the design domain (Ω_D), Ω and Γ are the structural domain and boundary, respectively.

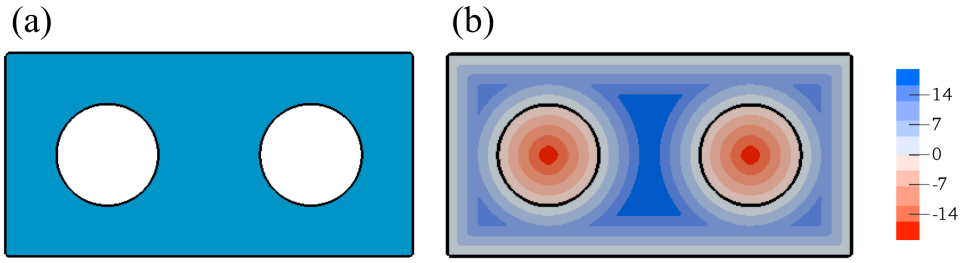


Figure1. a) Original structure, b) Contours of the signed distance function.

Another way to interpret this representation of the structure is that the boundary is the zero-level set of a higher dimensional implicit function. Therefore, we can say that the boundary is the intersection of a plane (defined by all points where $\phi = 0$) and the implicit function surface in the higher dimension. This is shown in Figure 2a for the 2D structure in Figure 1. To change the position of the structural boundary we can fix the position of the intersecting plane and change the implicit function, which is approach in most level-set topology optimization methods [5]. Alternatively, we could change the position of the intersecting plane. This second approach is demonstrated in Figure 2b, where the intersecting plane has been moved, resulting in the structural boundary moving position, as shown in Figure 2c.

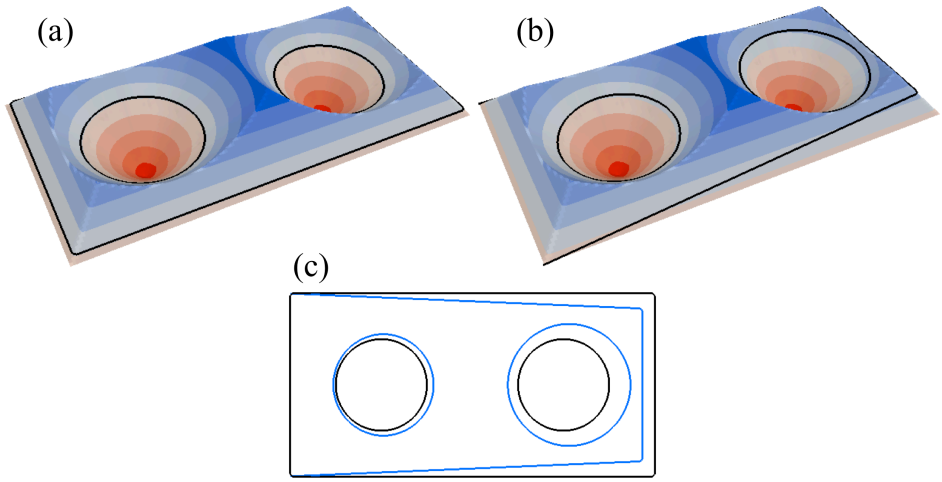


Figure 2. a) Implicit function surface with zero level-set contour, b) new position of the intersecting plane, resulting in a change of the structural boundary, c) comparison of structure boundaries: zero level-set (black) and new position of intersecting plane (blue).

The aim of this paper is to create a topology optimization method that exploits the moving intersecting plane mechanism for changing the position and connectivity of the structural boundary. The position of a plane in 3D can be described by 3 variables and we could proceed to develop an optimization method for 2D structures that uses just 3 variables. However, the design space would be very limited and thus this approach is not practical.

To provide more design freedom, the intersecting plane is replaced by a cutting surface that is defined in the higher dimension. Therefore, the method introduced in this paper uses two scalar valued functions: the implicit signed-distance function of the initial structure, $\alpha(x)$,

and the cutting surface, $\beta(x)$. The boundary of the structure is then defined by the set of points where the value of the signed-distance function equals the value of the cutting surface. Alternatively, the structural boundary is defined as the zero level-set of the implicit scalar-valued function that is the sum of the signed-distance and cutting surface functions. Thus, the definition of $\phi(x)$ in Eq. (1) can be written as:

$$\phi(x) = \alpha(x) + \beta(x) \tag{2}$$

The definition of the cutting surface is parameterized and the parameters become the design variables during optimization. There are several possible methods to parameterize the cutting surface, such as radial basis functions, polynomials or spline functions. In this paper, the cutting surface is parameterized using finite element shape functions, which is detailed in the next section.

During optimization, the signed-distance function remains fixed and the structure boundary changes position (and possibly connectivity) as the parameters of the cutting surface change. However, the design space is then limited by the current signed-distance function. Therefore, once an optimal cutting surface has been found for the current signed-distance function, a new signed-distance function is generated from the boundary of the current structure. The optimal cutting surface for the new signed-distance function is then found. This leads to an optimization method with an inner and outer loop. The aim of inner loop is to find the optimal cutting surface for the current signed-distance function. The outer loop then generates a new signed-distance function from the new structural boundary (as defined by the old signed-distance function and optimal cutting surface). At the start of the inner loop, the cutting surface is initialized as: $\beta(x) = 0$ (identical to the zero level-set), so that the structure boundary is the same as the optimum found by the previous inner loop (or initial structure).

In this paper, the inner loop optimization problem is solved by efficient gradient-based mathematical programming methods. Therefore, at least the first order derivatives of the objective and constraint functions need to be computed. It is usually most efficient to use analytical derivatives. Thus, an expression for the derivatives in terms of the cutting surface parameters is required. Structural quantities of interest can often be stated as integrals of a functional over the design domain, for example, compliance:

$$C\left(u,\phi\left(\alpha,\beta(\mathbf{b})\right)\right) = \int_{\Omega_D} c(u,x)\cdot H\left(\phi\left(x,\alpha,\beta(\mathbf{b})\right)\right)dx \tag{3}$$

where $C(u, \phi)$ is the total compliance of the structure, u is the displacement vector, $c(u, x)$ is the dot product of the stress and strain tensors at point x, $\mathbf{b}$ is a vector of the cutting surface parameters (the design variables) and $H(\phi)$ is the Heaviside function:

$$H(\phi) = \begin{cases} 1 & \phi \geq 0 \\ 0 & \phi < 0 \end{cases} \tag{4}$$

To compute the integral in Eq. (3) numerically the design domain is usually discretized using finite elements:

$$C\left(u,\phi\left(\alpha,\beta(\mathbf{b})\right)\right) = \sum_{e=1}^{N}\sum_{g=1}^{M_e} c\left(u_{e,g}\right)\cdot H\left(\phi\left(\alpha_{e,g},\beta_{e,g}(\mathbf{b})\right)\right)\cdot w_{e,g} \tag{5}$$

where N is the total number of elements, M_e is the number of integration points for element e and $w_{e,g}$ is an integration point weight. Taking the derivative of Eq. (5) with respect to design variable b_i, gives:

$$\frac{dC}{db_i} = \sum_{e=1}^{N}\sum_{g=1}^{M_e}\left[\frac{\partial c\left(u_{e,g}\right)}{\partial u_{e,g}}\cdot\frac{\partial u_{e,g}}{\partial b_i}\cdot H\left(\phi_{e,g}\right) + c\left(u_{e,g}\right)\cdot\frac{\partial H\left(\phi_{e,g}\right)}{\partial \phi_{e,g}}\cdot\frac{\partial \phi_{e,g}\left(\beta_{e,g}\right)}{\partial \beta_{e,g}}\cdot\frac{\partial \beta_{e,g}}{\partial b_i}\right]\cdot w_{e,g} \tag{6}$$

Therefore, the derivative of the compliance function (and many functions of interest in structural design) requires the derivative of the Heaviside function with respect to the implicit function, which is the Dirac delta function:

$$\frac{\partial H(\phi)}{\partial \phi} = \delta(\phi) = \begin{cases} 1 & \phi = 0 \\ 0 & else \end{cases} \tag{7}$$

Therefore, the derivative in Eq. (6) is discontinuous and only non-zero at the boundary of the structure (where $\phi = 0$). Thus, only integration points that lie exactly on the boundary will contribute to the derivative. The discontinuous nature of the derivatives using the exact Heaviside approach, Eq. (7), makes it difficult for gradient-based optimizers to find a solution. To avoid this problem, the Heaviside function can be approximated with a smooth function and this is the approach used in this paper. The choice of smoothed Heaviside function and inner loop optimization are discussed in Section 3.

3 NUMERICAL IMPLEMENTATION

3.1 Cutting surface parameterization

The proposed method uses two meshes that cover the design domain: one to discretize the cutting surface (cutting surface mesh) and one to perform the finite element analysis (analysis mesh), as shown in Figure 3.

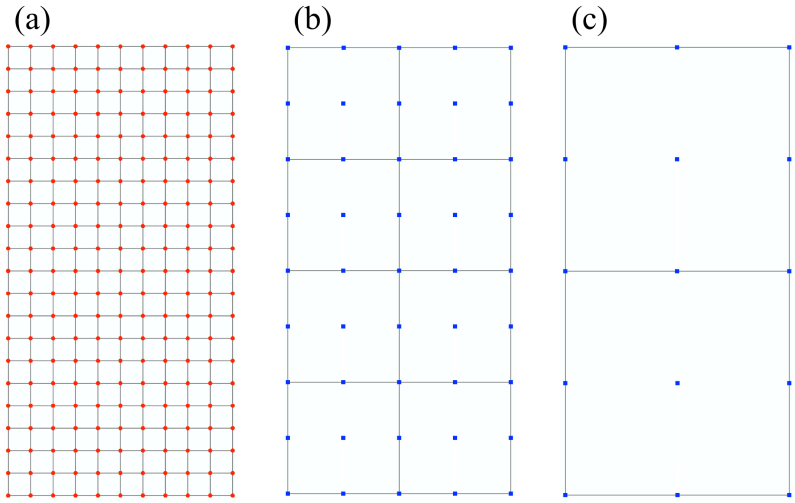


Figure 3. a) Analysis mesh using 200 Q4 elements, b) Cutting surface mesh using Q9 elements and mesh ratio 5 (45 design variables), c) cutting surface mesh with mesh ratio 10 (15 design variables).

The value of $\beta(x)$ is defined at nodes in the cutting surface mesh and interpolated using polynomial shape functions. The nodal values of $\beta(x)$ are then the design variables **b**. The implicit function, ϕ, and signed-distance function, α, are discretized on the analysis mesh. Values of ϕ at analysis element integration points are obtained by interpolating ϕ using the analysis element shape functions. Equation (2) can be written in terms of the discretized variables:

$$\boldsymbol{\phi} = \boldsymbol{\alpha} + \mathbf{A} \cdot \mathbf{b} \tag{8}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ are vectors of the discretized implicit and signed-distance functions, respectively, and **A** is a coefficient matrix that maps the cutting surface values to the nodes of the analysis mesh. This matrix remains constant during optimization and is only computed once.

Typically, the analysis and cutting surface meshes are different and there are advantages to decoupling the analysis discretization and design parameterization, such as reducing the number of design variables [16]. If the same mesh is used for both analysis and cutting surface, then **A** is the identity matrix and the method resembles one using nodal design variables. To reduce the number of design variables, the cutting surface mesh uses fewer elements than the analysis mesh. To aid future discussion, the ratio of analysis elements to cutting surface elements in one direction is called the "mesh ratio."

In this paper, 4 node bilinear elements (Q4) are used to discretize the analysis mesh and second order polynomial elements with 9 terms (Q9) are used to discretize the cutting surface, Figure 3. Thus, in this case, a mesh ratio of 2 results in **A** becoming the identity matrix.

Finally, the signed-distance function is computed before the start of each inner loop using the fast marching method developed by Adalsteinsson and Sethian [17].

3.2 Smoothed Heaviside function

In this paper a polynomial function is used to approximate the Heaviside function [026]:

$$H(\phi) = \begin{cases} \rho_{\min} & , \ \phi < -\Delta \\ \dfrac{3(1-\rho_{\min})}{4}\left(\dfrac{\phi}{\Delta} - \dfrac{\phi^3}{3\Delta^3}\right) + \dfrac{1+\rho_{\min}}{2} & , \ -\Delta \le \phi < \Delta \\ 1 & , \ \phi \ge \Delta \end{cases} \tag{9}$$

where Δ is the smoothing length of the approximation and $\rho_{\min}$ is a parameter that defines the stiffness of the void region, in a similar way to the minimum density value in an element-based topology optimization method.

One of the benefits of using an implicit function for topology optimization, compared with element-based methods, is that the solution has smooth, well-defined boundaries. However, this is partially lost when using a smoothed Heaviside function, as the boundary becomes effectively blurred over the smoothing length [5]. To counter this, two strategies are proposed. Firstly a small smoothing length is used throughout the optimization, ideally smaller than the element edge length in the analysis mesh. However, this strategy may cause convergence problems with some examples, especially those that have a small volume ratio. Thus, a second strategy is proposed where the optimization starts with a large smoothing length that is then reduced each time the outer loop convergences. The optimization continues until a converged solution using a lower limit of the smoothing length is obtained. This strategy is similar to continuation techniques used in density-based element methods to reduce the number of grey elements [3].

3.3 Inner loop optimization problem

The inner loop optimization problem can be stated as:

$$\begin{aligned} &\underset{\mathbf{b}}{\text{Minimize}} : f\left(\phi(\mathbf{b})\right) \\ &\text{Subject to} : g_i\left(\phi(\mathbf{b})\right) \le 0 \quad , i = 1 \cdots m \\ &\mathbf{b}_{\min} \le \mathbf{b} \le \mathbf{b}_{\max} \end{aligned} \tag{10}$$

where $f(\phi)$ is the objective function, $g_i(\phi)$ the constraint functions, m the total number of constraints, $\mathbf{b}_{\min}$ and $\mathbf{b}_{\max}$ are the side limits on the design variables.

The solution to the inner loop problem is affected by the choice of optimizer, design variable side limits and convergence criteria. In this paper the inner loop problem is solved using the method of moving asymptotes, as implemented in the package NLOPT [18]. The side limits are defined from the maximum and minimum signed-distance values of the current structure, with considerations for the boundary of the design domain, Γ_D , and fixed, non-designable regions within the design domain, Ω_{fixed}. First, a parameter is obtained from the discrete values of the signed-distance function:

$$\hat{\alpha} = \max\left\{\min\left\{\max\left\{\alpha_i\right\}, -\min\left\{\alpha_i\right\}\right\},\ 2h\right\} \tag{11}$$

The side limits for a design variable are then defined as:

$$b_{i,\min} = \begin{cases} 0 & if\ b_i \in \Omega_{fixed} \\ -\hat{\alpha} & else \end{cases} \qquad b_{i,\max} = \begin{cases} -\phi_i & if\ b_i \in \Gamma_D \\ \hat{\alpha} & else \end{cases} \tag{12}$$

Three convergence criteria are used for the inner loop: the maximum number of iterations, the relative change of the objective function and relative change of the design variables. The relative change criteria are both set to 10^{-3} and the maximum number of iterations to 10. The inner loop stops when any of these criteria are met. The criteria are intentionally slack, as this reduces the number of inner loop iterations spent fine tuning the structure, whereas quicker progress is often made by obtaining an approximate solution and generating a new signed-distance function for the next inner loop.

4 EXAMPLES

4.1 Minimization of compliance

First, the method introduced in this paper is used to solve two classic problems. The objective is to minimize compliance, with an upper limit on the total volume. The first problem is a cantilever, ratio 2:1, shown in Figure 4a, with volume constraint set to 50% of the design domain. The second is a Michell arch, shown in Figure 4b, with a volume constraint of 40%. The grey areas show fixed regions within the design domain. For both examples, the material properties are Young's modulus 1.0 and Poisson's ratio 0.3, and the load magnitude is 10 units. The initial design for both examples is shown in Figure 4c. The parameters for the smoothed Heaviside, Eq. 9, are: $\Delta = 0.25h$ and $\rho_{\min} = 10^{-6}$. Both examples use an analysis mesh composed of 160 × 80 unit sized plane stress elements. They are solved using different cutting surface meshes, defined by mesh ratios: 5, 10 and 20. This corresponds to problems with: 2145, 561 and 153 design variables, respectively.

The solutions for the cantilever are shown in Figure 5 and the solutions for the Michell arch are shown in Figure 6. The topologies and general shapes of the solutions compare well with the known analytical solutions and the solutions obtained using other methods [3, 8, 19]. For the cantilever, the solution using a mesh ratio of 5 is more topologically complex than the other solutions that use fewer cutting mesh elements. The compliance values for the three solutions are 5985, 6040 and 6131, for mesh ratios 5, 10 and 20, respectively. Thus, the objective value increases with a decrease in the number of design variables (or number of cutting surface elements). In this case using 153 design variables returns a compliance value 2.4% higher than the solution obtained using 2145 design variables. Also, for the mesh ratio 20 so-

lution, some parts of the structure boundary are not as straight as the boundaries of the other two solutions.

For the Michell arch, the topology of the main structural members is the same for all three solutions. The compliance values are 1095, 1090 and 1128, for mesh ratios 5, 10 and 20, respectively. Thus, there is a 3.5% increase in compliance for the mesh ratio 20 solution, compared with the mesh ratio 10 solution. Again, parts of the boundary for the mesh ratio 20 solution are not straight, which contributes to the increase in compliance.

Overall, these two examples demonstrate that the proposed method can obtain solutions to well-known compliance minimization problems. Also, there is a limit to how far the design space can be reduced (by reducing the number of cutting mesh elements), before the objective value and quality of the boundary become significantly affected. Note that, for these examples, a mesh ratio of 10 uses 23 times less design variables than the number of analysis mesh elements.

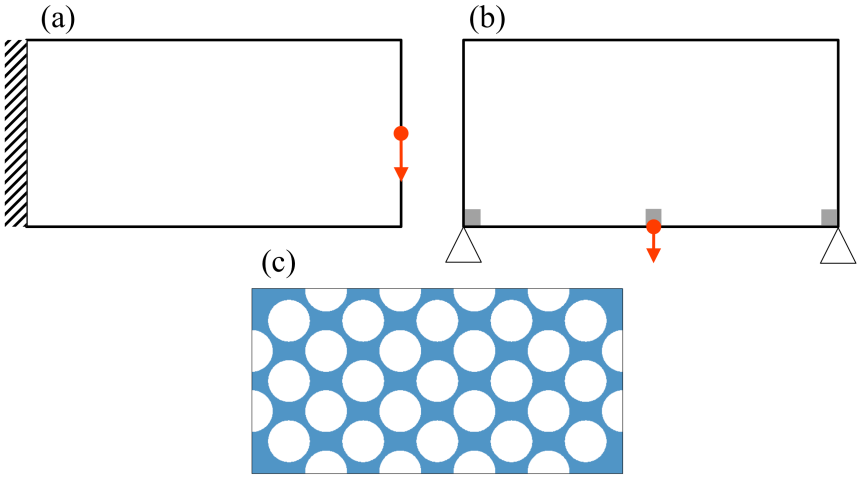


Figure 4. Examples. a) Cantilever, b) Michell arch, c) Initial design.

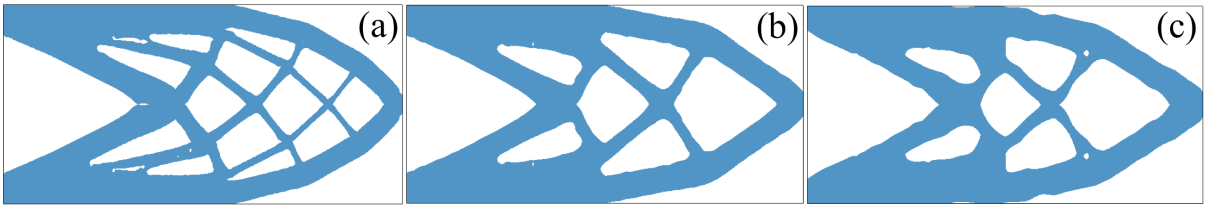


Figure 5. Cantilever minimization of compliance solutions. a) Mesh ratio 5, b) mesh ratio 10, c) mesh ratio 20.

Figure 6. a) Michell arch minimization of compliance solutions a) Mesh ratio 5, b) mesh ratio 10, c) mesh ratio 20.

4.2 Minimization of volume

The two examples used in the previous section for compliance minimization and now solved for the complimentary problem: minimization of volume, subject to an upper limit on compliance. For the cantilever, the upper limit is 6000 units and for the Michell arch the limit is 1100 units. These limits are approximately the optimal compliance values obtained in the previous section. Again, both problems are solved using different cutting surface meshes, defined by mesh ratios: 5, 10 and 20. All other parameters are the same as the previous section.

The solutions for the cantilever are shown in Figure 7 and the solutions for the Michell arch are shown in Figure 8. In all cases, a feasible solution is obtained. For the cantilever, the final volume values are 6353, 6454 and 6838, for mesh ratios 5, 10 and 20, respectively. Thus, there is a 7.6% increase in volume for the mesh ratio 20 solution, compared with the mesh ratio 5 solution. A similar trend is observed for the Michell arch, where the volume values are 5085, 5313 and 5597 for mesh ratios 5, 10 and 20, respectively. In this case the volume increase is 10.1%. The percentage increase in objective value is higher than that observed for the compliance minimization problems. This is perhaps because the compliance constraint is nonlinear and thus poses a more difficult problem to solve than one with a linear volume constraint. Again, the quality and definition of the boundary decreases as the number of design variables decreases.

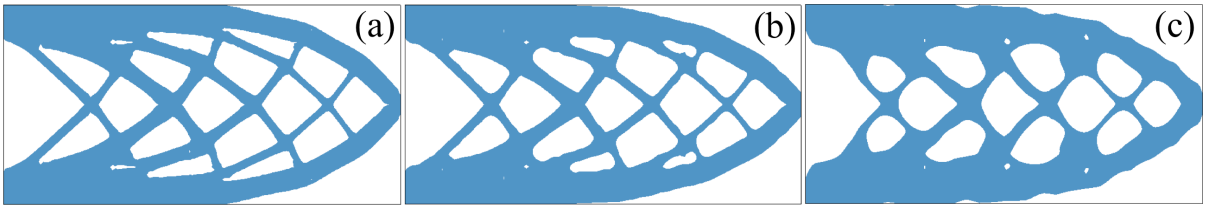


Figure 7. Cantilever minimization of volume solutions. a) Mesh ratio 5, b) mesh ratio 10, c) mesh ratio 20.

Figure 8. a) Michell arch minimization of volume solutions a) Mesh ratio 5, b) mesh ratio 10, c) mesh ratio 20.

4.3 Compliant mechanism

The new method is now used to solve an inverter compliant mechanism problem [14], as shown in Figure 9a. The objective is to maximize the output displacement, which is defined as positive in the opposite direction to the input force, with an upper limit on the total volume, set to 15% of the design domain. The material properties are Young's modulus 100.0 and Poisson's ratio 0.3, and the input load magnitude is 10 units. The stiffness of the input and output springs are: $k_{in} = 0.001$, $k_{out} = 1.0$. The initial design is shown in Figure 9b. The design domain is discretized using 160 × 160 plane stress elements, with element edge length 0.5.

For this example, a continuation strategy is used to adapt the smoothing length during optimization (as discussed in Section 3.2). The initial value is $\Delta = 2h$ and the lower limit is

0.25h, where h is the analysis mesh grid spacing. The smoothing length is halved each time the outer loop convergences. The value of $\rho_{\min} = 10^{-2}$.

The solutions for mesh ratios 5 and 10 (4225 and 1089 design variables, respectively) are shown in Figure 10. These correspond well with solutions for similar problems obtained using other methods [3, 20, 21]. The output displacement values are 1.37 and 1.25, for mesh ratios 5 and 10 respectively. Both solutions have the same topology, with small differences in shape. This is particularly noticeable near the input and output locations, where the mesh ratio 5 solution uses very thin structural members to create compliant hinges, whereas the hinges for the mesh ratio 10 solution are thicker. This difference results in the 9.8% increase in output displacement for mesh ratio 5, compared with mesh ratio 10. However, this increase is partly obtained by exploiting the numerical discretization, where a very compliant hinge is created around one node in the analysis mesh [22]. It would be interesting to add minimum length scale control to the proposed method to avoid this issue. The problem was also solved using a mesh ratio of 20. The solution was feasible (the volume constraint was met), but there was no physical link between the input and output locations, so it is not a practical design.

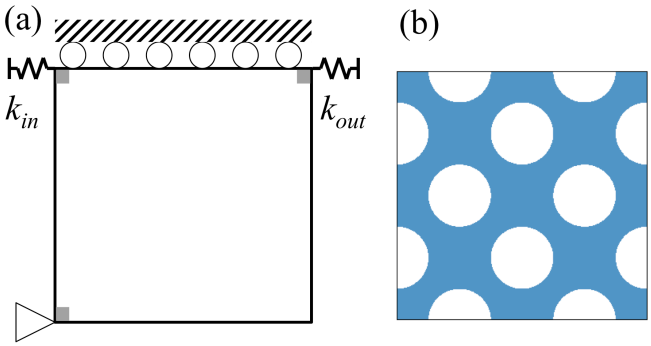


Figure 9. a) Inverter example, b) Inverter initial design.

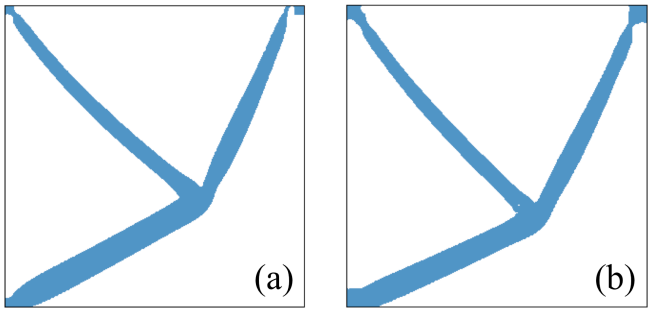


Figure 10. Inverter solutions. a) Mesh ratio 5, b) mesh ratio 10.

5 CONCLUSIONS

A new topology optimization method is proposed that uses an implicit signed-distance function and cutting surface, where the structure boundary is defined by their intersection. During optimization, the implicit function is fixed and the cutting surface modified to change the position and connectivity of the boundary. The cutting surface is parameterized using finite element shape functions and the nodal values become the design variables during optimization. The analysis and cutting surface meshes are different, where the cutting surface mesh uses fewer elements than the analysis mesh, thus reducing the number of design variables compared with a method using element-wise design variables. A critical aspect of the new method is the use of a smoothed Heaviside function for integrating the structural analysis ma-

trices. This approach avoids the discontinuous derivatives associated with an exact Heaviside approach. The benefits of the new method are that: solutions have clear, smooth boundaries, efficient mathematical programming methods can be employed and the number of design variables is reduced.

The new method is demonstrated by solving well-known minimization of compliance, minimization of volume and complaint mechanism problems. In each case, feasible, smooth boundary solutions are obtained that agree well with solutions obtained using other methods. However, there is a limit to how far the number of design variables can be reduced before the objective value and quality of the boundary become significantly affected. Despite this, good solutions to all problems are obtained using a number of design variables around 23 times fewer than the number of elements used for analysis.

REFERENCES

[1] J.D. Deaton, R.V. Grandhi, A survey of structural and multidisciplinary continuum topology optimization: post 2000. *Structural and Multidisciplinary Optimization*, **49**, 1-38, 2014.

[2] O. Sigmund, K. Maute, Topology optimization approaches: A comparative review. *Structural and Multidisciplinary Optimization* ,**48**(6), 1031-1055, 2013.

[3] M.P. Bendsoe, O. Sigmund, Topology optimization: Theory, Methods and Applications. Springer-Verlag, 2004.

[4] Y.M. Xie, G.P. Steven, Evolutionary structural optimization. Springer, 1997.

[5] N. P. van Dijk, K. Maute, M. Langelaar, F. van Keulen, Level-set methods for structural topology optimization: a review. *Structural and Multidisciplinary Optimization*, **48**, 437-472, 2013.

[6] M.Y. Wang, X. Wang, D. Guo, A level set method for structural topology optimization. *Computer Methods In Applied Mechanics and Engineering*, **192**(1-2), 227-246, 2003.

[7] G. Allaire, F. Jouve, A.M. Toader, Structural optimization using sensitivity analysis and a level-set method. *Journal of Computational Physics*, **194**(1), 363-393, 2004.

[8] P.D. Dunning, H.A. Kim. Introducing the sequential linear programming level-set method for topology optimization. *Structural and Multidisciplinary Optimization*, **51**(3), 631-643, 2015.

[9] S. Wang, M. Wang, Radial basis functions and level set method for structural topology optimization. *International Journal For Numerical Methods In Engineering*, **65**(12), 2060-2090, 2006.

[10] W. Zhang, J. Zhang, X. Guo, Lagrangian description based topology optimization - a revival of shape optimization. *Journal of Applied Mechanics*, **83**(4), 2016.

[11] A.A. Gomes, A. Suleman, Application of spectral level set methodology in topology optimization. *Structural and Multidisciplinary Optimization*, **31**(6), 430-443, 2006.

[12] N.P. van Dijk, M. Langelaar, F. van Keulen, Explicit level-set-based topology optimization using an exact Heaviside function and consistent sensitivity analysis. *International Journal for Numerical Methods in Engineering*, **91**(1), 67-97, 2012.

[13] J.K. Guest, Topology optimization with multiple phase projection. *Computer Methods in Applied Mechanics and Engineering*, **199**(1-4), 123-135, 2009.

[14] S. Rojas-Labanda, M. Stolpe, Benchmarking optimization solvers for structural topology optimization. *Structural and Multidisciplinary Optimization*, **52**(3), 527-547, 2015.

[15] S. Wang, E.D. Sturler, G.H. Paulino, Large-scale topology optimization using preconditioned Krylov subspace methods with recycling. *International journal for numerical methods in engineering*, **69**(12), 2441-2468, 2007.

[16] J.K. Guest, L.C. Smith Genut, Reducing dimensionality in topology optimization using adaptive design variable fields. *International journal for numerical methods in engineering*, 81: 1019–1045, 2010.

[17] D. Adalsteinsson, J. A. Sethian, The fast construction of extension velocities in level set methods. *Journal of Computational Physics*, **148**(1), 2-22, 1999.

[18] The NLopt nonlinear optimization package. http://ab-initio.mit.edu/nlopt. 2014.

[19] A.L. Gain, G.H. Paulino, A critical comparative assessment of differential equation-driven methods for structural topology optimization. *Structural and Multidisciplinary Optimization*, **48**(4), 685-710, 2013.

[20] O. Sigmund, On the design of compliant mechanisms using topology optimization. *Mechanics of Structures and Machines*, **25**(4), 493-524, 1997.

[21] Z. Luo, L. Tong, M.Y. Wang, S. Wang, Shape and topology optimization of compliant mechanisms using a parameterization level set method." *Journal of Computational Physics*, **227**(1), 680-705, 2007.

[22] T.A. Poulsen, A simple scheme to prevent checkerboard patterns and one-node connected hinges in topology optimization. *Structural and Multidisciplinary Optimization*, **24**(5), 396-399, 2002.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOPOLOGY OPTIMIZATION OF PERIODIC STRUCTURES FOR COUPLED ACOUSTIC-STRUCTURE SYSTEMS

William Martins Vicente[1], **Renato Picelli**[1], **Renato Pavanello**[1], **Yi Min Xie**[2]

[1] Department of Computational Mechanics, Faculty of Mechanical Engineering
University of Campinas
Rua Mendeleyev 200, 13083-860 Campinas, Brazil
vicente@fem.unicamp.br; picelli@fem.unicamp.br; pava@fem.unicamp.br

[2] Centre for Innovative Structures and Materials, School of Engineering
RMIT University
GPO Box 2476, Melbourne 3001, Australia
mike.xie@rmit.edu.au

Keywords: Topology optimization, Fluid-structure interaction, Frequency response, Periodic structures.

Abstract. *This work applies the topology optimization technique to an acoustic-structure coupled system with periodic geometry constraint in order to obtain the optimal layout of the design domain for the minimization of the pressure frequency response in the acoustic fluid. The displacement-pressure formulation (u–p) is used for the finite element analysis of the coupled system and external harmonic excitations are applied in the system. The design domain of the coupled system is considered to be composed of identical unit cells. A periodic geometry constraint is applied in the design domain considering the fluid-structure interaction and the objective function. Appling the modified bi-directional evolutionary structural optimization (BESO) technique to the system, the design domain is evolving towards the optimal topology of the unit cells through removing/adding material accordingly to the sensitivity analysis. The influence of the total number of unit cells composing the periodic structure and the aspect ratio of the unit cells are investigated in the minimization of the objective function. In order to show the capability and efficiency of the proposed formulation, two acoustic-structure systems are optimized for several cell configurations, different aspect ratios of the periodic unit cells and excitation frequencies.*

1 INTRODUCTION

In recent years, the optimization of multiphysics systems has been considered in several areas of the computational engineering. Concerning the acoustic-structure systems, the minimization of the pressure frequency response in a region of the fluid domain could be particularly important for engineering problems like noise reduction in interior of rooms, in passenger compartments of automobile, airplane, etc. Different topology optimization approaches, [1, 2, 3, 4, 5, 6, 7, 8], have been developed in the last decades and can be used to obtain a better performance of multiphysics systems.

The frequency response problems in multiphysics system have been investigated using different approaches for circumvent the difficulty of optimize systems having more than one material domain [9, 10, 11, 12].

Extending the traditional bi-directional evolutionary structural optimization (BESO) method to a three phases system (structure, fluid end void) and using the displacement-pressure formulation (*u–p*) for the finite element analysis of the coupled system, the problem concerning the interface between the domains during the optimization process have been studied [13, 14, 15]. Due the characteristics of the *u-p* formulation and the evolutionary optimization technique, the coupled system during the optimization procedure does not present intermediate densities of structure, thus, all phases, solid, void and fluid, as well as the interface between the domains are explicitly defined.

Concerning the optimization involving periodic structures, the topology optimization has been applied to find optimal design of the periodic cells using the based density approach, [16, 17] and more recently with the evolutionary technique, [18, 19, 20], considering static and dynamic problems.

Recently, the topology optimization using BESO technique was applied to minimize the frequency response considering the pressure [21] and displacement [22] for a range of frequency of excitation.

This work uses the modified BESO to fluid-structure coupled system to optimize acoustic-structure problems where the design domain is subject to a periodic constraint. The analysis of the system is carried out using the displacement-pressure formulation (*u–p*), where the solid domain is governed by the elasticity equation and the acoustic fluid domain by Helmholtz equation. The surface-coupling integral are used to fully coupled the domains.

This paper is organized as follows: Section 2 presents the acoustic-structure formulation used and the finite element modeling of the coupled system. In Section 3, the topology optimization problem for the pressure frequency response minimization is developed and the sensitivity analysis is carried out considering the periodicity of the unit cells in the design domain. Section 4 shows some numerical results reached using the proposed methodology. Finally, conclusions are drawn in Section 5.

2 ACOUSTIC-STRUCTURE INTERACTION

The fluid-structure interaction considered in this work is represented generically in Fig. 1, where Ω_s is the structure domain; Ω_f is the fluid acoustic domain and Γ_{sf} represents the interface between the domains.

All boundary conditions and the external force are also represented in Fig. 1. A detailed description about the formulation of the system can be found in the [21].

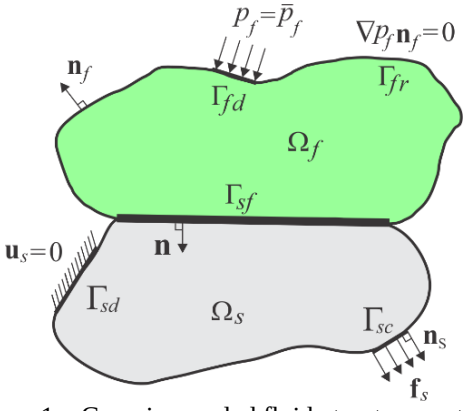


Figure 1 – Generic coupled fluid-structure system.

2.1 Equilibrium equations and boundary conditions

Considering the structural domain composed of a homogeneous and isotropic material and under a small linear deformation, the linear elastodynamic equation can be used to describe the continuum medium:

$$\nabla \cdot \boldsymbol{\sigma}_s - \rho_s \frac{\partial^2 \mathbf{u}_s}{\partial t^2} = 0 \quad \text{in} \quad \Omega_s \tag{1}$$

where $\nabla \cdot \boldsymbol{\sigma}_s$ is the divergence of the Cauchy stress tensor; $\mathbf{u}_s$ is the displacement vector field and ρ_s is the mass density. The subscript *s* denotes to structural variables.

The boundary conditions considered in the structural domain, Fig. 1, is the displacement constraint

$$\mathbf{u}_s = 0 \quad \text{in} \quad \Gamma_{sd} \tag{2}$$

the prescribed external load, $\mathbf{f}_s$, applied in the structure

$$\boldsymbol{\sigma}_s \mathbf{n}_s = \mathbf{f}_s \quad \text{in} \quad \Gamma_{sc} \tag{3}$$

and the pressure force produced by the fluid and transmitted to the structure at the interface between the domains

$$\boldsymbol{\sigma}_s \mathbf{n}_s = p_f \mathbf{n}_f \quad \text{in} \quad \Gamma_{sf} \tag{4}$$

The fluid medium can be described using the wave equation, assuming the fluid to be inviscid, irrotational and in a small translation:

$$\frac{1}{c_0^2} \frac{\partial^2 p_f}{\partial t^2} - \nabla^2 p_f = 0 \quad \text{in} \quad \Omega_f \tag{5}$$

where p_f represents the acoustic pressure scalar field and c_0 is the constant speed of sound in the fluid.

In the fluid acoustic domain, the boundary conditions considered, Fig. 1, is the prescribed pressure, $\bar{p}_f$, in the fluid

$$p_f = \bar{p}_f \quad \text{in} \quad \Gamma_{fd} \tag{6}$$

the rigid wall boundary condition

$$\nabla p_f \mathbf{n}_f = 0 \quad \text{in} \quad \Gamma_{fr} \tag{7}$$

and the kinematic compatibility of the normal displacements at the interface of the fluid and structural domains

$$\nabla p_f \mathbf{n}_f = \rho_f \frac{\partial^2 \mathbf{u}_s}{\partial t^2} \mathbf{n}_s \quad \text{in} \quad \Gamma_{sf} \tag{8}$$

where ρ_f is the mass density of the fluid medium. Them, the stated boundary value problem for acoustic-structure interaction is find $\mathbf{u}_s$ and p_f that satisfy the equations [1] - [8].

2.2 Finite element discretization

The finite element discretization of the system is reached through taking the weak form of the boundary value problem and applying the Galerking method in the weighted residual approximation, [23]. Considering the presented coupled system of equations, the finite element matrices are written as follows

$$\mathbf{K}_s = \int_{\Omega_s} \left(\nabla \mathbf{N}_s\right)^T \mathbf{D}_s \nabla \mathbf{N}_s dV \tag{9}$$

$$\mathbf{M}_s = \rho_s \int_{\Omega_s} \mathbf{N}_s^T \mathbf{N}_s dV \tag{10}$$

$$\mathbf{K}_f = \int_{\Omega_f} \left(\nabla \mathbf{N}_f\right)^T \nabla \mathbf{N}_f dV \tag{11}$$

$$\mathbf{M}_f = \frac{1}{c_0^2} \int_{\Omega_f} \mathbf{N}_f^T \mathbf{N}_f dV \tag{12}$$

$$\mathbf{f} = \int_{\Gamma_{sc}} \mathbf{N}_s^T \mathbf{f}_s dS \tag{13}$$

$$\mathbf{L}_{sf} = \int_{\Gamma sf} \mathbf{N}_s^T \mathbf{n} \mathbf{N}_f dS \tag{14}$$

where $\mathbf{K}_s$ and $\mathbf{M}_s$ are the stiffness and mass matrices of the structural domain; $\mathbf{K}_f$ and $\mathbf{M}_f$ are the stiffness and mass matrices of the fluid domain; $\mathbf{f}$ is the nodal vector of the external forces applied to structure; $\mathbf{L}_{sf}$ is the coupling matrix at the interface between the domains; $\mathbf{N}_s$ and $\mathbf{N}_f$ are the shape functions for the structural and fluid domains, respectively.

Letting ω be the harmonic excitation frequency of the system, the coupled fluid-structure system can be described in terms of the finite element matrices as:

$$\begin{bmatrix} \mathbf{K}_s - \omega^2 \mathbf{M}_s & -\mathbf{L}_{sf} \\ -\rho_f \mathbf{L}_{sf}^T \omega^2 & \mathbf{K}_f - \omega^2 \mathbf{M}_f \end{bmatrix} \begin{Bmatrix} \mathbf{u}_s \\ \mathbf{p}_f \end{Bmatrix} = \begin{Bmatrix} \mathbf{f} \\ 0 \end{Bmatrix} \tag{15}$$

where $\mathbf{u}_s$ and $\mathbf{p}_f$ are the nodal approximation of the frequency response of displacement and pressure, respectively.

3 EVOLUTIONARY OPTIMIZATION WITH PERIODIC CONSTRAINT

This work aims to find the optimum layout of the periodic unit cells in the design domain in order to minimize the pressure frequency response in an acoustic-structure coupled system. Following is presented the mathematical approach of the problem and the optimization procedure used in this work.

3.1 Optimization problem

A periodic system is exemplified in Fig. 2; in this generic 2D representation, the system is divided into $m = m_1 \times m_2$ equal unit cells. m_1 unit cells in horizontal direction and m_2 unit cell in vertical direction, a cell mode of $m_1 \times m_2$. The sensitivity number in the optimization problem is given in terms of the design variable $x_{i,j}$, where the subscript i is related to cell number and the subscript j refers to element number in the unit cell.

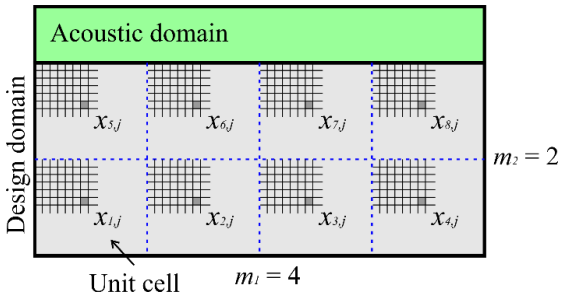


Figure 2 – Coupled system with the design domain composed of $m = 8$ identical unit cells, in the cell mode $m_1 \times m_2 = 4 \times 2$, with 4 unit cells along the horizontal direction and 2 unit cells along the vertical direction.

The optimization problem proposed here for minimizing the pressure frequency response for a periodic system can be stated as

Minimize: $$|\mathbf{p}_f| \tag{16}$$

Subject to: $$\begin{bmatrix} \mathbf{K}_s - \omega^2 \mathbf{M}_s & -\mathbf{L}_{sf} \\ -\rho_f \mathbf{L}_{sf}^T \omega^2 & \mathbf{K}_f - \omega^2 \mathbf{M}_f \end{bmatrix} \begin{Bmatrix} \mathbf{u}_s \\ \mathbf{p}_f \end{Bmatrix} = \begin{Bmatrix} \mathbf{f} \\ 0 \end{Bmatrix} \tag{17}$$

$$V_f - mV_i = 0 \tag{18}$$

$$V_i - \sum_{j=1}^{n} v_{i,j} x_{i,j} = 0 \tag{19}$$

$$x_{1,j} = x_{2,j} = x_{3,j} = \cdots = x_{m,j} \tag{20}$$

$$x_{i,j} \in \{0,1\} \quad (j = 1,\ 2,\ 3,\ \cdots,\ n) \quad (i = 1,\ 2,\ 3,\ \cdots,\ m) \tag{21}$$

where $|\mathbf{p}_f|$ is the mean pressure frequency response in a point or area of the fluid domain; Eq. 17 refers to the equilibrium equation of the coupled system; V_f is final prescribed volume fraction of the structural part in the design domain; V_i the volume fraction of the *i*th unit cell; $v_{i,j}$ is the volume fraction of the *j*th element in the *i*th unit cell, *n* is the total number of elements in the unit cell and $x_{i,j}$ is the design variable with 1 for structural elements and 0 for void or fluid elements. Due the periodic constraint, the design variable is enforced to have the same value for elements with the same position in the unit cell as indicate in Eq. 20.

3.2 Sensitivity analysis

Following the evolutionary optimization procedure, the sensitivity number of each element based on the objective function should be evaluated for all elements in the design domain. Then, the elements are ranked according to the they efficiency. To obtain this efficiency of each element, the equilibrium equation, Eq. 17, should be derived with respect to the design variable, $x_{i,j}$, for each element.

$$\frac{\partial \begin{bmatrix} \mathbf{K}_s - \omega^2 \mathbf{M}_s & -\mathbf{L}_{sf} \\ -\rho_f \mathbf{L}_{sf}^T \omega^2 & \mathbf{K}_f - \omega^2 \mathbf{M}_f \end{bmatrix} \begin{Bmatrix} \mathbf{u}_s \\ \mathbf{p}_f \end{Bmatrix}}{\partial x_{i,j}} = \frac{\partial \begin{Bmatrix} \mathbf{f} \\ 0 \end{Bmatrix}}{\partial x_{i,j}} \tag{22}$$

Based on [22]; introducing the location vector to identify de degree of freedom for the minimization of the frequency response, neglecting the alteration in fluid global matrices due the structural element remove, the Eq. 22 can be rewritten as

$$\frac{\partial u_k}{\partial x_{i,j}} = -\mathbf{u}_{s,k}^T \frac{\left(\mathbf{K}_s - \omega^2 \mathbf{M}_s\right)}{x_{i,j}} \mathbf{u}_s \tag{23}$$

where u_k is the frequency response in the specific degree of freedom of interesting to minimize and $\mathbf{u}_{s,k}$ is frequency response of the structural domain when only a unitary load is applied in the *k*th degree of freedom.

Appling the material interpolation proposed by [24], the sensitivity number, from Eq. 23, to minimize the proposed objective function can be state as

$$\alpha_{i,j} = \begin{cases} \left(\mathbf{u}_{s,k}^{i,j}\right)^T \left(p\mathbf{K}_s^{i,j} - \omega^2 \mathbf{M}_s^{i,j}\right) \mathbf{u}_s^{i,j} & x_{i,j} = 1 \\ \\ -\left(\mathbf{u}_{s,k}^{i,j}\right)^T \left(\omega^2 \mathbf{M}_s^{i,j}\right) \mathbf{u}_s^{i,j} & x_{i,j} = x_{\min} \end{cases} \tag{24}$$

where the superscripts i, j indicate that the matrices and vector are in the elemental level. *p* is the penalty exponent factor. Following the soft-kill procedure, the design variable is 1 for the structure elements and $x_{\min}$ for the void or fluid elements, in this work the $x_{\min} = 10^{-8}$.

Due the imposed periodicity of the unit cells, the elements should be removed/added at the same time in the same position of the unit cells. In this way, the updating of the unit cell design can be done choosing a representative unit cell. And the sensitive number of the *j*th element in the representative unit cell is calculated as the overall sensitivity number of all *j*th elements from the other unit cells with the same position.

$$\alpha_j = \sum_{i=1}^{m} \alpha_{i,j} \tag{25}$$

where α_j is elemental sensitivity number of the representative unit cell, which can be choose from any unit cell once all unit cells are identical.

3.3 Filtering and sensitivity history

In order to prevent the instability during the topology optimization a number of strategy can be applied. A heuristic procedure that helps to avoid checkerboard patterns is filtering the elemental sensitivity number found by Eq. 25, [8]. A spatial linear filter is used here based on

the procedure shown in [8]. Due the periodicity unit cells in the design domain the filter scheme can be applied only at the representative unit cell.

The filtering procedure starts calculating the nodal sensitivity number, α_l^n, which is the averaging of the sensitivity number of all connected element at the node. Then, the nodal sensitivity numbers are used to calculate the elemental smoothed sensitivity numbers, α_j^s, using a circular (for 2D, or spherical for 3D) sub-domain defined by a radius $r_{\min}$ and centered at the elements center. Finally, for each element, the nodes inside the element sub-domain have their nodal sensitivity numbers averaged into the element as follows

$$\alpha_j^s = \frac{\sum_{l=1}^{nod} w(r_{lj})\alpha_l^n}{\sum_{l=1}^{nod} w(r_{lj})} \tag{26}$$

where nod is the total number of nodes inside the sub-main; $w(r_{lj})$ is the linear weight factor defined as

$$w(r_{lj}) = r_{\min} - r_{lj} \quad (l = 1,\ 2,\ \cdots,\ nod) \tag{27}$$

where r_{lj} is the distance between the *j*th element centre and the node l.

Helping in the convergence and stability of the optimization procedure [25], a history averaging of the sensitivities is implemented and the final sensitivity elemental number, α_j^f, for the representative cell can be calculated as follows

$$\alpha_j^f = \frac{\left(\alpha_j^s\right)^m + \left(\alpha_j^s\right)^{m-1}}{2} \tag{28}$$

where *m* is the current iteration number.

3.4 Convergence criteria

After reached the final prescribed volume, the optimization procedure evolves until the convergence of the objective function. The prescribed error tolerance to define the convergence of the objective function is defined as [25]

$$error = \frac{|\sum_{i=1}^{N} (\mathbf{p}_f)^{m-i+1} - \sum_{i=1}^{N} (\mathbf{p}_f)^{m-N-i+1}|}{|\sum_{i=1}^{N} (\mathbf{p}_f)^{m-i+1}|} \leq \tau \tag{29}$$

where $(\mathbf{p}_f)^m$ is the value of the objective function in the *m*th iteration and *N* is the number of iterations involved in the convergence error estimation.

4 NUMERICAL RESULTS

The structural optimization of two acoustic-structure systems are presented in this section considering the periodicity of the design domain and applying the formulation developed in the previous sections. In the finite element discretization of the systems, the four-node quadrilateral elements are used for the both domains.

4.1 Example 1

In this example, the formulation proposed in this work is applied to a coupled acoustic-structure system shown in Fig. 3 for minimization of the acoustic pressure. Simulating a rotating machine, like rotary engine, turbine, electric motor, etc., a distributed harmonic load is applied to a clamped structure as presented in Fig. 3. The optimization procedure aims to minimize the acoustic pressure response in the point P_{out} generated by the vibration of the structure in the frequency of 50 Hz.

The structural domain is divided in two parts, the design domain which is considered composed by periodic unit cells and the non-design domain which is fixed and will not be modified during the optimization procedure. The non-design domain is shown in black in Fig. 3.

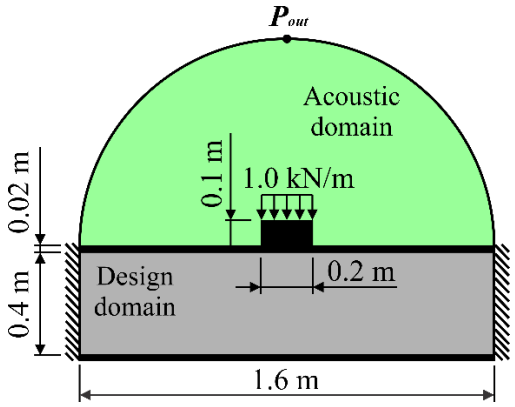


Figure 3 - Coupled system under a harmonic loading in the structure domain for minimization of the acoustic pressure at the top of the acoustic domain.

The follow material properties are adopted for the structural domain: Young's modulus $E = 200$ GPa, Poisson's ratio $\mu = 0.3$ and density $\rho = 7700$ kg/m^3. For the acoustic fluid domain, the air properties are considered: density $\rho = 1.2$ kg/m^3 and the speed of sound $c = 343$ m/s. The volume of the structural domain is kept constant during the optimization procedure. Four different cell modes configuration of the initial periodic design are optimized, 4 x 1, 8 x 2, 12 x 3 and 16 x 4. For a fair comparison of the unit cell final topology, the unit cell for all configuration has a mesh of 140 x 140 elements.

The BESO parameters [8] used are: $ER = 0.02$, $Ar = 0.02$, $\tau = 0.005$, $N = 5$ and $p = 3$. For all cell mode configurations, the filter radius is defined as 1/10 of the side length of the unit cell. The initial and final volume fraction of the design domain is defined as $V_f = V_i = 0.55$.

Figure 4 shows the pressure distribution for the initial configurations of the system and for the optimized design domain considering the frequency excitation of 50 Hz. The sensitivity number distribution in the optimized unit cell is also shown in Fig. 4 for all configurations of the periodic structure. The unit cell final topology converges for a symmetric pattern as the number of unit cell increases in horizontal and vertical directions. For the cell modes 12 x 3 and 16 x 4 the optimized unit cells show only a slight difference in the distribution of the material and can be considered as the converged pattern.

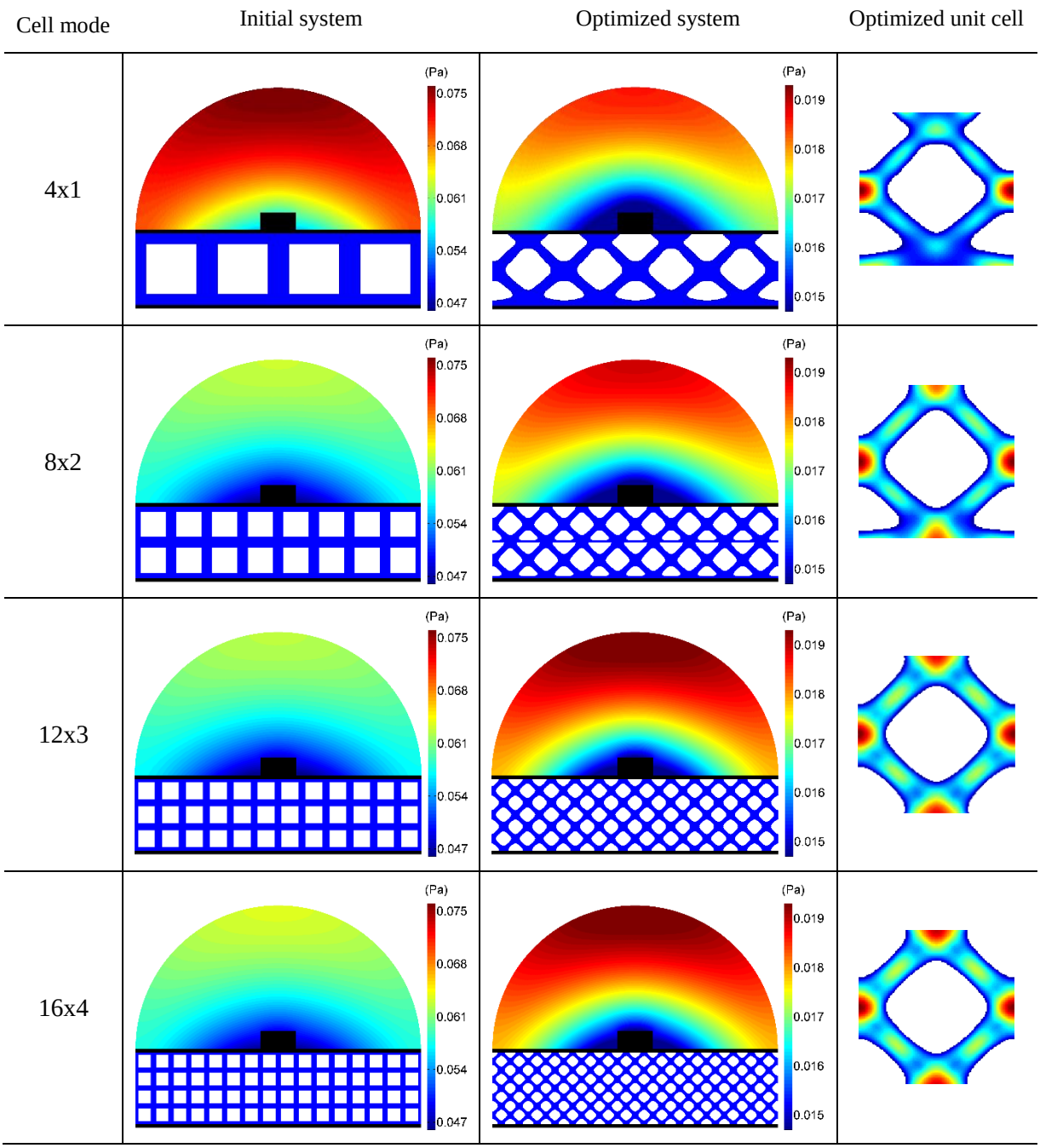


Figure 4 – Pressure distribution for the initial guess design and for the optimized configurations considering four different cell modes: 4 x 1, 8 x 2, 12 x 3 and 16 x 4. Evolution of the optimized unit cell for different cell modes.

The history of the objective function during the evolutionary optimization can be seen in Fig. 5 for all cases analyzed. The volume of the unit cell is kept constant for all cases during the optimization. Smooth convergences are observed for all configurations studied, the objective function curves do not present any peaks or discontinuities, the small value of the parameter *Ar* assure that only a small percentage of the structural material will be swapped in each iteration on the unit cells. The maximum number of iteration needed to satisfied to convergence criteria was 34 in the first case.

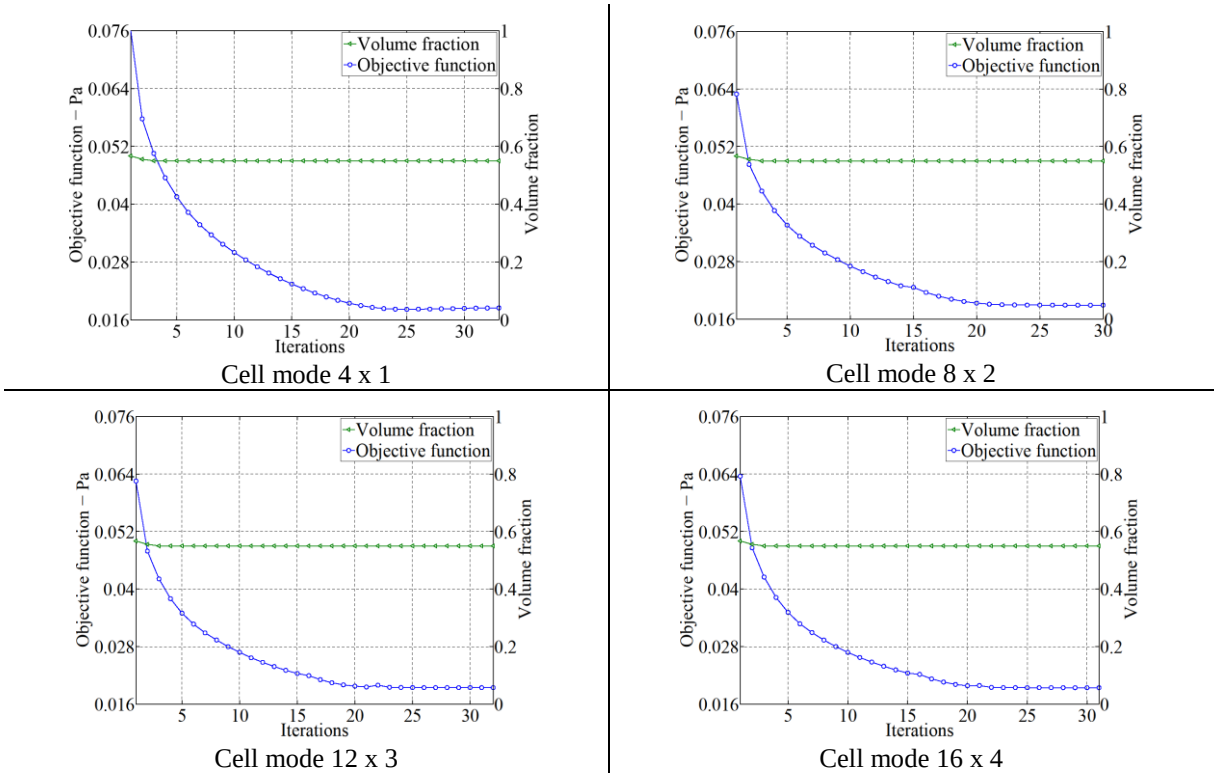
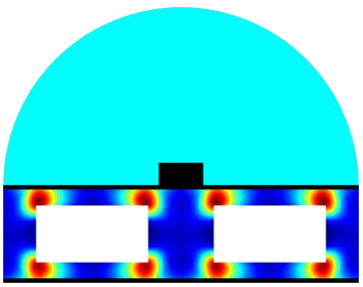
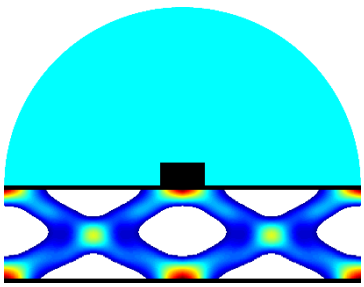
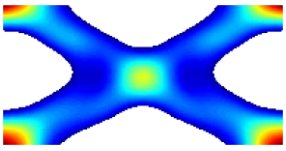
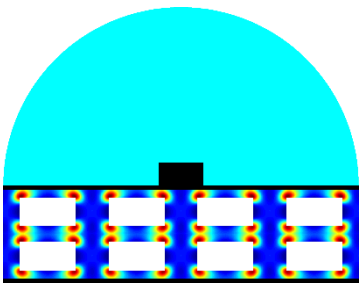
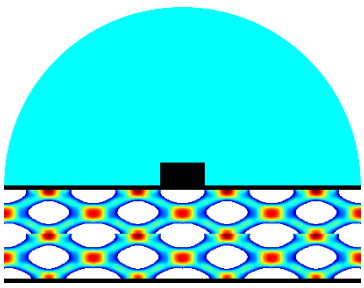
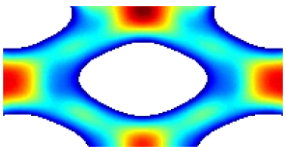


Figure 5 – Evolutionary histories of the objective function, the acoustic pressure in the point P_{out}, for the four different cell modes analyzed.

A new periodic unit cell is considered to be studied in the same coupled system for minimizing the acoustic pressure. The design domain is considered to be composed of periodic unit cell with the ratio length/high equal to 2. The same materials properties and the BESO parameters for the previous analyses are used. Figure 6 shows initials and optimized topologies of the periodic system with the new unit cell, 4 cell modes are presented 2 x 1, 4 x 2, 6 x 3 and 8 x 4. For all cases the mesh of the unit cell is discretized with 200 x 100 elements.

Cell mode	Initial system	Optimized system	Optimized unit cell
2 x 1			
4 x 2			

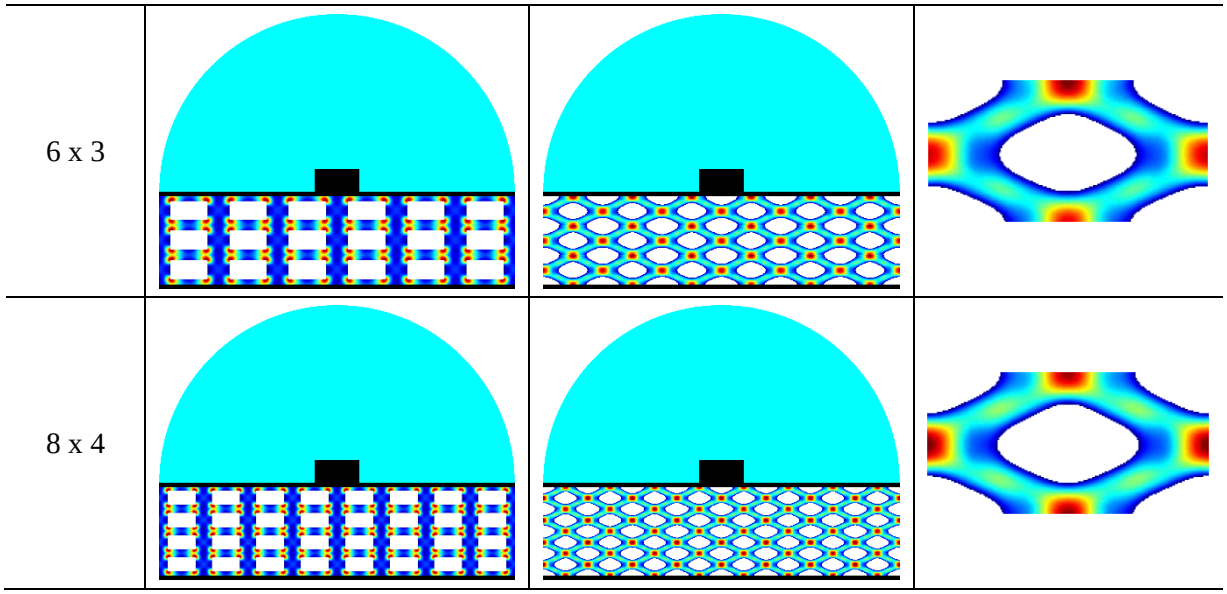


Figure 6 – Initial guess design and optimized final topologies of the coupled system considering four different cell modes: 2 x 1, 4 x 2, 6 x 3 and 8 x 4. Optimized unit cell for different configuration of the design domain.

The convergence of the unit cell to pattern is observed for the cell mode 6 x 3 and 8 x 4 and is expected to be repeated in cell modes with higher number of cells. Figure 7 presents the history of the objective function in the optimization procedure for all cell modes investigated. The volume of the unit cell is kept constant during the optimization, $V_f = V_i = 0.65$.

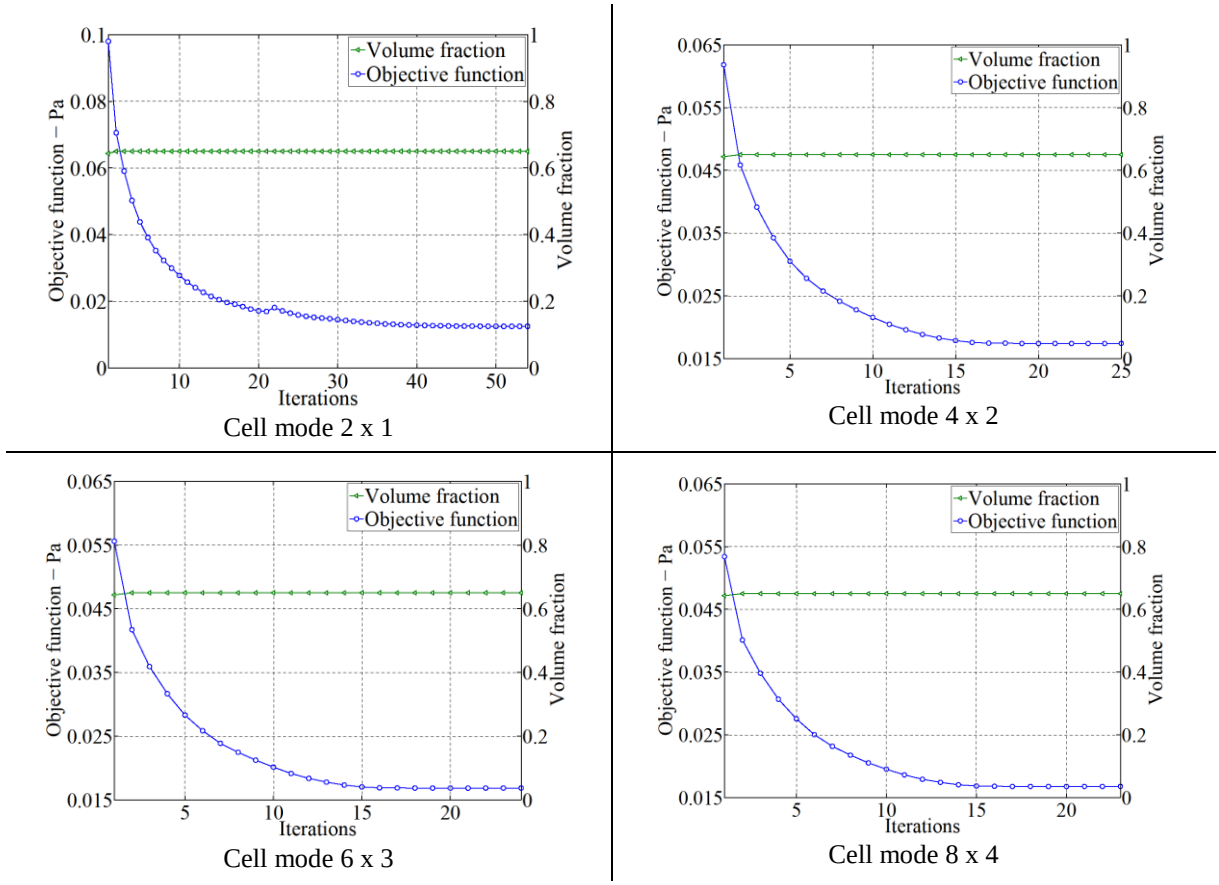


Figure 7 – Evolutionary histories of the objective function for the different cell modes analyzed with the new unit cell aspect ratio.

Can be observed in Fig. 7, that the number of iteration needed to reach the convergence criteria is decreasing with the increase of the number of periodic cells.

The topology optimization of system is now explored for a higher frequency of excitation. Instead of the frequency of 50 Hz previous analyzed, now a frequency of 350 Hz is taking into account to the optimization of the periodic design domain of the coupled system showed in Fig. 3. The cell mode of 32 unit cells (4 x 8) is used to build the design domain. The same previous material properties and BESO parameters are applied in this investigation. The initial guess design of the unit cell start with a structural volume fraction of 0.5, however, the void area in the unit cell is fill with acoustic fluid. In this way, the couple system presents now only two phases, acoustic fluid or structure material in all stages of the optimization procedure.

Figure 8 presents the initial and final topologies of the coupled system considering the optimization for the frequency of 350 Hz. The pressure distributions in the acoustic domain and in the design domain for both initial and final topologies. The vibration operational modes are presented considering the initial and the optimized system.

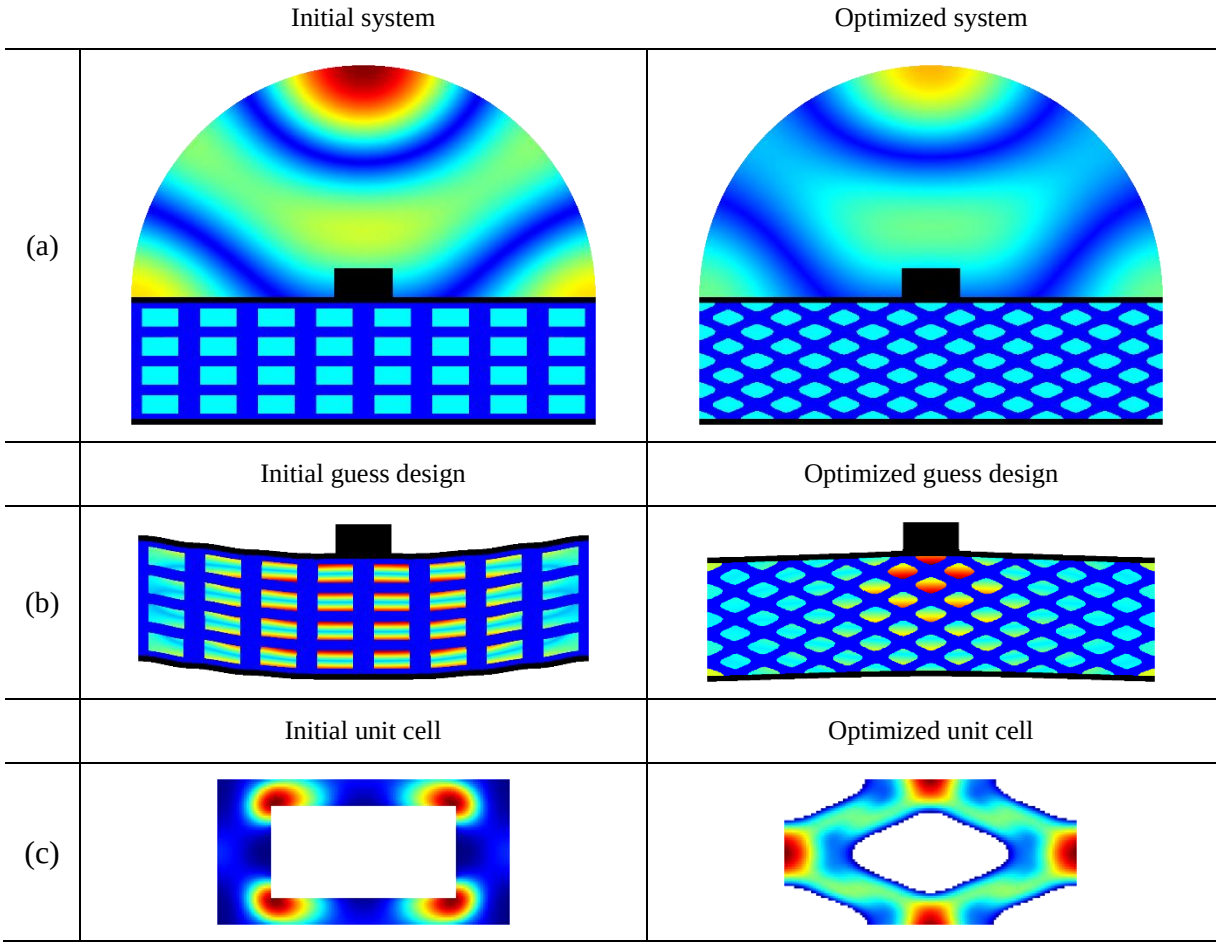


Figure 8 – Topology optimization of the couple acoustic-structure system considering the frequency of 350 Hz: (a) Pressure distribution on the acoustic domain for the initial and the optimized system; (b) Frequency response of displacement and pressure distributions in the design domain; (c) Sensitivity number distribution in the initial guess design and in the optimized unit cell.

The system frequency response, mean acoustic pressure, in the point P_{out} is show in Fig. 9 for initial system and for the optimized system. The response of the system indicates a significant reduction of the acoustic pressure in the point P_{out} for the frequency range of $(0-380)$ Hz.

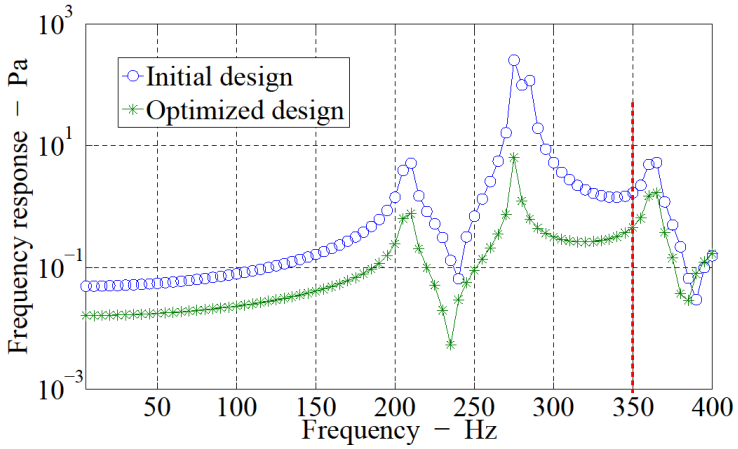


Figure 9 – Frequency response in the P_{out} in the acoustic domain considering the initial and the optimized design of the coupled system.

4.2 Example 2

The optimization of a coupled acoustic-structure system composed of two acoustic cavities separated by a structural domain is investigated in this example. As shown in Fig. 10, a harmonic pressure is imposed in the first acoustic cavity, P_{in}, and transmitted to the second cavity by the vibration of the clamped structure. The optimization procedure aims to minimize the pressure in the second cavity in the point P_{out}. The boundary conditions and geometric dimensions of the system are also shown in Fig. 10. The excitation frequency of the system is $f = 50$ Hz. The whole structural part is considered to be the design domain and its composed of periodic unit cells. A number of cell modes is investigated in order to obtain a convergent pattern of the unit cell. In all configurations, each unit cell is squared and discretized using 200 x 200 elements.

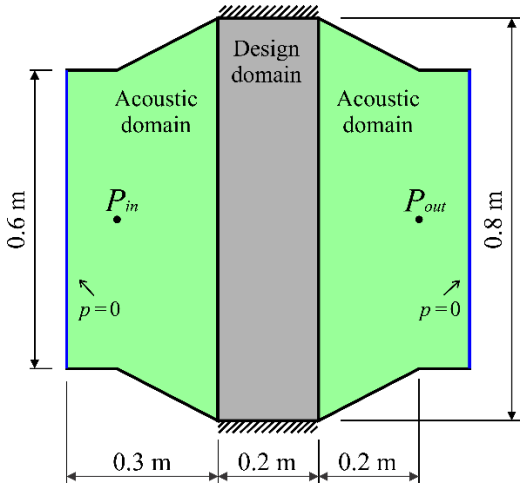


Figure 10 – Acoustic-structure system with a harmonic imposed pressure, P_{in}, for minimize of the acoustic pressure at specific point, P_{out}, in the other cavity.

The acoustic fluid has the follow properties: density $\rho = 1.2\,\text{kg/m}^3$ and the speed of sound $c = 343\,\text{m/s}$. The design domain material has Young's modulus $E = 100\,\text{MPa}$, Poisson's ratio $\mu = 0.3$ and density $\rho = 100\,\text{kg/m}^3$.

The BESO parameters used are: $ER = 0.02$, $Ar = 0.02$, $\tau = 0.01$, $N = 5$ and $p = 3$. For all cell mode configurations, the filter radius is defined as 6% of the side length of the unit cell. The optimization is performed with a constant volume fraction of the design domain $V_f = V_i = 0.40$.

Three different cell modes are investigated in order to obtain the convergence of the unit cell. Figure 11 presents the pressure distribution in the acoustic cavities and inside unit cells in the design domain for cases analyzed, cell modes 1 x 4, 2 x 8 and 3 x 12.

The optimized unit cells for the cell modes 2 x 8 and 3 x 12 are similar with one a slender variation of the topology.

The sensitivity number distribution in the optimized unit cell is also shown in Fig. 11 for all configurations of the periodic structure. The unit cell final topology converges for a symmetric pattern as the number of unit cell increases in horizontal and vertical directions. For the cell modes 2 x 8 and 3 x 12 the optimized unit cells show only a slender difference in the distribution of the material and can be considered as the converged pattern.

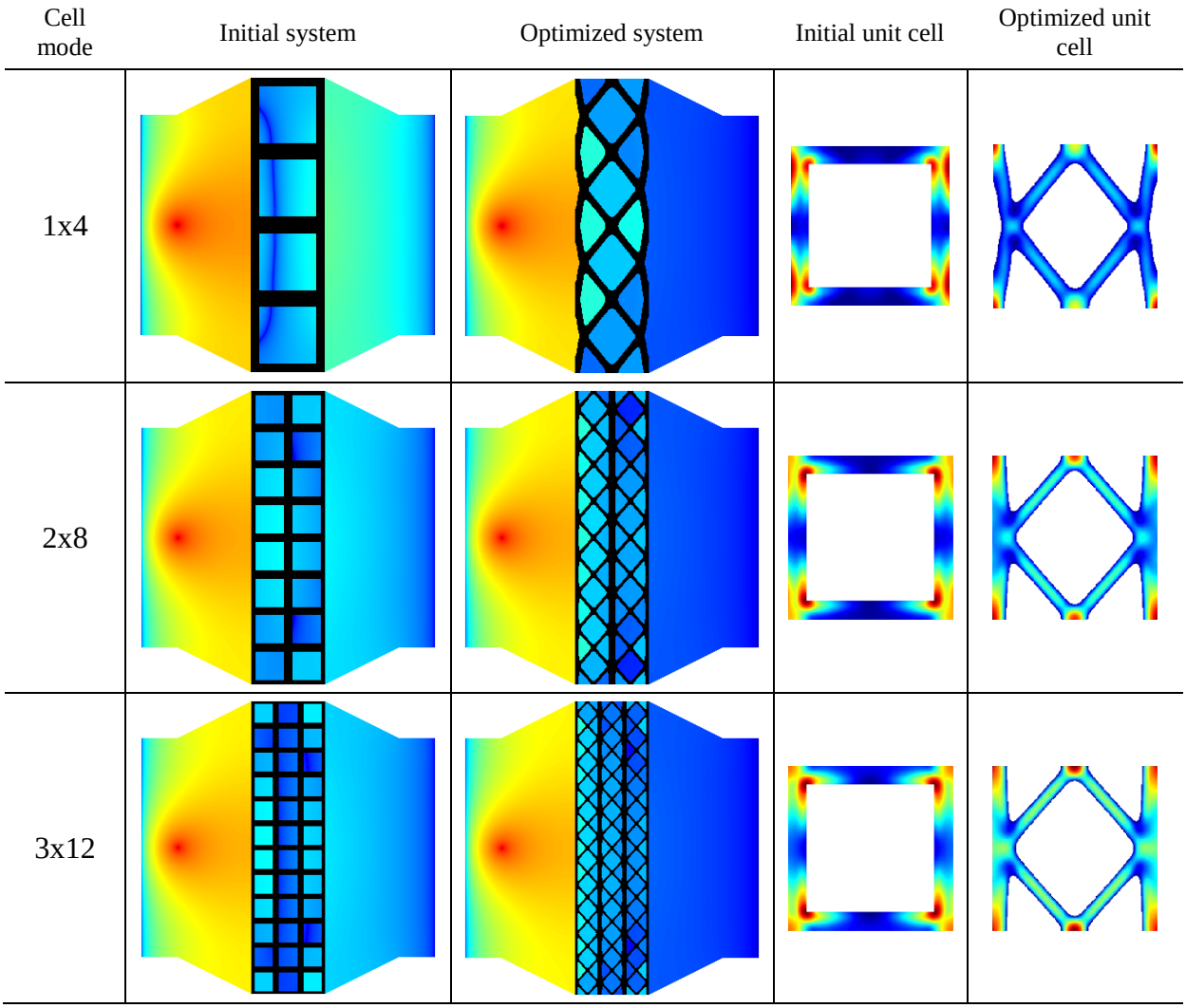


Figure 11 - Initial guess design and optimized final topologies of the fluid-structure system considering different cell modes: 1 x 4, 2 x 8 and 3 x 12. Initial and optimized unit cell for three cell modes different configurations.

Evolutionary history of the topology for the unit cell optimization is presented in Fig. 12 considering the first cell mode configuration 1 x 4.

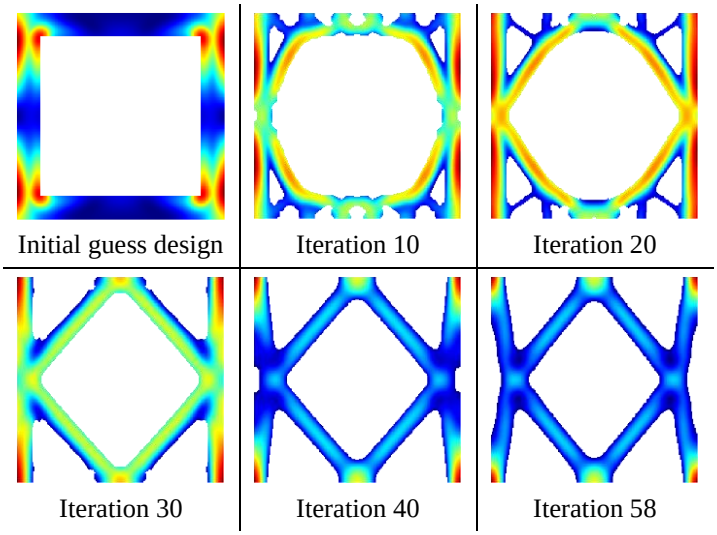


Figure 12 – Sensitivity number distribution on the evolutionary history of topology for the cell mode 1 x 4 case.

Smoothed convergences of the objective function are observed in Fig. 13 for all cell modes. The cell mode configuration 1 x 4 needed the highest number of iteration to satisfied the convergent criteria, 58 iterations. The initial and final objective function for the cell modes 1 x 4, 2 x 8 and 3 x 12 are $(0.787-0.014)\cdot 10^{-2}$ Pa, $(0.211-0.016)\cdot 10^{-2}$ Pa and $(0.170-0.016)\cdot 10^{-2}$ Pa, respectively. The cell mode 1 x 4 shows the best performance in this case. The configuration with more number of periodic unit cells, 3 x 12, required the lowest number of iterations in order to converge, 45 iterations.

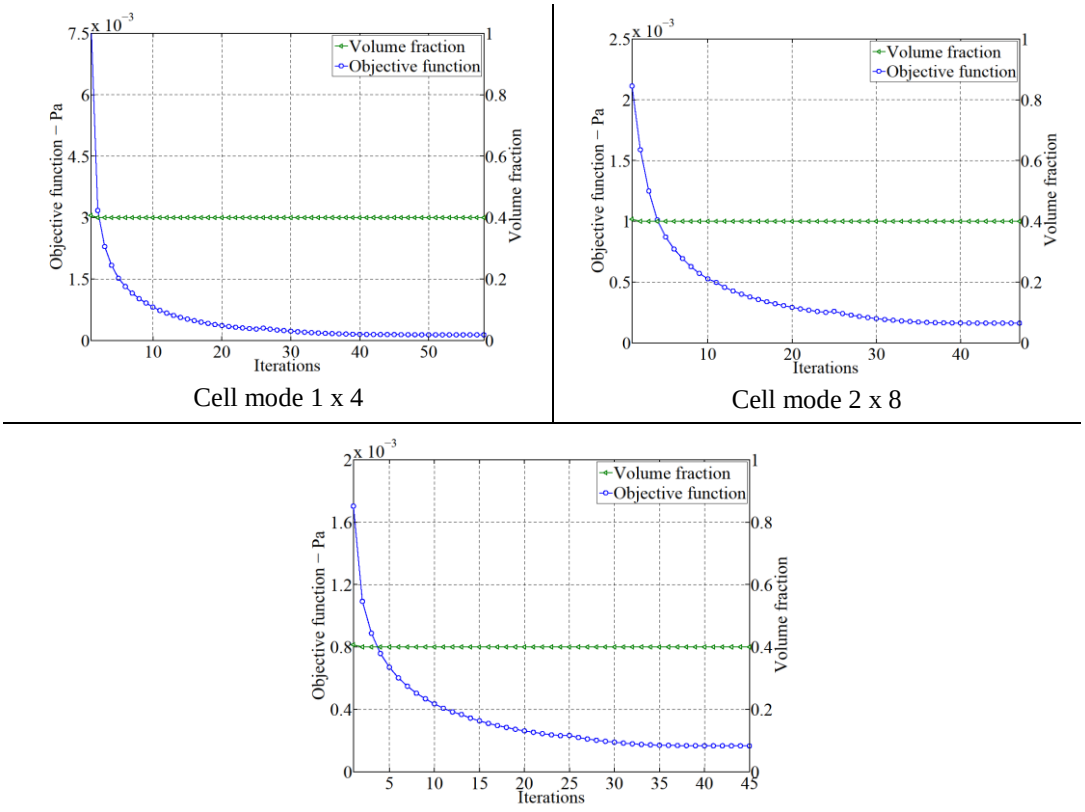


Figure 13 – Evolutionary histories of the objective function for all three cell modes of the design domain. The constant structural volume fraction of 0.4 for all cases during the optimization procedure.

The frequency response of the objective function, the mean acoustic pressure at point P_{out}, is presented in Fig. 14 considering the initial and the optimized coupled system. The frequency of interesting in the optimization procedure, 50 Hz, is indicated with the red line. Can be seen that the first natural frequency of the system increased from 60 to 185 Hz, approximately, as a consequence of the optimization process. The objective function decreased not only in the specific frequency, but in the range of $(0-140)$ Hz.

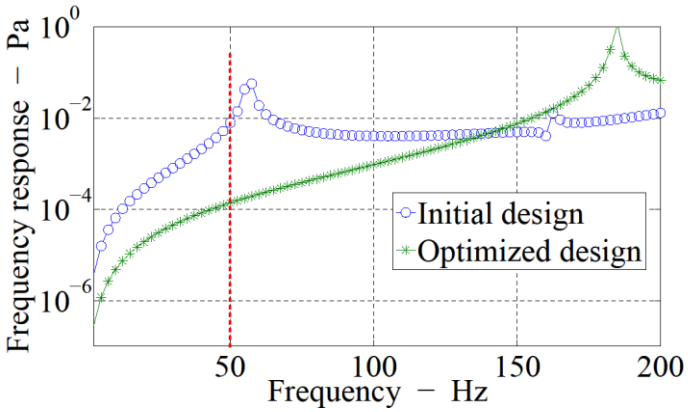


Figure 14 – Frequency response considering the acoustic pressure in the point P_{out} for the initial and the optimized design of the coupled acoustic-structure system.

5 CONCLUSIONS

A topology optimization methodology of periodic acoustic-structure systems has been presented in this work. The pressure frequency response is defined as the objective function in the optimization formulation. Two examples exploring different aspects of the proposed methodology are presented. The following conclusions are drawn:

- The proposed optimization procedure can effectively minimize the pressure frequency response in the acoustic domain of coupled acoustic-structure systems.
- A periodic geometry constraint has been implemented for the optimization of the coupled systems to obtain periodic unit cells in the design domain.
- Considering the unit cell composed of structural or void areas or composed of fluid and structural material, the minimization of the objective function and the convergence of the unit cell to a pattern have been reached in all analyzed examples.
- Analyzing systems with several different cell modes, the value of the objective function and the optimal topology tend to converge as the total number of periodic unit cell grows. In the cases studied, at least three unit cells in each direction is need to obtain the convergence of the unit cell topology.
- The minimum values of the objective function in the analyzed cases have been found for the configurations with the lowest numbers of unit cells. Nevertheless, the configurations with higher numbers of unit cells could be more versatile and have lower cost to construct using additive manufacturing once the unit cell topology have converged to a pattern.

ACKNOWLEDGMENTS

The authors are grateful to FAPESP (São Paulo Research Foundation, grant numbers 2011/09730-6, 2013/08293-7 and 2013/00085-6) for the financial support of this work.

REFERENCES

[1] M. P. Bendsoe and N. Kikuchi, "Generating optimal topologies in structural design using a homogenization method," *Comput. Methods in Appl. Mech. Eng.,* vol. 71, pp. 197-224, 1988.

[2] M. Zhou and G. I. N. Rozvany, "The COC algorithm, part II: topological, geometrical and generalized shape optimization," *Comput. Methods in Appl. Mech. Eng.,* vol. 89, pp. 309-336, 1991.

[3] M. P. Bendsoe and O. Sigmund, Topology Optimization - Theory, Methods and Applications, Berlin: Springer Verlag, 2003, p. 370.

[4] Y. M. Xie and G. P. Steven, "A simple evolutionary procedure for structural optimization," *Comput. Struct.,* vol. 49, pp. 885-896, 1993.

[5] Y. M. Xie and G. P. Steven, Evolutionary Structural Optimization, Springer, London, 1997.

[6] J. Sethian and A. Wiegmann, "Structural Boundary Design via Level Set and Immersed Interface Methods," *J. Comput. Phys.,* vol. 163, no. 2, pp. 489-528, 2000.

[7] M. Y. Wang, X. Wang and D. Guo, "A level set method for structural topology optimization," *Comput. Methods in Appl. Mech. Eng.,* vol. 192, no. 1-2, pp. 227-246, January 2003.

[8] X. Huang and Y. M. Xie, Evolutionary Topology Optimization of Continuum Structures: Methods and Applications, Chichester: John Wiley & Sons, Ltd, 2010, p. IX+223.

[9] G. H. Yoon, J. S. Jensen and O. Sigmund, "Topology optimization of acoustic-structure problems using a mixed finite element formulation," *Int. J. Numer. Meth. Eng.,* vol. 70, pp. 1049-1075, 2007.

[10] L. Shu, M. Y. Wang and Z. Ma, "Level set based topology optimization of vibrating structures for coupled acoustic-structural dynamics," *Comput. Struct.,* vol. 132, pp. 34-42, 2014.

[11] J. Du and N. Olhoff, "Minimization of sound radiation from vibrating bi-material structures using topology optimization," *Struct. Multidisc. Optim.,* vol. 33, pp. 305-321, 2007.

[12] J. Du and N. Olhoff, "Topological design of vibrating structures with respect to optimum sound pressure characteristics in a surrounding acoustic medium," *Struct. Multidisc. Optim.,* vol. 42, pp. 43-54, 2010.

[13] W. M. Vicente, R. Picelli and R. Pavanello, "An Evolutionary Structural Optimization applied to Fluid-Structure Problems," in *Proceedings of the Eleventh International Conference on Computational Structures Technology, CST 2012, Dubrovnik, Croatia.,* 2012.

[14] R. Picelli, W. M. Vicente and R. Pavanello, "Bi-directional evolutionary structural optimization for design-dependent fluid pressure loading problems," *Eng. Optimiz.,* vol. 47, no. 10, pp. 1324-1342, 2015.

[15] R. Picelli, W. M. Vicente, R. Pavanello and Y. M. Xie, "Evolutionary topology optimization for natural frequency maximization problems considering acoustic-structure interaction," *Finite Elem. Anal. Des.,* vol. 106, pp. 56-64, 2015.

[16] O. Sigmund, "Materials with prescribed constitutive parameters: An inverse homogenization problem," *Int. J. Solids. Struct.,* vol. 31, no. 17, pp. 2313-2329, 1994.

[17] O. Sigmund and S. Torquato, "Design of materials with extreme thermal expansion using a three-phase topology optimization method," *J. Mech. Phys. Solids,* vol. 45, no. 6, pp. 1037-1067, 1997.

[18] X. Huang and Y. M. Xie, "Optimal design of periodic structures using evolutionary topology optimization," *Struct. Multidisc. Optim.,* vol. 36, no. 6, pp. 597-606, 2007.

[19] Z. H. Zuo, X. Huang, X. Yang, J. H. Rong and Y. M. Xie, "Comparing optimal material microstructures with optimal periodic structures," *Comp. Mater. Sci.,* vol. 69, pp. 137-147, 2013.

[20] Z. H. Zuo, Y. M. Xie and X. Huang, "An improved bi-directional evolutionary topology optimization method for frequencies," *Int. J. Struct. Stab. Dy.,* vol. 10, pp. 55-75, 2010.

[21] W. M. Vicente, R. Picelli, R. Pavanello and Y. M. Xie, "Topology optimization of frequency responses of fluid-structure interaction systems," *Finite Elem. Anal. Des.,* vol. 98, pp. 1-13, 2015.

[22] W. M. Vicente, Z. H. Zuo, R. Pavanello, T. K. L. Calixto, R. Picelli and Y. M. Xie, "Concurrent topology optimization for minimizing frequency responses of two-level hierarchical structures," *Comput. Methods in Appl. Mech. Eng.,* vol. 301, pp. 116-136, 2016.

[23] K. J. Bathe, Finite Element Procedures in Engineering Analysis, Englewood Cliffs, NJ: Prentice-Hall, 1982.

[24] X. Huang, Z. H. Zuo and Y. M. Xie, "Evolutionary topological optimization of vibrating continuum structures for natural frequencies," *Comput. Struct.,* vol. 88, pp. 357-364, 2010.

[25] X. Huang and Y. M. Xie, "Convergent and mesh-independent solutions for the bi-directional evolutionary structural optimization method," *Finite Elem. Anal. Des.,* vol. 43, pp. 1039-1049, 2007.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

BESO APPROACH TO TOPOLOGY OPTIMIZATION OF GAN PHONONIC CRYSTALS

Luca D'Alessandro[1,2*], **Bichoy Bahr**[2], **Luca Daniel**[3], **Dana Weinstein**[4], **Raffaele Ardito**[1]

[1] Dept. of Civil and Environmental Engineering, Politecnico di Milano
Milan, Italy
e-mail: {luca.dalessandro, raffaele.ardito}@polimi.it

[2] Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
Cambridge (MA), USA
{lucadale, bichoy}@mit.edu

[3] Research Laboratory of Electronics, Massachusetts Institute of Technology
Cambridge (MA), USA
luca@mit.edu

[4] Dept. of Electric and Computer Engineering, Purdue University
West Lafayette (IN), USA
danaw@purdue.edu

Keywords: Phononic Crystals, Resonators, Acoustic waves, Bandgap, Topology optimization, Micromechanics

Abstract. *The use of Phononic Crystals (PnC) in suspended structures and microstructures, such as plates and slabs, has gained a lot of attention in the past years for the wide range of feasible applications (acoustic waveguides, acoustic insulation, acoustic cloaking) and for the easy fabrication technique. Since the performance of the device is related to the band of frequencies reflected by the PnC and since this band (called bandgap) depends on the geometric and material properties of the fundamental unit cell of the PnC, a useful tool for the design of those structures is topology optimization. This paper is focused on a novel and fast engineering use of Bidirectional Evolutionary Structural Optimization for the definition of the optimal hole configuration in air-solid PnC. The technique adopted finds the optimal shape of the hole in less than 20 iterations, and it is easy implemented in a 2D plane strain Matlab finite element solver.*

1 INTRODUCTION

Phononic Crystals (PnCs) gained lots of attention over the past several years, owing to their versatility and wide spectrum of applications: acoustic isolation, noise suppression, vibration attenuation, acoustic waveguides, acoustic super-lenses, negative refraction and acoustic cloaking [1],[2].

This paper is focused on the use of PnCs in "unreleased" resonators [3], which overcome many challenges faced by traditional MEMS resonators, including non-linearity, power handling, residual stresses and yield [4]. PnCs are represented by periodic media, which exhibit bandgaps, i.e. bounded frequency regions for which the propagation of sound waves through the crystal is impeded. The central frequency and the size of the bandgap basically depend on the materials employed in PnCs and on the shape of the unit cell.

Most of the abovementioned applications will significantly benefit from PnCs with larger bandgap. For this reason, it is highly desirable to optimize the shape of a PnC unit cell, in order to obtain the largest bandgap while complying with manufacturability constraints.

The present paper aims at the achievement of the best shape for the gallium nitride (GaN) solid-hole PnC unit cell, producing the largest possible phononic bandgap around a given center frequency.

The procedure starts from a given geometry, endowed with a certain bandgap, and proceeds with an automatic shape optimization algorithm in order to evolve this geometry into an optimal one for maximum bandgap around the requested central frequency. Several research groups have treated the topological optimization of PnCs, e.g. by using the SIMP approach and the MMA algorithm [5], genetic algorithm [6]–[8]. In this work, we consider the application of Bidirectional Evolutionary Structural Optimization (BESO) that was previously applied for Photonic Crystals [9] and is now used for the first time in the solid-hole PnCs with an engineering perspective. The algorithm is able to modify the shape by adding and removing elements, using the FEM-based sensitivity analysis for the acoustic wave frequency.

The proposed optimization algorithm can successfully obtain the optimal topology in less than 20 iterations: the optimization process involves a 500% gain with respect to the initial bandgap.

The optimal dimensionless bandgap depends only on the hole's shape and on the ratio between elastic constants of the material. Attention is paid to the robustness of the achieved solution, in view of the construction tolerances, which might alter the optimality of the real device.

2 FINITE ELEMENT PROBLEM FORMULATION

PnC structures are periodic structures generally composed of two or more elastic material phases, solid or fluid. The theoretical framework is wave propagation in elastic inhomogeneous media, and it is governed by the following elasto-dynamic equations [10], [11]:

$$\begin{aligned}
\rho\frac{\partial^2 u_x}{\partial t^2} &= \frac{\partial}{\partial x}\left(\lambda\nabla[u_x,u_y,u_z]^T\right)+\nabla\left(\mu\left(\nabla u_x+\frac{\partial[u_x,u_y,u_z]^T}{\partial x}\right)\right)\\
\rho\frac{\partial^2 u_y}{\partial t^2} &= \frac{\partial}{\partial y}\left(\lambda\nabla[u_x,u_y,u_z]^T\right)+\nabla\left(\mu\left(\nabla u_y+\frac{\partial[u_x,u_y,u_z]^T}{\partial y}\right)\right)\\
\rho\frac{\partial^2 u_z}{\partial t^2} &= \frac{\partial}{\partial z}\left(\lambda\nabla[u_x,u_y,u_z]^T\right)+\nabla\left(\mu\left(\nabla u_x+\frac{\partial[u_x,u_y,u_z]^T}{\partial z}\right)\right)
\end{aligned} \tag{1}$$

where ρ is the material density, λ and μ are Lamé's coefficients, u_x, u_y, u_z the cartesian components of the displacement vector.

The case study analyzed in this paper is referred to PnC slabs composed of air-solid phase materials, it is possible to assume some simplifications such that:

- the material parameters ρ, λ and μ depend only on the x and y (in-plane) coordinates: $\rho = \rho(x, y)$, $\lambda = \lambda(x, y)$, $\mu = \mu(x, y)$;
- being the thickness of the slab small compared to the in-plane dimensions, plane strain hypothesis can be exploited: $\partial[u_x, u_y, u_z]^T / \partial z = \mathbf{0}$, $u_z = 0$.

Due to the previous hypotheses, the elasto-dynamic problem (1) can be formulated as:

$$\begin{aligned}\rho\frac{\partial^2 u_x}{\partial t^2} &= \frac{\partial}{\partial x}\left((2\mu+\lambda)\frac{\partial u_x}{\partial x}+\lambda\frac{\partial u_y}{\partial y}\right)+\frac{\partial}{\partial y}\left(\mu(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x})\right)\\ \rho\frac{\partial^2 u_y}{\partial t^2} &= \frac{\partial}{\partial x}\left(\mu(\frac{\partial u_y}{\partial x}+\frac{\partial u_x}{\partial y})\right)+\frac{\partial}{\partial y}\left((2\mu+\lambda)\frac{\partial u_y}{\partial y}+\lambda\frac{\partial u_x}{\partial x}\right)\end{aligned} \tag{2}$$

where the only unknowns are $u_x = u_x(\mathbf{r}, \mathbf{k}, t)$, $u_y = u_y(\mathbf{r}, \mathbf{k}, t)$, where $\mathbf{r} = [x, y]^T$ is the vector of coordinates in the plane, $\mathbf{k} = [k_x, k_y]^T$ is the plane wave number vector, t is the time.

In the frequency domain, the equations (2) become:

$$\begin{aligned}-\rho u_x\omega^2 &= \frac{\partial}{\partial x}\left((2\mu+\lambda)\frac{\partial u_x}{\partial x}+\lambda\frac{\partial u_y}{\partial y}\right)+\frac{\partial}{\partial y}\left(\mu(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x})\right)\\ -\rho u_y\omega^2 &= \frac{\partial}{\partial x}\left((2\mu+\lambda)\frac{\partial u_x}{\partial x}+\lambda\frac{\partial u_y}{\partial y}\right)+\frac{\partial}{\partial y}\left(\mu(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x})\right)\end{aligned} \tag{3}$$

where ω is the angular frequency, $u_x = u_x(\mathbf{r}, \mathbf{k})$, $u_y = u_y(\mathbf{r}, \mathbf{k})$ are no more function of time.

Due to the periodicity of the PnC, it's possible to study the elasto-dynamic problem referring to the single cell only, exploiting the Floquet-Bloch theorem [12]:

$$\begin{aligned}u_x &= u_x(\mathbf{r}, \mathbf{k}) = u_x(\mathbf{r})e^{i\mathbf{kr}}\\ u_y &= u_y(\mathbf{r}, \mathbf{k}) = u_y(\mathbf{r})e^{i\mathbf{kr}}\end{aligned} \tag{4}$$

In the literature, there are two approaches to introduce (4) in the problem under investigation:

- direct substitution in (3) and proceed to the weak formulation of the equations and the finite element (FE) discretization ([5],[10],[11]);
- first weak formulation (as usual for elasto-dynamic problems), then FE discretization, and at the end enforce the equations (4) as periodic boundary conditions ([13],[14],[15]).

Being the aim of this work the implementation of a topology optimization technique, it is more suitable to refer to the second approach since in this way the assembly of the general mass and stiffness matrices of the problem is not depending on the wave vector $\mathbf{k}$ and therefore it is done only once for each $\mathbf{k}$ vector.

It is possible to discretize the weak form of equations (3) into finite elements, being the in-plane displacements expressed like:

$$u_x = u_x(\mathbf{r},\mathbf{k}) = \sum_{el}\sum_{j=1}^{n} N_j u_{xj}^{el}$$
$$u_y = u_y(\mathbf{r},\mathbf{k}) = \sum_{el}\sum_{j=1}^{n} N_j u_{yj}^{el} \quad (5)$$

where el is the index for the elements of the mesh, j is the index for the nodes of each element, N_j are the shape functions for the quadrilateral plane element: $j = 1,2,3,4$, u_{xj}^{el} and u_{yj}^{el} are the components of the displacement of each node of an element.

The linear quadrilateral element has been chosen and, in particular, a square FE shape in order to build a regular grid of elements over the unit cell: in this way the algorithm of topological optimization is easily implemented, dealing with each element as a "pixel" that can be switched between fill and void. Being the aim of this work the optimization of a solid/air PnC the finite element formulation is such that the only elements that contribute to the matrices are the solid elements: the air domain is not considered.

The FE discretization (5) of equations (3) yields the governing FE governing equation:

$$\mathbf{u}^T(\mathbf{k},\mathbf{r})(\mathbf{K} - \omega^2\mathbf{M})\mathbf{u}(\mathbf{k},\mathbf{r}) = 0$$
$$\mathbf{u}(\mathbf{k},\mathbf{r}) = [u_{jx}(\mathbf{k},\mathbf{r}), u_{jy}(\mathbf{k},\mathbf{r}), \ldots \mathrm{j} = 1,2..\mathrm{N}_{nodes}]^T \quad (6)$$

where $\mathbf{K}$ and $\mathbf{M}$ are the assembled global stiffness and mass matrices that do not depend on $\mathbf{k}$, $\mathbf{u}(\mathbf{k},\mathbf{r})$ is the global vector of unknowns containing in order the x and y components of each node displacement.

In order to formulate the complete FE problem, some suitable boundary conditions must be defined together with equations (6):

- free boundary condition is applied to the boundary elements of the solid domain in the interface with air domain;
- periodic boundary conditions are defined after reordering the nodes of the mesh as shown in Figure 1.

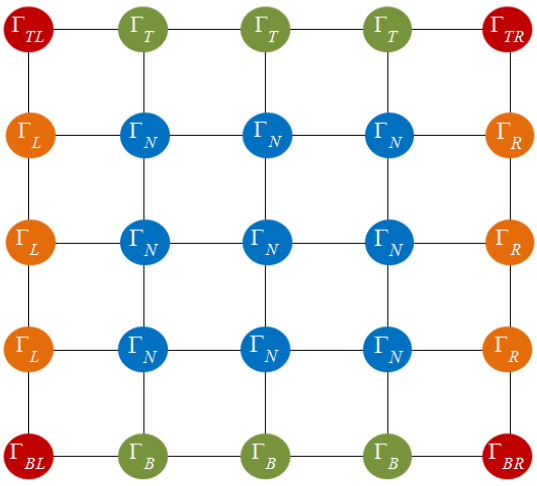


Figure 1: FE grid with nodes identified by color and type with respect to periodicity boundary conditions. Γ stands for boundary, T for Top, B for Bottom, L for Left, R for Right, , N for Internal. The nodes of different types identified with the same color are related by the bc.

By means of this organization, it is possible to define the periodic boundary conditions in a matrix form:

$$\begin{bmatrix} \mathbf{u}|_{\Gamma_N} \\ \mathbf{u}|_{\Gamma_B} \\ \mathbf{u}|_{\Gamma_L} \\ \mathbf{u}|_{\Gamma_{BL}} \\ \mathbf{u}|_{\Gamma_T} \\ \mathbf{u}|_{\Gamma_R} \\ \mathbf{u}|_{\Gamma_{TL}} \\ \mathbf{u}|_{\Gamma_{BR}} \\ \mathbf{u}|_{\Gamma_{TR}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{(n_N \times n_N)} & \mathbf{0}_{(n_N \times n_B)} & \mathbf{0}_{(n_N \times n_L)} & \mathbf{0}_{(n_N \times n_{BL})} \\ \mathbf{0}_{(n_B \times n_N)} & \mathbf{I}_{(n_B \times n_B)} & \mathbf{0}_{(n_B \times n_L)} & \mathbf{0}_{(n_B \times n_{BL})} \\ \mathbf{0}_{(n_L \times n_N)} & \mathbf{0}_{(n_L \times n_B)} & \mathbf{I}_{(n_L \times n_L)} & \mathbf{0}_{(n_L \times n_{BL})} \\ \mathbf{0}_{(n_{BL} \times n_N)} & \mathbf{0}_{(n_{BL} \times n_B)} & \mathbf{0}_{(n_{BL} \times n_L)} & \mathbf{I}_{(n_{BL} \times n_{BL})} \\ \mathbf{0}_{(n_T \times n_N)} & e^{ik_y \Delta y}\mathbf{I}_{(n_T \times n_B)} & \mathbf{0}_{(n_T \times n_L)} & \mathbf{0}_{(n_T \times n_{BL})} \\ \mathbf{0}_{(n_R \times n_N)} & \mathbf{0}_{(n_R \times n_B)} & e^{ik_x \Delta x}\mathbf{I}_{(n_R \times n_L)} & \mathbf{0}_{(n_R \times n_{BL})} \\ \mathbf{0}_{(n_{TL} \times n_N)} & \mathbf{0}_{(n_{TL} \times n_B)} & \mathbf{0}_{(n_{TL} \times n_L)} & e^{ik_y \Delta y}\mathbf{I}_{(n_{TL} \times n_{BL})} \\ \mathbf{0}_{(n_{BR} \times n_N)} & \mathbf{0}_{(n_{BR} \times n_B)} & \mathbf{0}_{(n_{BR} \times n_L)} & e^{ik_x \Delta x}\mathbf{I}_{(n_{BR} \times n_{BL})} \\ \mathbf{0}_{(n_{TR} \times n_N)} & \mathbf{0}_{(n_{TR} \times n_B)} & \mathbf{0}_{(n_{TR} \times n_L)} & e^{i(k_x \Delta x + k_y \Delta y)}\mathbf{I}_{(n_{TR} \times n_{BL})} \end{bmatrix} \begin{bmatrix} \mathbf{u}|_{\Gamma_N} \\ \mathbf{u}|_{\Gamma_B} \\ \mathbf{u}|_{\Gamma_L} \\ \mathbf{u}|_{\Gamma_{BL}} \end{bmatrix} \quad (7)$$

where the symbol | denotes "on the boundary", and Δx and Δy denote the offset between the coordinates of the nodes in x and y direction, respectively.

Equation (7) can be rewritten as:

$$\mathbf{u}(\mathbf{k},\mathbf{r}) = \mathbf{T}(\mathbf{k},\mathbf{r})\tilde{\mathbf{u}}(\mathbf{r}) \quad (8)$$

where $\tilde{\mathbf{u}}(\mathbf{r}) = \left[\mathbf{u}|_{\Gamma_N}{}^T, \mathbf{u}|_{\Gamma_B}{}^T, \mathbf{u}|_{\Gamma_L}{}^T, \mathbf{u}|_{\Gamma_{BL}}{}^T \right]^T$ that collects only independent displacements.

Substituting equation (8) in equation (6):

$$\tilde{\mathbf{u}}^T(\mathbf{r})\mathbf{T}^H(\mathbf{k},\mathbf{r})(\mathbf{K} - \omega^2\mathbf{M})\mathbf{T}(\mathbf{k},\mathbf{r})\tilde{\mathbf{u}}(\mathbf{r}) = 0 \quad (9)$$

where ^{H} stands for Hermitian transpose.

The following problem matrices are Hermitian:

$$\begin{aligned} \tilde{\mathbf{K}} &= \mathbf{T}^H(\mathbf{k},\mathbf{r})(\mathbf{K})\mathbf{T}(\mathbf{k},\mathbf{r}) \\ \tilde{\mathbf{M}} &= \mathbf{T}^H(\mathbf{k},\mathbf{r})(\mathbf{M})\mathbf{T}(\mathbf{k},\mathbf{r}) \end{aligned} \quad (10)$$

Introducing (10) in (9) one obtains:

$$\tilde{\mathbf{u}}^T(\mathbf{r})(\tilde{\mathbf{K}} - \omega^2\tilde{\mathbf{M}})\tilde{\mathbf{u}}(\mathbf{r}) = 0 \quad (11)$$

Equation (11) must be solved in order to find eigenvalues $\omega_n{}^2$, eigenvectors $\tilde{\mathbf{u}}_n(\mathbf{r})$ and the associated dispersion relation at varying $\mathbf{k}$.

To obtain a good description of the dynamic behavior of the PnC for all the possible $\mathbf{k}$ vectors in the $\mathbf{k}$ space (plane), it is possible to choose a limited number of $\mathbf{k}$ vectors, discretizing the boundary of the first irreducible Brillouin zone [16]. In Figure 2, the rectangle in $\mathbf{k}$ space is the first Brillouin zone. Due to the square symmetry of the unit cell, it is possible to focus the attention only on the colored triangle (first irreducible Brillouin zone). It is claimed then from many authors that from the engineering point of view the maximum and minimum points of the bandgap can be determined only focusing the attention on the $\mathbf{k}$ vectors along the boundary of the First Irreducible Brillouin zone, following the path indicated in Figure 3.

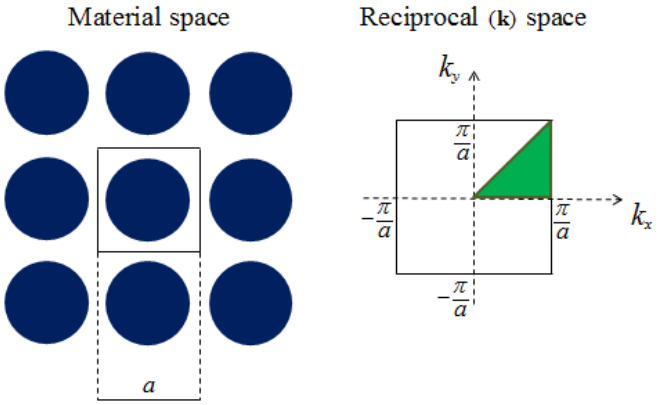


Figure 2: Material and reciprocal space, First Brillouin Zone.

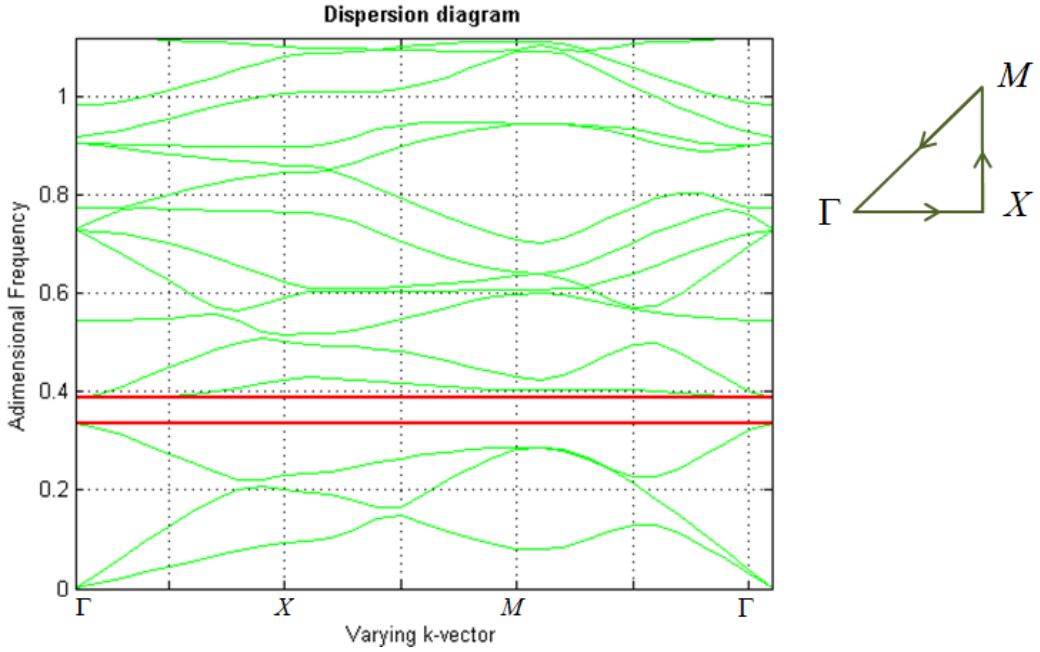


Figure 3: Dispersion Diagram for k-vector in the First Irreducible Brillouin Zone.

3 OPTIMIZATION PROBLEM FORMULATION

3.1 Objective function

The aim of this work is to find the optimal topology that maximizes an existing bandgap around a certain central frequency for solid/air PnC.

The objective function of this optimization is the bandgap width normalized with respect to the central frequency of the bandgap:

$$\max_{\boldsymbol{\rho}} \left\{ bg(\boldsymbol{\rho}) = \frac{\Delta\omega^2(\boldsymbol{\rho})}{\omega^2_{centr}(\boldsymbol{\rho})} = 2\frac{\min\limits_{\mathbf{k}} \omega^2_{n+1}(\mathbf{k},\boldsymbol{\rho}) - \max\limits_{\mathbf{k}} \omega^2_{n}(\mathbf{k},\boldsymbol{\rho})}{\min\limits_{\mathbf{k}} \omega^2_{n+1}(\mathbf{k},\boldsymbol{\rho}) + \max\limits_{\mathbf{k}} \omega^2_{n}(\mathbf{k},\boldsymbol{\rho})} \right\} \tag{12}$$

where $\boldsymbol{\rho}$ is a vector with length equal to number of elements containing the densities of each element, and n is the number of the bottom eigenvalue that defined the bandgap analyzed.

It is important to notice that in this work, for engineering purposes, the densities in the vector $\boldsymbol{\rho} = \{\rho_e, e = 1, \ldots, N_{elements}\}$ can only assume the values of 0 or 1. No interpolation between the two values is considered, the element could be "alive" (taken into account in the FE formulation) or "dead" (part of the air domain that is not considered in the FE formulation).

3.2 Sensitivity analysis

The goal is to capture how much the bandgap varies switching on/off each single element, this is a well-suited problem for the Evolutionary Structural Optimization framework, since the answer can be derived analytically [17].

It is possible to define each eigenvalue of the problem as follows:

$$\begin{aligned} \omega_n^2 &= \frac{k_n}{m_n} \\ k_n &= \mathbf{u}_n^T \mathbf{K} \mathbf{u}_n \\ m_n &= \mathbf{u}_n^T \mathbf{M} \mathbf{u}_n \end{aligned} \tag{13}$$

where k_n is the modal stiffness, m_n is the modal mass and $\mathbf{u}_n = \mathbf{T}\tilde{\mathbf{u}}_n$ is the eigenvector.

The sensitivity with respect to the variables is:

$$\frac{\partial \omega_n^2}{\partial \rho_e} = \frac{1}{m_n}\frac{\partial k_n}{\partial \rho_e} - \frac{k_n}{m_n^2}\frac{\partial m_n}{\partial \rho_e} = \frac{1}{m_n}\left(\frac{\partial k_n}{\partial \rho_e} - \omega_n^2 \frac{\partial m_n}{\partial \rho_e}\right) = \frac{1}{m_n}\left(\frac{\partial (\mathbf{u}_n^T \mathbf{K} \mathbf{u}_n)}{\partial \rho_e} - \omega_n^2 \frac{\partial (\mathbf{u}_n^T \mathbf{M} \mathbf{u}_n)}{\partial \rho_e}\right) \tag{14}$$

Assuming that modifying only few elements the shape of the eigenvector will not change significantly:

$$\frac{\partial \mathbf{u}_n}{\partial \rho_e} \approx 0 \tag{15}$$

equation (14) becomes:

$$\frac{\partial \omega_n^2}{\partial \rho_e} = \frac{1}{m_n}\left(\frac{\partial (\mathbf{u}_n^T \mathbf{K} \mathbf{u}_n)}{\partial \rho_e} - \omega_n^2 \frac{\partial (\mathbf{u}_n^T \mathbf{M} \mathbf{u}_n)}{\partial \rho_e}\right) = \frac{1}{m_n}\mathbf{u}_n^T\left(\frac{\partial \mathbf{K}}{\partial \rho_e} - \omega_n^2 \frac{\partial \mathbf{M}}{\partial \rho_e}\right)\mathbf{u}_n \tag{16}$$

from which it is clear how the change of the density of a single element can contribute to the change of the modulus of each eigenvalue by means of a combination of kinetic and potential energy.

It is possible to define a parameter representing the sensitivity obtained:

$$\alpha_n^e = \frac{1}{m_n}\mathbf{u}_n^T\left(\frac{\partial \mathbf{K}}{\partial \rho_e} - \omega_n^2 \frac{\partial \mathbf{M}}{\partial \rho_e}\right)\mathbf{u}_n \tag{17}$$

By examining equation (17), it is clear that in order to increase the modulus of ω_n^2 it is necessary to set:

$$\rho_e = 0 \text{ for } e \mid \alpha_n^e > 0 \tag{18}$$

Since the optimization problem defined in (12) involves both the upper and the lower extrema of the bandgap, it is necessary to define also the sensitivity for the upper eigenvalue of the bandgap:

$$\alpha_{n+1}^{e} = \frac{1}{m_{n+1}} \mathbf{u}_{n+1}^{T} \left(\frac{\partial \mathbf{K}}{\partial \rho_e} - \omega_{n+1}^{2} \frac{\partial \mathbf{M}}{\partial \rho_e} \right) \mathbf{u}_{n+1} \tag{19}$$

By means of (17), (18) and (19) the sensitivity number referred to the bandgap width is:

$$\alpha_{bg(n,n+1)}^{e} = \alpha_{n+1}^{e} - \alpha_{n}^{e} \tag{20}$$

The Evolutionary Structural Optimization technique is such that the change in the density is only possible in one direction, i.e. from one phase material to the other one and not the vice versa.

A slight modification is the so-called Bidirectional ESO technique that allows both "directions" of change of the element density depending on the $\alpha_{bg(n,n+1)}^{e}$ value. In order to correctly implement the technique it is important to notice that:

$$\begin{aligned} \sum_{e=1}^{N_{elements}} \alpha_{n}^{e} &= 0 \\ \sum_{e=1}^{N_{elements}} \alpha_{n+1}^{e} &= 0 \end{aligned} \tag{21}$$

It is important to notice that in order to obtain a feasible design the general periodic structure must not have regions with suspended material, therefore this constraint must be added to the optimization algorithm.

Another important issue regards the numerical implementation: easily these kinds of algorithms end up with checkerboard topology, and the particular implementation for solid/air phases can lead to suspended regions of material. In order to avoid such effects, a filtering subroutine is introduced in the algorithm.

3.3 Optimization algorithm

Following the path of reasoning introduced in the previous section, equation (18), the core of the algorithm is:

$$\begin{aligned} \rho_e = 0 \quad \text{for} \quad e \,|\, \alpha_{bg(n,n+1)}^{e} > 0 \ \text{[for elements to be "killed"]} \\ \rho_e = 1 \quad \text{for} \quad e \,|\, \alpha_{bg(n,n+1)}^{e} < 0 \ \text{[for elements to be "activated"]} \end{aligned} \tag{22}$$

In this work, n is chosen at the beginning of the optimization procedure corresponding to the eigenvalue that defines the bottom boundary for the bandgap.

During the optimization procedure the value of n may change from the starting one but the bandgap will always be centered around the desired frequency.

A general observation is mandatory in relation to (22): since the optimization concerns solid-air PnCs and the air domain is treated by eliminating the related degrees of freedom from the matrix formulation, it is not possible to define the densities in the hole. Therefore, the algorithm is built so that the densities of the elements in the material domain in the neighborhood of the hole define which elements to "live" in the hole itself.

The optimization scheme is described in the next Figure 4.

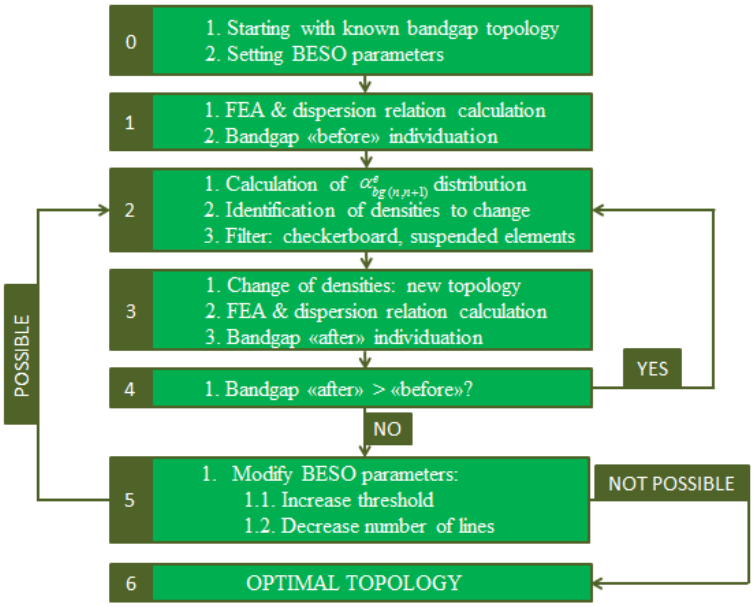


Figure 4: Optimization algorithm.

The parameters governing each BESO optimization iteration (see Figure 4) are:

- $nlines^d$ maximum number of lines of elements near to the hole to be "killed"
- $nlines^l$ maximum number of lines of elements in the hole to be "activated"
- $nelem^d$ maximum number of elements to be "killed"
- $nelem^l$ maximum number of elements to be "activated"
- thr^d threshold for elements to be "killed", $thr^d > 0$
- thr^l threshold for elements to be "activated", $thr^l < 0$

where ^{d} stands for "dying" and ^{l} stands for "living".

At each iteration, elements to be "killed" are:

$$\mathbf{e}^d = \{e \,|\, \alpha^e_{bg(n,n+1)} > thr^d, \text{within } nlines^d \text{ and } nelem^d\} \tag{23}$$

At each iteration, elements to be "activated" are:

$$\mathbf{e}^l = \{e_{new} \,|\, e_{new} \text{ is near to } e, e \,|\, \alpha^e_{bg(n,n+1)} < thr^l \text{within } nlines^d \text{ and } nelem^d, e_{new} \in nlines^l\} \tag{24}$$

4 NUMERICAL EXAMPLE

In this section, the optimization procedure is illustrated starting from the well-known circle topology shown in Figure 5. This topology presents a complete bandgap in the 2D plane strain case. The parameters for the BESO optimization are set as:

- $nlines^d = 2:1$
- $nlines^l = 2:1$
- $nelem^d = 400$
- $nelem^l = 400$
- $thr^d = 0.20:0.75$ normalized w.r.t. the max absolute value of the selected elements
- $thr^l = -0.05$ normalized w.r.t. the max absolute value of the selected elements

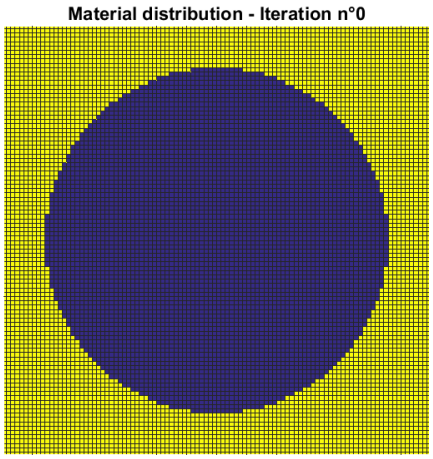


Figure 5: Material distribution for the PnC unit cell starting topology.
The hole has a circular shape of diameter 0.8 a, with a unit cell characteristic dimension (here $a = 5.5\mu m$).
The unit cell is discretized in 100x100 Q4 square FE.

It is worth to notice that the optimum engineering topology is found after 13 iterations, see Figure 6. From the starting iteration, it is possible to identify the final shape following the highlighted elements that will be removed and added (see Figure 7). The bandgap in relative terms w.r.t. the center frequency increases from 14% to 89%, as shown in Figure 8.

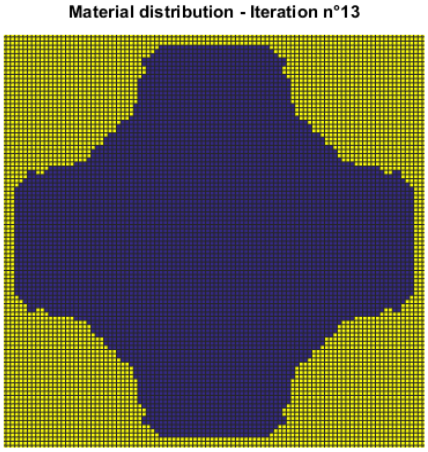


Figure 6: Material distribution for the PnC unit cell optimal topology, after 13 iterations.

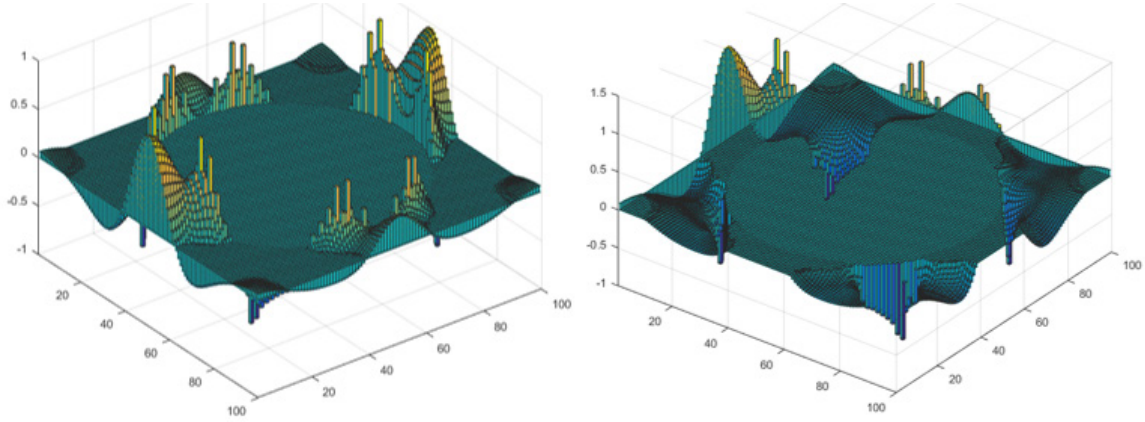


Figure 7: Distributions of $\alpha^{e}_{bg(n,n+1)}$: dying elements on left figure, living elements on right figure.

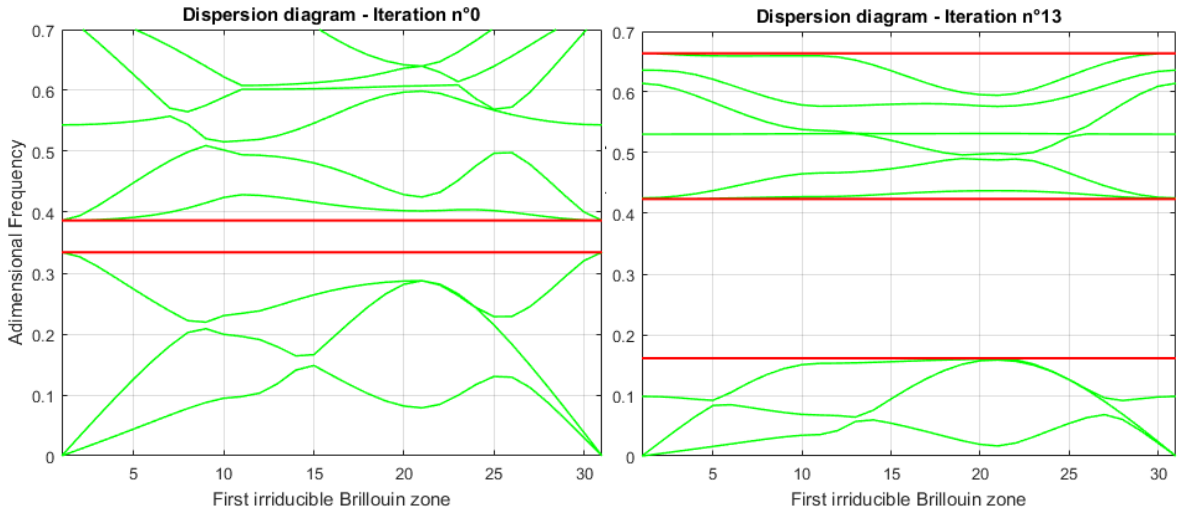


Figure 8: Comparison between dispersion relation diagrams corresponding to:
starting topology (bg,rel=14%) on the left and optimal topology (bg,rel=89%) on the right.

The dimensionless frequency is calculated as follows:

$$f_{a\,\mathrm{dim}} = \frac{f_{[Hz]}\ a_{[\mu m]}}{v_{p[\mu m/s]}}$$
$$v_p = \sqrt{\frac{\lambda + 2\mu}{\rho}} \tag{25}$$

The optimal topology of Figure 6 can be generalized also to different materials: it is only a matter of ratios between the elastic and the mass properties. In particular, assuming that the Poisson ratio varies between 0.2 and 0.3 the only important parameter is the square root of the ratio between the elastic Young Modulus and the density.

5 SENSITIVITY TO FABRICATION IMPRECISIONS

An important aspect of this work is the sensitivity with respect to fabrication imprecisions: the optimal bandgap must not be too sensitive with respect to little shape variation.

In order to prove that this optimal shape is "stable", the same optimization algorithm has been used "against" the optimal shape itself, choosing the BESO parameters to simulate the fabrication imprecisions.

This choice has been done starting from the element-mesh size and then deciding how many lines of material of the optimal shape could have been removed from a fabrication imprecision, namely two lines.

This information is set as parameter in the BESO algorithm, that chooses in a couple of iterations the worst fabrication imprecision possible in order to destroy the bandgap. Figure 9 shows the results of the analyses: owing to the fabrication imprecision, the bandgap is reduced to 77%, instead of its optimal value 89%.

6 CONCLUSIONS

In this work is described a new topology optimization procedure to find the optimal bandgap for solid/air phononic crystals.

First, the formulation is described in terms of finite element method, in particular the 2D plane strain simplification. Then the optimization procedure and algorithm are introduced,

focusing on the novelty of Bidirectional Evolutionary Structural Optimization applied to the air domain.

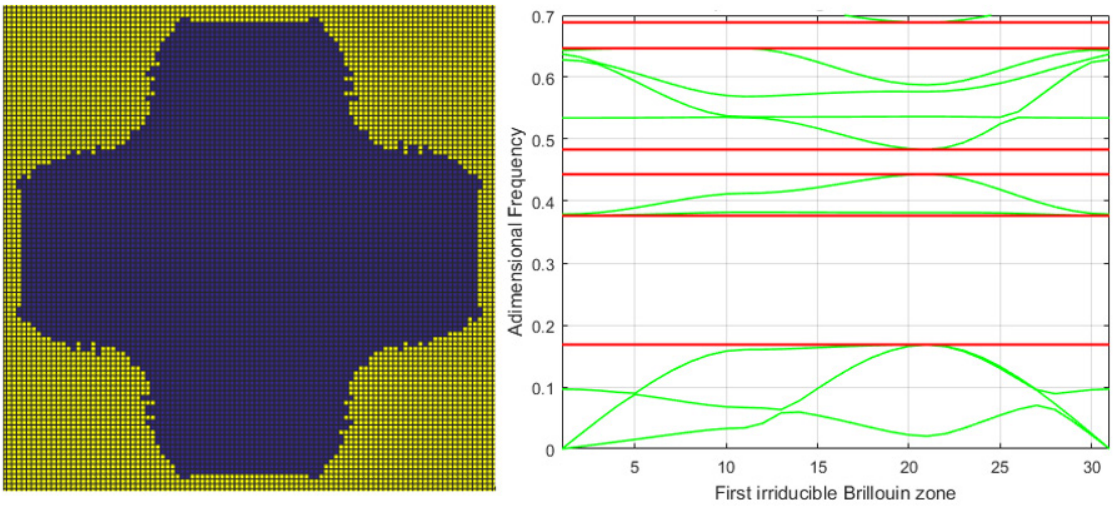


Figure 9: Results of sensitivity to fabrication process reversing the optimization algorithm: resulting topology on the left and related bandgap (bg,rel=77%) on the right.

A numerical example is then illustrated, with starting topology defined as the well known circle-hole, that presents a bandgap of 14% and the optimal topology is derived with 90% bandgap.

This result can be extended to different materials: due to the fact that the "internal" phase is not considered the bandgap is mostly influenced by the shape of the hole and material parameter ratios.

In the last section, a sensitivity analysis is introduced in order to show that the optimal shape is "stable" with respect to fabrication imprecisions: the optimization algorithm is forced "against" the optimal topology itself in order to destroy the bandgap. The result is a decrease of the bandgap from 89% to 77%, which is still a good result.

REFERENCES

[1] T. Gorishnyy, M. Maldovan, C. Ullal, and E. Thomas, "Sound ideas," *Phys. World*, vol. 18, no. 12, pp. 24–29, 2005.

[2] M. Maldovan, "Sound and heat revolutions in phononics.," *Nature*, vol. 503, no. 7475, pp. 209–17, 2013.

[3] S. Wang, L. C. Popa, and D. Weinstein, "GAN MEMS RESONATOR USING A FOLDED PHONONIC CRYSTAL STRUCTURE Massachusetts Institute of Technology , Cambridge , MA," pp. 72–75, 2014.

[4] B. Bahr, R. Marathe, and D. Weinstein, "Phononic crystals for acoustic confinement in CMOS-MEMS resonators," *IFCS 2014 - 2014 IEEE Int. Freq. Control Symp. Proc.*, no. iii, 2014.

[5] O. Sigmund and J. S. Jensen, "Systematic design of phononic band-gap materials and structures by topology optimization.," *Philos. Trans. A. Math. Phys. Eng. Sci.*, vol. 361, no. 1806, pp. 1001–1019, 2003.

[6] G. A. Gazonas, D. S. Weile, R. Wildman, and A. Mohan, "Genetic algorithm

optimization of phononic bandgap structures," *Int. J. Solids Struct.*, vol. 43, no. 18–19, pp. 5851–5866, 2006.

[7] O. R. Bilal and M. I. Hussein, "Optimization of phononic crystals for the simultaneous attenuation of out-of-plane and in-plane waves," *ASME 2011 Int. Mech. Eng. Congr. Expo. IMECE 2011*, vol. 8, pp. 969–972, 2011.

[8] H. W. Dong, X. X. Su, Y. S. Wang, and C. Zhang, "Topology optimization of two-dimensional asymmetrical phononic crystals," *Phys. Lett. Sect. A Gen. At. Solid State Phys.*, vol. 378, no. 4, pp. 434–441, 2014.

[9] F. Meng, X. Huang, and B. Jia, "Bi-directional evolutionary optimization for photonic band gap structures," *J. Comput. Phys.*, vol. 302, pp. 393–404, 2015.

[10] M. Collet, M. Ouisse, M. Ruzzene, and M. N. Ichchou, "Floquet–Bloch decomposition for the computation of dispersion of two-dimensional periodic, damped mechanical systems," *Int. J. Solids Struct.*, vol. 48, no. 20, pp. 2837–2848, 2011.

[11] M. I. Hussein, "Reduced Bloch Mode Expansion for Periodic Media Band Structure Calculations," *Proc. Math. Phys. Eng. Sci.*, vol. 465, no. 2109, pp. 2825–2848, 2009.

[12] J. H. Sun and T. T. Wu, "Propagation of acoustic waves in phononic-crystal plates and waveguides using a finite-difference time-domain method," *Phys. Rev. B - Condens. Matter Mater. Phys.*, vol. 76, no. 10, pp. 1–8, 2007.

[13] A. S. Phani, J. Woodhouse, and N. A. Fleck, "Wave propagation in two-dimensional periodic lattices," *J. Acoust. Soc. Am.*, vol. 119, no. 4, p. 1995, 2006.

[14] Y.-F. Wang, Y.-S. Wang, and X.-X. Su, "Large bandgaps of two-dimensional phononic crystals with cross-like holes," *J. Appl. Phys.*, vol. 110, no. 11, p. 113520, 2011.

[15] A. R. Diaz, A. G. Haddow, and L. Ma, "Design of band-gap grid structures," *Struct. Multidiscip. Optim.*, vol. 29, no. 6, pp. 418–431, 2005.

[16] L. Brillouin, *Wave Propagation in Periodic Structures*. New York, 1953.

[17] Y. M. Xie and G. P. Steven, *Evolutionary Structural Optimization*. 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

LOW BOOM/LOW DRAG SMALL SIZE SUPERSONIC AIRCRAFT DESIGN

Atsushi Ueno[1], Yasushi Watanabe[1], Itham Salah El Din[2], Richard Grenon[2] and Gérald Carrier[2]

[1] Japan Aerospace Exploration Agency
6-13-1 Osawa, Mitaka, Tokyo, 181-0015, Japan.
e-mail: {aueno, wata}@chofu.jaxa.jp

[2] Office National d'Etudes et de Recherches Aérospatiales
8 rue des Vertugadins, FR-92190, Meudon, France.
{itham.Salah_el_din, richard.grenon, gerald.carrier}@onera.fr

Keywords: Supersonic, Low boom/low drag design, Engine integration, Multidisciplinary design optimization.

Abstract. *The French Aerospace Lab (ONERA) and Japan Aerospace Exploration Agency (JAXA) conducted a collaborative research on low-boom and low-drag design of small size supersonic aircraft including engine integration perspective. In this collaborative research, the validation of low-boom design tools, the optimization study on engine position considering aerodynamics, acoustic, and propulsion performances, and the low-boom design of propulsion-airframe integrated (PAI) configuration were performed. In this paper, obtained results are summarized and the low-boom design of PAI configuration is focused on, because the sonic boom loudness is raised by the engine integration. The applied low-boom design tool is based on the free-form deformation and the equivalent area with consideration of the change of lift distribution due to deformations. Results show that the sonic boom loudness of PAI configuration can be reduced to almost the same loudness of glider configuration without engine by the horizontal tail optimization.*

1 INTRODUCTION

Research activities to realize environmental-friendly supersonic aircrafts are conducted all over the world [1, 2]. The major environmental constraint is the sonic boom. Pressure waves caused by objects flying at supersonic speed are integrated with each other during the propagation to the ground and finally coalesce into two shock waves called N wave, causing loud noise on the ground. The Concorde was prohibited from flying supersonically overland due to the sonic boom. Currently, ICAO (International Civil Aviation Organization) is trying to establish regulations on the sonic boom loudness and to set the threshold to allow overland supersonic flight, which accelerates research activities to realize low-boom as well as low-drag performances [3, 5]. Other environmental constraints are originated from the engine. They are the airport noise and the emission such as NOx and CO2. Reducing these environmental burdens depends largely on the engine itself. For instance, large bypass ratio decreases the engine exhaust velocity at take-off and improves the specific fuel consumption, which results in reduction of airport noise and emission. On the other hand, large bypass ratio increases engine weight and drag caused by the nacelle, which has negative impact on the emission. In addition, shock waves caused by the propulsion system potentially affect the low-boom performance. Thus, the engine should be considered in terms of integration with airframe. Based on the above discussions, the low-boom and low-drag design including engine integration perspective is the key to realize environmental-friendly supersonic aircrafts.

Japan Aerospace Exploration Agency (JAXA) conducted the flight experiment to demonstrate the low-boom design using axisymmetric bodies [6], and recorded the low-boom signature that can be used as the validation data for low-boom design tools. The French Aerospace Lab (ONERA) has experience of low-boom design through HISAC project. In order to integrate their experiences, ONERA and JAXA conducted the collaborative research on the low-boom and low-drag design including engine integration perspective from 2011 to 2016. The reference aircraft and its specifications are shown in Fig. 1. In this paper, obtained results are summarized and the low-boom design of propulsion-airframe integrated configuration is focused on.

Specification	
Length	53 m
Weight	60 ~ 70 ton
Cruise Speed	1.6 Mach
Range	3500 nm
Payload	36 ~ 50 passengers

Figure 1: Small size supersonic transport.

2 OVERVIEW OF COLLABORATIVE RESEARCH

In the collaborative research, the glider geometry designed by JAXA is used as the reference glider (Fig. 2). Design conditions are Mach number of 1.6 and lift coefficient of 0.15. The sonic boom signature of this glider shows double shock waves at front and aft booms (Fig. 2), which contributes to the low-boom performance. Here, the reflection factor on the ground is 1.9. In this research, ONERA is mainly in charge of the low-boom design based on the equivalent area. JAXA is mainly in charge of the engine integration. Work packages (WP) are shown in Fig. 3. WP1 is ONERA's activity, and low-boom design tools are validated using the experimental data acquired in the D-SEND#1 flight test conducted by JAXA [6]. WP2 is

JAXA's activity, and the engine integration is examined in terms of aerodynamics, acoustic, and propulsion perspectives. WP3 is conducted collaboratively by ONERA and JAXA to perform the low-boom and low-drag design including engine integration. Results are summarized in the following sections.

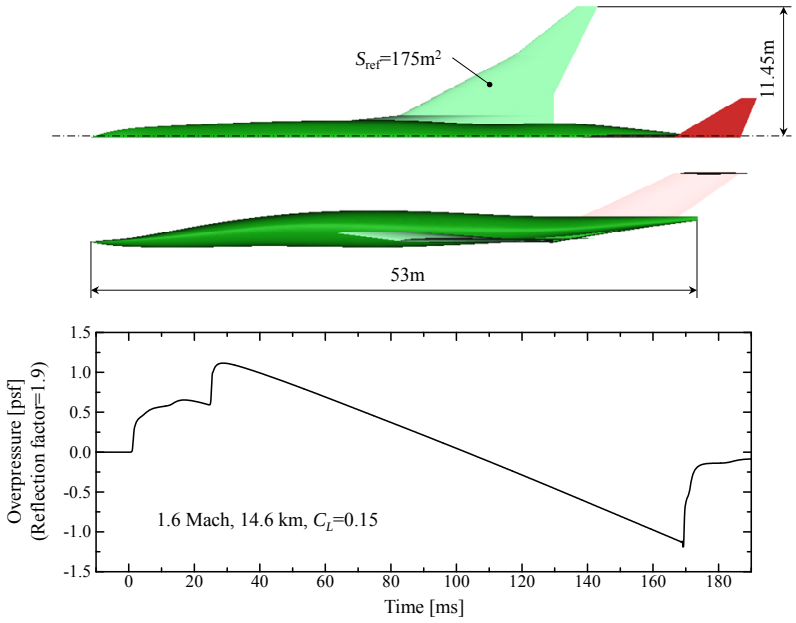


Figure 2: Reference glider and sonic boom signature.

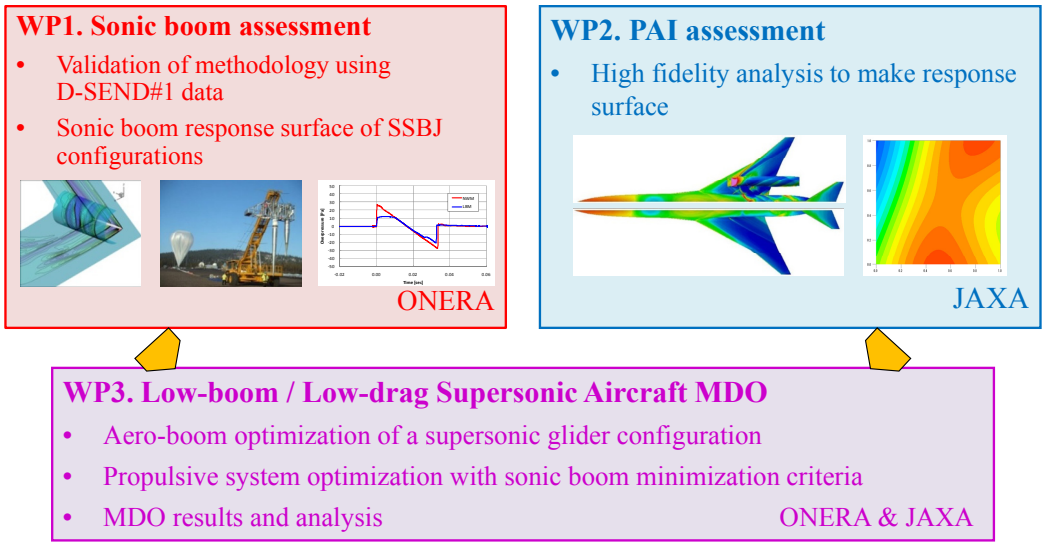


Figure 3: Work packages.

2.1 WP1: Sonic boom assessment

2.1.1. Sonic boom physics

Sonic boom signals are impulsive, high amplitude sound signals felt by a static observer at ground level. Sonic booms are the results of the nonlinear propagation through the atmosphere, down to the ground, of shock waves and pressure disturbances generated by supersonically travelling air vehicles such as aircraft, missiles or artillery projectiles [7]. The

displacement imposed by the supersonic vehicle to the air particles along its trajectory in its close vicinity are accompanied by pressure perturbations which radiate in all directions away from the vehicle along the three dimensional characteristics planes of the supersonic flow. A fraction of this pressure disturbances directed downward will eventually reach the ground after travelling through the atmosphere. During their propagation, the shape of the time signal associated to these pressure perturbations evolves under the competing non-linear and dissipative effects (Fig. 4). The non-linear effects tend to generate steep shocks through coalescence of successive pressure perturbations. On the contrary, the different dissipative effects, dominated by molecular relaxation phenomena, tend to thicken and dissipate shocks.

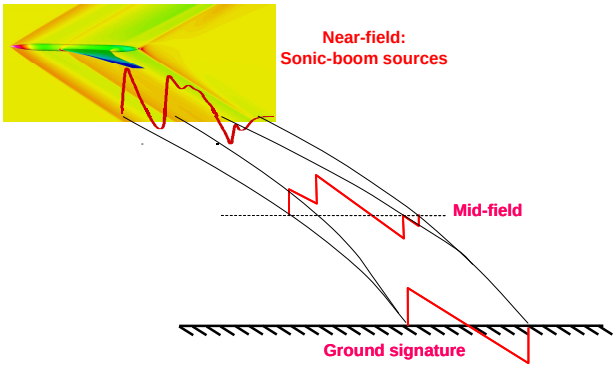


Figure 4: Sonic boom physics.

2.1.2. Methods and tools for calculating sonic boom

Simulating the sonic boom is an intricate task involving complex physical phenomena and very different scales. An accurate modeling of sonic boom first requires an adequate prediction of its sources, the aerodynamic pressure disturbances generated by the aircraft in its close vicinity. In this region, the aerodynamic flow is governed by the non-linear Euler equations and includes three-dimensional features with scales proportional to the aircraft length. Therefore CFD methods are perfectly suited and necessary to predict these near-field aerodynamic perturbations giving birth to the sonic boom. Because the prediction of the sonic boom requires propagating the pressure signal from the aircraft down to the ground over tens of kilometers, i.e. several hundreds of the aircraft length, conventional CFD methods are inadequate to perform this long distance propagation. They would require a tremendous number of mesh points and would eventually fail to capture important effect occurring during the propagation (such as the molecular relaxation). Therefore, a specific acoustic code is necessary to carry out this long distance propagation of the sonic boom signal through the standard atmosphere that is stratified and include temperature and density evolution with altitude.

The ONERA sonic boom prediction methodology is based on a three-layer approach, as illustrated in Fig. 5. Layer 1 corresponds to the near-field aerodynamic flow prediction with CFD. Layer 3 is the atmospheric propagation of the sonic boom with an acoustic code, while layer 2 ensures a natural matching between the near-field aerodynamic data and the native inputs of the acoustic code.

Layer 1: Near field aerodynamic calculation by CFD

First, the pressure perturbations in the close vicinity of the aircraft flying in supersonic cruise condition, which are the origin of the sonic boom, are calculated by solving the three-dimensional steady Euler equations. The ONERA *elsA* [8] or *Cedre* [9] CFD softwares are

used to perform these calculations. To obtain an accurate evaluation of the near-field aerodynamic pressure perturbations, a specific care must be given to the quality of the CFD mesh which must have characteristics adapted to the flow physics of these perturbations travelling along flow characteristics. The computational domain of this CFD calculation typically extends between twice and four times the aircraft length around the aircraft.

Layer 2: Multipole matching method

The CFD-aerodynamic pressure field is extracted on a cylinder surrounding the aircraft, aligned with the flow direction, whose radius is a user-specified parameter varying between one half to one body length. The pressure perturbations on this cylinder are then post-processed using the multipole decomposition method originally introduced by Plotkin and Page [11] and applied by Salah El Din [12]. This decomposition method proceeds through a development of the near field pressure signature on the cylinder according to the azimuth variable **θ.** Thereby, it allows to rebuild a Whitham function equivalent, at long distance, to the pressure perturbation generated by the aircraft, while cumulating the diffraction effects associated to the non-axisymmetrical near-field flow. The ground signature computed from this equivalent rebuilt Whitham function is observed to converge much faster with the matching distance between the CFD and the acoustic theory (i.e. the radius of the cylinder) than the ground signature from a direct CFD/acoustic would match. This justifies the use of this second layer which, further to provide a theoretically correct near-field/far-field match, greatly reduces the CFD grid size needed for the near field computation, saving significant computing time.

Layer 3: Atmospheric non-linear acoustic propagation method

Finally, the ground signature is computed by propagating the near-field aerodynamic pressure perturbations matched by the multipole decomposition method (layer 2) using the acoustic propagation code *TRAPS* [13]. The *TRAPS* code is a dedicated sonic boom propagation code based on the inviscid non-linear acoustic theory. It uses a ray-tracing approach to account for the refraction phenomena occurring during the propagation through a stratified atmosphere with vertical temperature and density gradients and to evaluate the extent of the "primary carpet", i.e. the width of the corridor underneath the aircraft trajectory directly affected by sonic boom. Along each acoustic ray, the Whitham theory (first order correction to the linear supersonic theory) is used to predict the evolution of the shape of the sonic boom pressure signal.

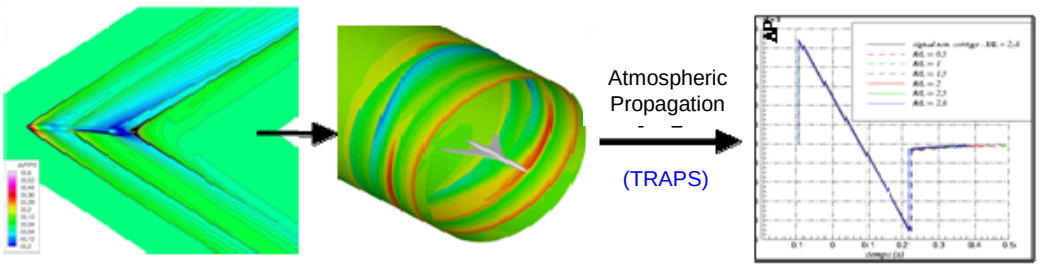


Figure 5: Three-layer sonic-boom prediction methodology.

2.1.3. Validation of the numerical sonic boom prediction tools

The sonic boom prediction tools used by ONERA have been validated using the JAXA D-SEND#1 experimental data [14]. These experiments consisted in a vertically drop test of two different models from a balloon high altitude: an N-Wave Model (NWM) and a Low-Boom Model (LBM) were tested (Fig. 6). During these drop tests, sonic booms have been measured at different locations including two microphones, one at ground level and the second at about 1km altitude, just above atmospherically turbulent boundary layer (Fig. 7).

The exact "flight" conditions of the model during its drop at the time when it generated the registered sonic boom signal on ground were determined through an analysis of the trajectory data using an inverse ray-tracing procedure. The near field aerodynamic signatures generated by both models in the previously identified flight conditions (Mach, altitude) were then computed using CFD calculations performed with the [9] software and mesh adaptations performed with INRIA mesh adaptation tool FEFLOA [10]. The sonic boom has then been computed at ground level using the methods and tools described in the previous section and the results compared to the D-SEND#1 tests data (Fig. 8).

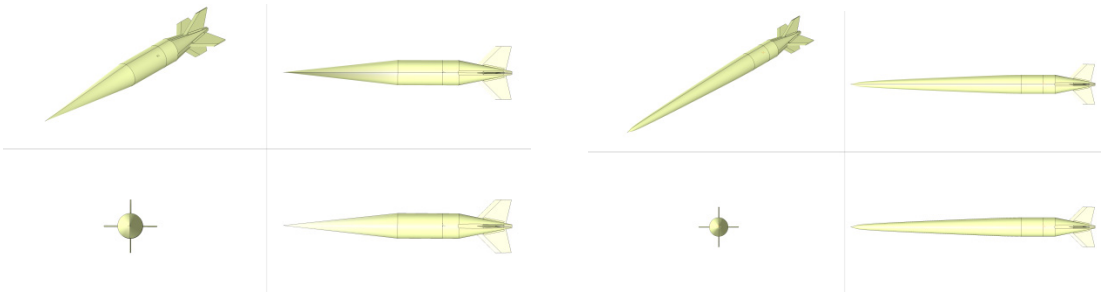


Figure 6: D-SEND#1 models geometry: conventional N-wave signature model (left) and low-boom flat top model (right).

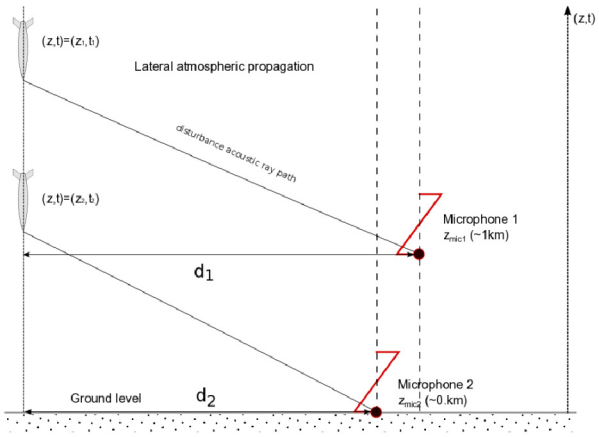


Figure 7: Schematic of JAXA D-SEND#1 drop tests sonic boom propagation: vertical descending trajectory of the model and oblique acoustic ray pointing downward; Position of the two microphones measuring the sonic boom at ground level and immediately above turbulent atmospheric boundary layer.

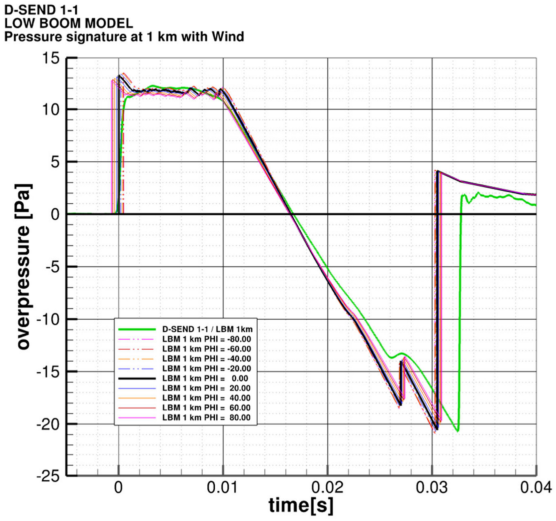


Figure 8: Validation of numerical sonic boom prediction tools using the JAXA D-SEND#1 experimental test results for the Low-boom model.

2.2 WP2: PAI assessment

In WP2, the engine integration is examined in terms of low-boom and low-drag [15, 16]. Supersonic engines generally tend to have smaller bypass ratio than subsonic engines, and increase airport noise. In this study, the engine specification was optimized to maximize the range performance while the engine exhaust velocity is constrained to be less than the specified value that can satisfy ICAO regulation (chapter 4). As a result, the bypass ratio and the fan diameter become 3.4 and 1.7 m, respectively. The inlet/nacelle geometry is designed based on the optimized engine specification and is integrated with the reference glider shown in Fig. 2. The design space of engine position is determined in this integration (Fig. 9), and response surfaces of lift-to-drag ratio and sonic boom loudness are constructed by the Euler analysis. Moreover, the engine position affects the propulsion performance such as inlet pressure recovery, thrust, and specific fuel consumption, which in turn affects the range performance. Thus, the inlet pressure recovery evaluated by the Navier-Stokes analysis is added to objective functions of response surfaces. These three objective functions are evaluated at the design lift coefficient (0.15). Response surfaces of side-body engines are shown in Fig. 10 as a typical example. There are 9 samples denoted by 01 ~ 09 in Fig. 10. Response surfaces are initially constructed by samples 01 ~ 08. Sample 09 is added considering the convergence of nondominated front in the optimization of engine position performed in WP3. Each axis in Fig. 10 is non-dimensionalized by the range of design space, that is, 0 corresponds to x_{min} and z_{min}, and 1 corresponds to x_{max} and z_{max}. The sonic boom loudness is evaluated in terms of the Stevens Perceived Level Mark VII (PL), and the difference of PL from that of the reference glider (97.1 dB) is the objective function. The inlet pressure recovery is the ratio of total pressure at engine fan face to total pressure of uniform flow. According to Fig. 10, the position of maximum lift-to-drag ratio and the position of minimum sonic boom loudness come close to each other, thus, the side-body engine can realize good compromise between low-boom and low-drag. On the other hand, the inlet pressure recovery is low at the position of maximum lift-to-drag ratio, which indicates the necessity of further investigation considering the propulsion performance. Therefore, the optimization in terms of low-boom and range performance including propulsion performance is conducted and is described in WP3.

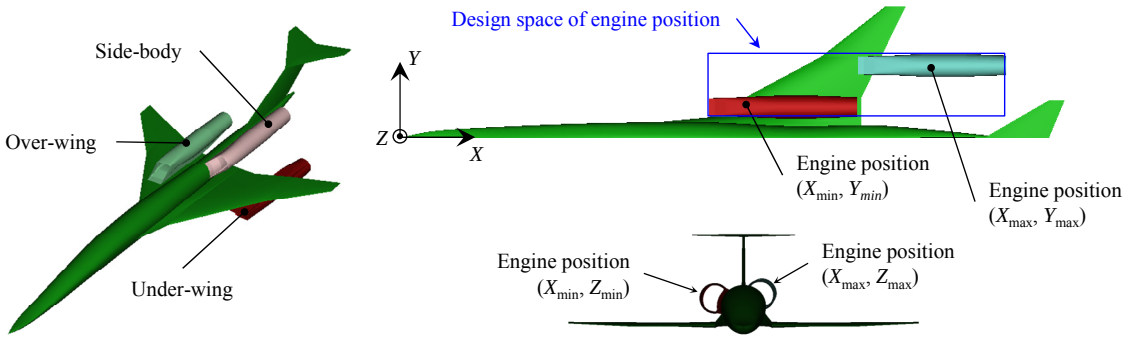


Figure 9: Design space of engine position.

Over- and under-wing		Side-body	
X_{min}	28.514 m	X_{min}	28.514 m
X_{max}	41.508 m	X_{max}	41.508 m
Y_{min}	2.756 m	Z_{min}	1.908 m
Y_{max}	6.306 m	Z_{max}	2.332 m

Table 1: Design space of engine position.

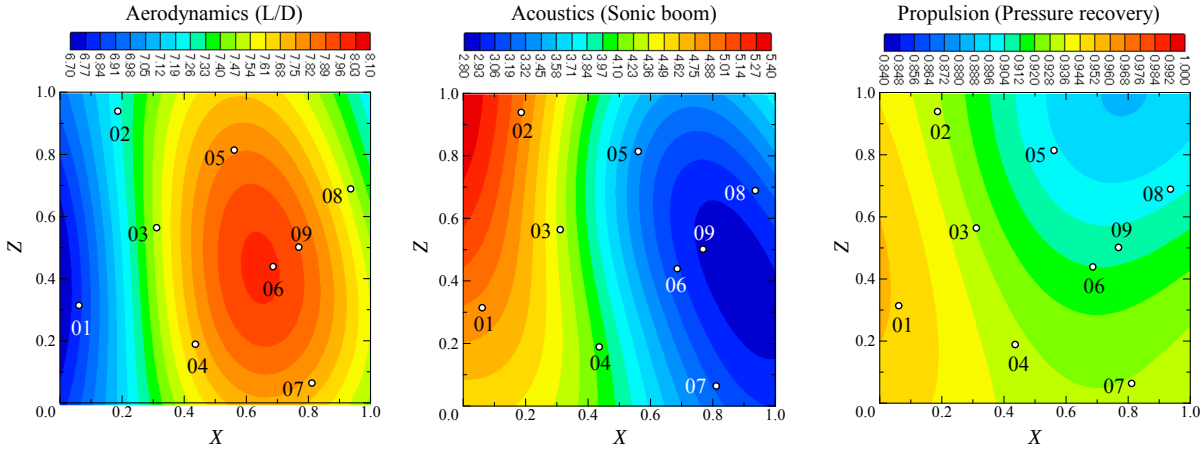


Figure 10: Response surfaces of side-body engine.

2.3 WP3: Low-boom/Low-drag Supersonic Aircraft MDO

Work package 3 includes the low-boom design of the reference glider that was conducted by ONERA and the optimization of engine position using response surfaces in WP2 that was conducted by JAXA. The low-boom design of engine-airframe integrated configuration based on results obtained here is described in the next chapter.

Aero-boom optimization of a supersonic glider configuration

A sonic boom minimization process of the initial wing-body-tail (glider) configuration has been performed by ONERA. This has been done using a two-step approach:

1. First, an inverse design approach, AIDA [4], implementing the Jones-Seebass-Darden shaped sonic boom theory, has been used to generate the minimum sonic

boom equivalent area ruling corresponding to the JAXA small supersonic aircraft characteristics (Mach number, fuselage length, weight, wing position, …);

2. Then, starting from the actual geometry of the initial glider configuration, a free-form deformation approach has been used to modify the actual fuselage geometry in order to reach the target equivalent area distribution issued from step 1). This was done through a numerical optimization process that used the vertical displacement of the different control points of the free-form lattice embedding the fuselage geometry as design parameter (Fig. 11). In this optimization process, only the impact of the geometry modifications on the equivalent area distribution due to volume was evaluated (the contribution of lift to the equivalent area was assumed to be constant).

The results of this design process applied to the glider configuration are shown in Fig. 12. The front part of the equivalent area was successfully mapped to the target one obtained by inverse design with AIDA module [4]. This enables to obtain a ground sonic boom whose front part resembles the target flat top signal even if it includes a succession of small shocks. The aft part of the signature was not significantly modified due to a limited design freedom (the control box was defined in order to keep the tail geometry unchanged).

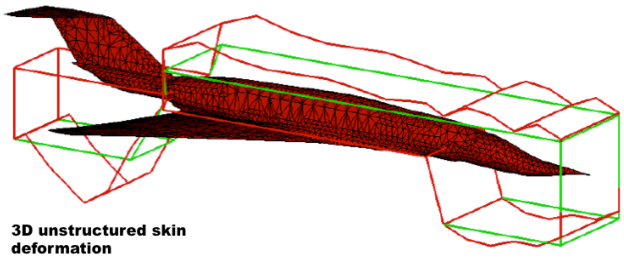


Figure 11: Free-form deformation control lattice used to optimize the fuselage geometry in order to reach the target equivalent area distribution that provides optimal sonic boom signature.

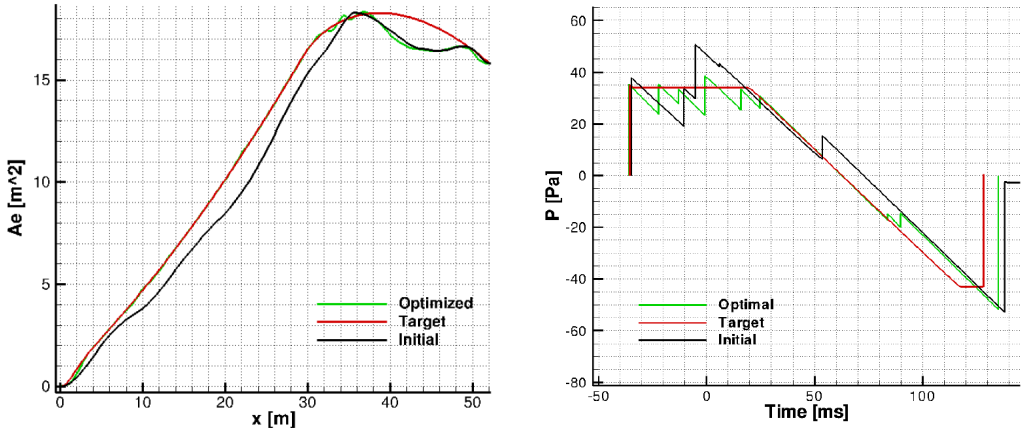


Figure 12: Left: initial, optimized and target equivalent area distribution. Right: ground propagated sonic boom signals corresponding to the initial glider, optimized glider and target optimal (flat-top) sonic boom.

Latter, such a similar design procedure using a free-form deformation parameterization and a target equivalent area distribution was reapplied to the complete aircraft geometry including propulsion system (Fig. 13).

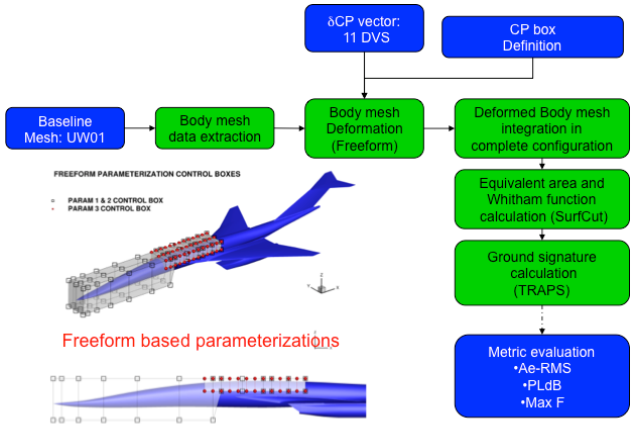
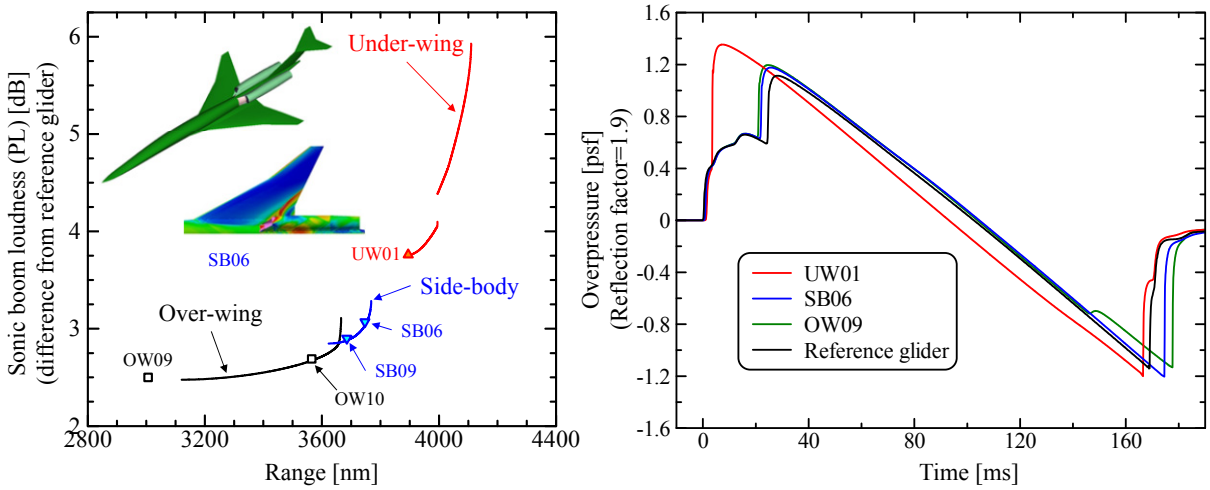


Figure 13: Application of a free form deformation based optimization process to reduce the sonic boom of the aircraft configuration with propulsion system.

Propulsive system optimization with sonic boom minimization criteria

The optimization of engine position is based on the response surfaces obtained in WP2 (Fig. 10) [16], which means that the glider geometry is fixed at the reference glider geometry and that the glider optimization tool mentioned above is not applied. In this optimization, two objective functions are set. One is the sonic boom loudness that is obtained directly from the response surface. The other is the range performance that is evaluated by the Breguet range equation using the lift-to-drag ratio and the specific fuel consumption that includes the performance degradation caused by the inlet pressure recovery. The nondominated front is shown in Fig. 14.



(a) Nondominated front with respect to sonic boom and range. (b) Sonic boom signature.

Figure 14: Optimization of engine position.

The under-wing engines realize large lift-to-drag ratio thanks to the compression lift produced by the nacelle shock wave, and realize large inlet pressure recovery because the stream line under the wing is nearly parallel to the inlet axis. Thus, the under-wing engines realize better range performance. However, the nacelle shock wave propagates to the ground and causes louder sonic boom. The over-wing engines show the best low-boom performance by shielding the nacelle shock wave by the wing upper surface, but the cross-flow over the wing degrades the inlet pressure recovery. Smaller lift-to-drag ratio due to the negative lift caused by the na-

celle shock wave along with low inlet pressure recovery leads to low range performance. The side-body engines overcome drawbacks mentioned above. The nacelle shock wave acts around the trailing edge of the wing, which reduces the wing pressure drag and prevents the nacelle shock wave from propagating to the ground. Thus, the low-boom and low-drag are realized at the same time. The inlet pressure recovery is relatively low due to the thick boundary layer on the fuselage; however, the side-body engines show better compromise between low-boom and long-range.

Results mentioned above are integrated, in other words, the glider optimization tool is applied to the side-body engine in order to realize further low-boom performance. The optimization results are described in the next section.

3 AFT BOOM SUPPRESSION BY HORIZONTAL TAIL OPTIMIZATION

Prior to the low-boom design, the sample 06 of side-body engine denoted by SB06 (Fig. 14) is slightly modified in order to reduce the wave drag. The engine position of modified SB06 denoted by PAI (propulsion-airframe integrated) configuration is almost the same as that of SB06, moved rearward and upward by 0.5% and 0.2% of fuselage length, respectively. The engine nozzle of SB06 is tangential to the fuselage at its exit. On the other hand, the part of engine nozzle (7% of the nozzle exit diameter) of PAI configuration is buried into the fuselage, and the wave drag is reduced due to the reduction of fuselage volume. The PAI configuration has larger lift-to-drag ratio by 3.2% than SB06. The PAI configuration and SB06 show almost the same sonic boom signature (Fig. 14), causing louder sonic boom than the reference glider due to the single shock wave at the aft boom. The engine nacelle produces negative lift, and the angle of attack is increased to meet design lift coefficient. As a result, the expansion over the wing is enhanced and the peak negative pressure at the aft boom is increased. Another reason to the louder aft boom is the expansion wave at the rear part of the nacelle. The shock wave at the leading edge of the horizontal tail plays an important role in fragmenting the aft boom into double shock waves. In the PAI configuration, the leading edge shock wave is weakened by the expansion wave originated from the nacelle (Fig. 15), and pressure waves coalesce into single shock wave at the aft boom. As shown in Fig. 14, over-wing engines realizing quieter sonic boom than side-body engines also show the single shock wave at the aft boom. These results indicate the importance of the aft boom suppression of propulsion-airframe integrated configurations. Thus, the low-boom design of the PAI configuration is conducted by applying the glider optimization tool described in WP3.

Equivalent area distributions of reference glider and PAI configuration are shown in Fig. 16. Both configurations have the peak value of equivalent area at x of around 60 m. This peak is made up by the shock wave and expansion wave of the horizontal tail. The aft boom of the reference glider is fragmented into double shock waves by the shock and expansion waves corresponding to the front and rear parts of the peak, respectively. This fragmentation is the key to reduce the aft boom loudness. In the PAI configuration, the shock wave is weakened by the expansion wave generated at the rear part of the nacelle as described above, which lowers the peak and degrades low-boom performance. To suppress the aft boom of the PAI configuration, the target equivalent area distribution with higher peak is determined (Fig. 16), and the profile of the horizontal tail that is originally the biconvex is optimized by free-form deformation (FFD) to meet the target.

In this optimization study, the low-boom design tool applied in WP3 is updated to consider the change of lift distribution, because the horizontal tail has impact on the lift. In this update, the change of lift caused by the deformation of horizontal tail is evaluated by the simple method, that is, the conical flow theory and the Prandtl-Meyer expansion. Firstly, the horizon-

tal tail surface is divided into triangle panels, and the angle between the panel and uniform flow is calculated for each panel. When the panel is facing the uniform flow, the pressure coefficient is evaluated by the conical flow theory. Otherwise, the pressure coefficient is evaluated by the Prandtl-Meyer expansion. After the deformation, the pressure coefficient is evaluated in the same manner, and the difference of the pressure coefficient due to the deformation gives the change of lift. The flow diagram of the updated tool is shown in Fig. 17. The horizontal tail is deformed by FFD and the volume component of equivalent area (AE_vol) is calculated. Then, the difference of the pressure coefficient is evaluated and the lift component of equivalent area (AE_lift) is calculated. The sum of AE_vol and AE_lift gives the equivalent area (AE), and FFD is iterated until the AE meets the target equivalent area (AE_target). After the iteration, the Euler analysis is conducted for the optimized shape to evaluate the low-boom performance.

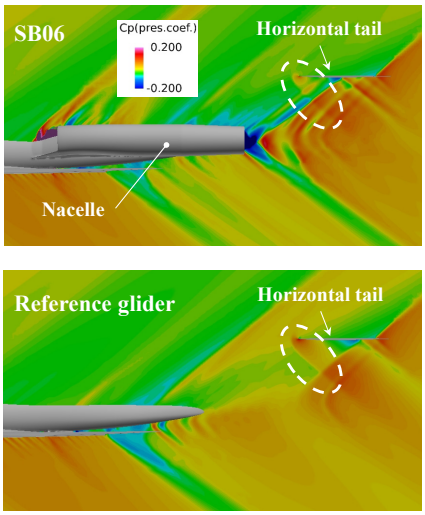


Figure 15: Pressure contour plot around nacelle.

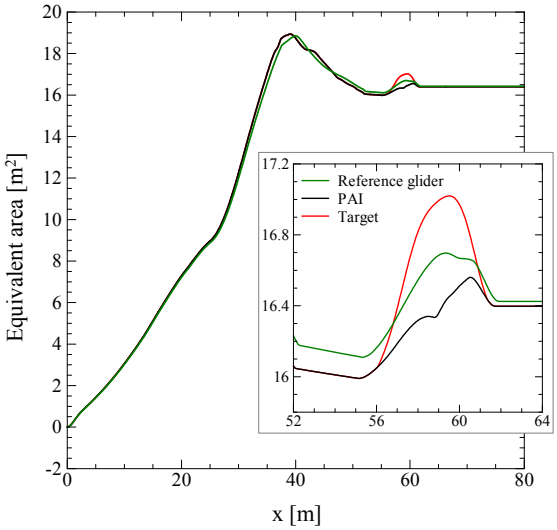


Figure 16: Equivalent area distribution.

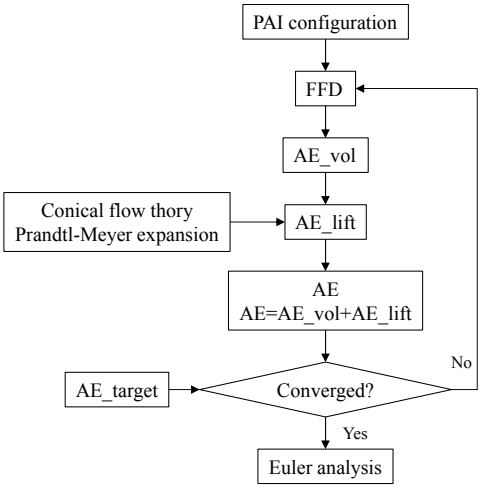


Figure 17: Flow diagram of horizontal tail optimization.

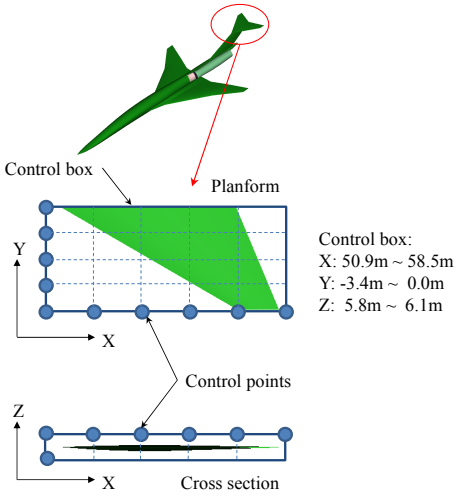


Figure 18: Free form deformation.

In FFD optimization, the rectangular parallelepiped control box is set to enclose the horizontal tail (Fig. 18). Numbers of control points are 6 in the *x* direction, 5 in the *y* direction, and 2 in the *z* direction. Design variables are the displacement of control points in the *z* direc-

tion, except for the control points on the symmetric surface (i.e., y=0). In order to check the convergence, the difference of equivalent area from the target is calculated at 81 sections between x of 56.0475 m and 61.3475 m where the equivalent area is affected by the horizontal tail deformation. The root mean square of these differences is the objective function (F). In this optimization, the drag increment due to the deformation is added to the objective function in order to realize both low-boom and low-drag (Eq. 1).

$$F = \left(\frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{AE}_i - \mathrm{AE\_target}_i \right)^2 \right)^{0.5} + 10 \cdot \Delta C_D \tag{1}$$

Here, N is the number of sections where the difference of equivalent area from the target is calculated and is 81 as described above. The objective function is minimized by the optimizer based on the genetic algorithm. The constraint function is not defined, however, the lift coefficient is implicitly constrained to be unchanged, because AE_target has the same equivalent area at the tail end (at x=63 m) as the PAI configuration.

As a result of the optimization, the root tail-section has larger thickness to chord ratio and larger incidence angle (Fig. 19), causing larger lift and stronger shock wave. At the tip section, the incidence angle becomes smaller to cancel the lift increment at the root section, because lift coefficient is implicitly fixed at design C_L (0.15). The Euler analysis of the optimized shape shows that the lift coefficient of horizontal tail is changed by only 0.2%, which demonstrates the effectiveness of simplified method in the evaluation of lift change. Thanks to these changes of horizontal tail profile, the peak value of equivalent area is raised to meet the target (Fig. 20), and the aft boom is successfully fragmented into double shock waves (Fig. 21). The aft boom loudness, which is evaluated using the pressure signature where the pressure is negative, is reduced by 4.5 dB. The overall loudness is reduced by 2.9 dB. On the other hand, the drag is increased by 3 drag count mainly due to the thicker profile at the root section, though the smaller incident angle at the tip section contributes to the drag reduction. The nondominated front with respect to lift-to-drag ratio and sonic boom loudness obtained in WP3 is shown in Fig. 22. In this figure, the PAI configuration and the optimized shape are added. The reduction of lift-to-drag ratio is 1.8%. Though the optimization without the drag increment in the objective function (Eq. 1) is not performed, the nondominated front with respect to low-boom and low-drag is expected to be obtained by changing the scale factor of drag increment (10 in this study). As for the sonic boom loudness, it is largely reduced to almost the same level of the reference glider, which indicates the effectiveness of the low-boom design tool.

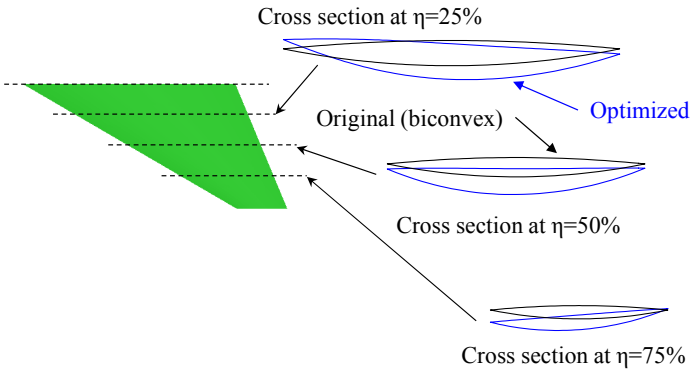


Figure 19: Optimized profile of horizontal tail.

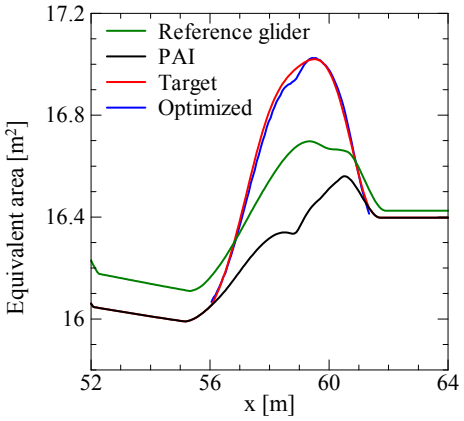


Figure 20: Optimized equivalent area distribution.

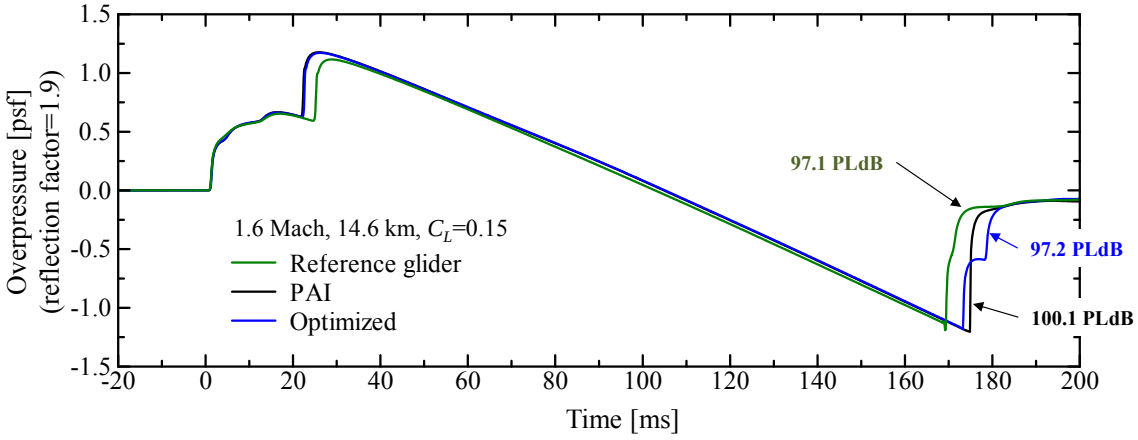


Figure 21: Optimized sonic boom signature.

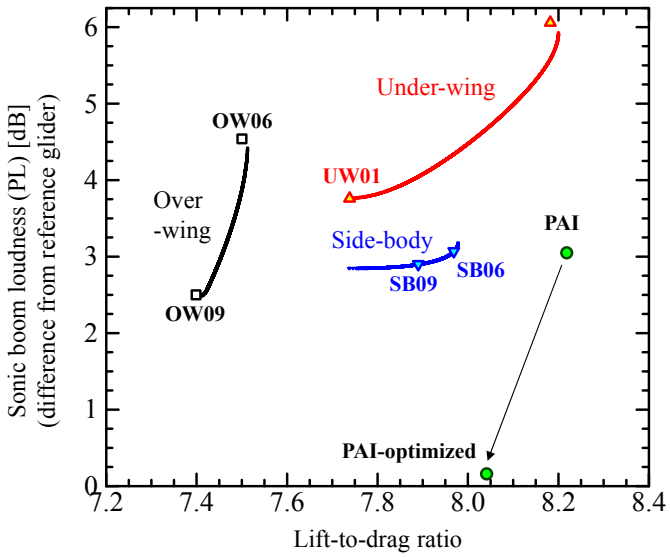


Figure 22: Nondominated front with respect to sonic boom and lift-to-drag ratio.

4 CONCLUSION

The low-boom and low-drag design of small size supersonic aircraft including engine integration perspective was conducted collaboratively by ONERA and JAXA. The validation of the numerical sonic boom prediction methodologies implemented at JAXA and ONERA has first been conducted using the experimental results of the D-SEND#1 drop test. This validation exercise showed an excellent code-to-code comparison and a good prediction of the experimentally measured sonic booms. These sonic boom prediction tools were then used in a numerical optimization process to improve the low boom characteristics of a initial glider aircraft configuration. Regarding the engine integration, the side-body engine out of three engine layouts (i.e., over-wing, under-wing, and side-body engines) showed the best compromise between low-boom and low-drag by utilizing the nacelle shock wave. The engine integration has large impact on the sonic boom loudness especially at the aft boom showing a single shock wave, even when the engine is mounted on side-body. The equivalent area distribution revealed the importance of the shock wave generated at the horizontal tail in reducing the aft boom loudness. Thus, the profile of horizontal tail was optimized to meet the target equivalent area distribution that was determined to realize double shock waves at the aft boom. The low-

boom design tool with consideration of the change of lift distribution successfully fragmented the aft boom into double shock waves by strengthening the leading edge shock wave at horizontal tail. The optimized shape realizes almost the same sonic boom loudness as the reference glider, which indicates the effectiveness of the low-boom design tool.

REFERENCES

[1] P. Coen, Supersonics Project Overview, Fundamental Aeronautics Program Technical Conference 2012.

[2] A. Murakami, Research Activities on Supersonic Technology at JAXA, The 2010 Asia-Pacific International Symposium on Aerospace Technology, 2010.

[3] A. Minelli, I. Salah el Din, G. Carrier, A. Zerbinati, J-A. Desideri, Cooperation and Competition Strategies in Multi-objective Shape Optimization - Application to Low-boom/Low-drag Supersonic Business Jet, 31st AIAA Applied Aerodynamics Conference, AIAA 2013-2678, 2013.

[4] Andrea Minelli. Aero-acoustic Shape Optimization of a Supersonic Business Jet. PhD Thesis, Analysis of PDEs [math.AP]. Université Nice Sophia Antipolis, 2013. HAL Id: tel-00938396, https://tel.archives-ouvertes.fr/tel-00938396v2, Submitted on 19 Feb 2014

[5] Y. Makino, K. Suzuki, M. Noguchi, and K. Yoshida, Nonaxisymmetrical Fuselage Shape Modification for Drag Reduction of Low-Sonic-Boom Airplane, AIAA Journal, Vol. 41, No. 8, 2003.

[6] M. Honda, K. Yoshida, D-SEND Project for Low Sonic Boom Design Technology, 28th Congress of the International Council of the Aeronautical Sciences, 2012.

[7] Plotkin, K. J., and Maglieri, D. J., Sonic Boom Research: History and Future, AIAA Paper 2003-3575, 2003.

[8] L. Cambier and M. Gazaix, elsA: An Efficient Object-Oriented Solution to CFD Complexity, AIAA 2002-0108, Reno, January 14-17, 2002.

[9] A. Refloch, B. Courbet, A. Murrone, P. Villedieu, C. Laurent, P. Gilbank, J. Troyes, L. Tessé, G. Chaineray, J.B. Dargaud, E. Quémerais, F. Vuillot, The CEDRE Software, AerospaceLab Journal, AL02-1, Issue 2 - March 2011 (http://www.aerospacelab-journal.org/sites/www.aerospacelab-journal.org/files/AL2-11.pdf)

[10] Adrien Loseille and Frédéric Alauzet. Continuous mesh framework, Part I: well-posed continuous interpolation error. SIAM Journal on Numerical Analysis, 49(1):38{60, 2011}

[11] K. Plotkin, J. Page, Extrapolation of sonic boom signatures from CFD solution, AIAA 2002-0922.

[12] I. Salah el Din, Contribution à l'optimisation de la forme aérodynamique d'un avion de transport supersonique en vue de la réduction du bang, Thèse de doctorat, Université de Poitiers, Décembre 2004.

[13] A.D. Taylor ; *The TRAPS sonic boom program* ; NOAA Technical memorandum ERL-ARL-87, July 1980 ;

[14] JAXA D-SEND#1 : http://www.aero.jaxa.jp/eng/research/frontier/sst/d-send.html

[15] A. Ueno, Y. Watanabe, Propulsion/airframe Integration Considering Low Drag and Low Sonic Boom, 29th Congress of the International Council of the Aeronautical Sciences, 2014.

[16] A. Ueno, Y. Watanabe, Multidisciplinary Design Optimization of Engine Layout for Supersonic Airliner. Koku-Uchu-Gijyutsu (Aerospace Technology Japan, the Japan Society for Aeronautical and Space Sciences), Vol. 14, 2015. (in Japanese)

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DEVELOPMENT OF PHYSICAL METHODS OF THE SUPERSONIC AIRPLANE NEAR-FIELD INVESTIGATION AIMED AT THE SONIC BOOM MINIMIZATION

Sergey L. Chernyshev[1], Alexander I. Ivanov[1], Andrey Ph. Kiselev[1], Vladimir E. Mosharov[1], Leonid L. Teperin[1]

[1]Central Aerohydrodynamic Institute n.a. prof. N.E. Zhukovsky (TsAGI),
1, Zhukovsky st., 140180, Zhukovsky, Moscow reg., Russia
ivanov_a_i@list.ru

Keywords: Sonic Boom Minimization, Supersonic Wind Tunnel, Pressure-Sensitive Paints, Near Field, Computational Methods.

Abstract. *New facility for experimental and computational studies of sonic boom from aircraft is designed and manufactured in TsAGI. The equipment was mounted in supersonic wind tunnel with 0.6×0.6 m^2 cross section. Tests were performed at Mach numbers 1.75, 2.0 and 2.25 with the use of two different models, the first being the cone-cylinder body and the second is the simplified supersonic airplane configuration with the triangle wing. The method implies the application of fluorescent pressure-sensitive paints for measurements of pressure distribution in the near field of the model. The experimental facility was designed using modern CFD approaches. Pressure distributions obtained in the near field of the model using the experimental method are compared with numerical simulations based on the software package ANSYS CFX. This comparison enables to estimate the possibilities of both the experiment and numerical simulation in prediction of the shock waves formation in the near-field of the disturbed flow.*

1 INTRODUCTION

The problem of the sonic boom influence on people and the environment became acute at the beginning of the 70th of the last century when the Soviet Tu-144 and the British-French "Concord" supersonic jets appeared in the sky. Flights of the civil passenger aircrafts with their regularity and proximity to big cities leads the sonic boom problem into the focus of interest. The opinion of its low importance and far-fetched nature was quickly replaced by anxiety and introduction of restrictive measures [1].

In the USA, the Federal Aviation Agency (FAA) prohibited flights of aircrafts at the supersonic speeds over the populated areas. The flights of aircrafts which can generate the sonic boom approaching the USA territory were also forbidden. In France and the UK such restrictions were not introduced at the state level, but the airlines did not allow flights with the sonic boom creation over the densely populated regions of these countries.

In the Soviet Union after the beginning of Tu-144 aircraft exploitation the state standard (GOST 23552-79) was introduced which allowed the flights of the supersonic commercial aircrafts over the populated regions. The sonic boom level of 90 Pa ±20 Pa in the real atmosphere conditions was pointed out as acceptable one.

After certain weakening of interest to the supersonic transport in the 80th due to its economical inefficiency, at the beginning of the 90th leading aviation countries and companies developed the investigations of possibility and expediency of creation of the second generation supersonic transport (SST-2). The topicality of this task is not decreasing nowadays. Special interest is shown in the small-size supersonic business jets (SSBJ) with the mass of 40 to 60 tons.

The aviation community closely approached to practical realization of the idea to create a new type of the airplane. Researches in this direction are made by "Lockheed-Martin" and "Gulfstream" companies in the USA, and by "Dassault Aviation" in the EU. In Russia such projects are supported by Sukhoi and Tupolev leading aviation design bureaus in cooperation with Central Aerohydrodynamic Institute (TsAGI) and Central Institute of Aviation Motors (CIAM).

Toughening of ecological demands to future generation of the aviation transport implies first of all the solution of the sonic boom problem for supersonic civil aviation. The International Civil Aviation Organization (ICAO) ordered airlines to avoid unacceptable situations for people with the sonic boom (ICAO resolution A33-7, 1998) and requested the ICAO Council to take action to achieve an international agreement about measuring the sonic boom, quantitative and qualitative definition of the expression "unacceptable situation for the population" and establish appropriate limits (ICAO resolution A38-18, 2013).

Since the requirements to aircraft configurations for reducing sonic boom and for increasing the aerodynamic quality are in contradiction, so for configurations with reduced level of sonic boom the problems occur to ensure acceptable aerodynamic characteristics and, consequently, economic efficiency of the aircraft.

Theoretical analysis of the phenomenon of sonic boom has now reached a sufficiently high level. In particular, in TsAGI algorithms and computer programs are developed on the basis of which formulated recommendations for the formation of the SST configurations of different classes with a reduced level of sonic boom. The redistribution of the fuselage thickness, the deformation of its axis and different V-shape of the wing root and tip parts are used [2].

2 METHODS OF EXPERIMENTAL STUDIES OF THE SONIC BOOM

At present, in view of the difficulties in direct numerical simulation of the sonic boom phenomena, the calculation is based on an integrated approach. This approach consists in de-

fining the parameters of disturbed flow in near field region using numerical methods based on integrating Euler equations or Reynolds-averaged Navier-Stokes (RANS) equations, and then calculating the parameters of the sonic boom at greater distances from the aircraft using the quasilinear theory [3]. When performing such calculations, a number of problems appear associated with both the description of the complex geometry of the aircraft and long computational domain (hundreds of plane lengths), where levels of disturbed pressure vary by several orders of magnitude. In this regard, the selection of complex configuration that provides low level of sonic boom and acceptable aerodynamic efficiency should go hand in hand with experimental research aimed at obtaining more reliable information.

Full information about the parameters of the sonic boom wave generated in the real atmosphere, may be provided by the flight experiment. However, at the stage of research on the formation of the aircraft configuration, flight tests are extremely expensive, and their results are heavily dependent on atmospheric conditions, precision of the flight regimes simulation and other factors, and the information obtained is limited to existing types of aircrafts.

Wind tunnels (WT) allow simulate with good accuracy the appearance of waves of sonic boom that provides the opportunity to explore:

− the influence of the body shape and flow regime;
− the influence of the lift force;
− the influence of spatial effects;
− the influence of flow non-uniformity;
− the interaction of shock waves;
− minimization of the sonic boom.

3 SONIC BOOM SIMULATION IN THE WIND TUNNELS

The possibility of modeling in WT the area of disturbed flow generated by a model of the aircraft, are determined by the transverse dimensions of the test section and the characteristic dimensions of the model. The size of this area (height) is determined by the similarity parameter, equal to the ratio of the realized model distance from the measuring base *(H)* to the length (or diameter) of model - $K = H/L$. The size of the near field is about the length of the aircraft $(K \sim L)$. The far zone, where the profile shape of the perturbed pressure acoustic shock wave is transformed into *N*-shaped wave, and the parameters are changed virtually asymptotic is implemented at distances corresponding to hundreds lengths of the aircraft. The middle zone, being intermediate between the near and far zones, is characterized by the presence of the intermediate shock waves generated by the aircraft layout basic elements, such as wing consoles and nacelles. When the model length $L = 0.1$ m, even for implementing the removal coefficient $K = 100$ the transverse dimension of the supersonic flow should be 10 m, which indicates a limited ability of the sonic boom modeling in existing WT.

3.1 Experimental and computational method of the sonic boom modeling

The combined experimental and computational method allowing to determine full disturbed flow field generated by the supersonic airplane was developed in ITAM SB RAS [4, 5].

The method is based on measurement of the flow characteristics in the near field of the model mounted in the wind tunnel test section, and then calculating the evolution of the measured pressure profiles at the long distances considering 2D or 3D character of flow [6].

Experimental part is based on modeling of the sonic boom near zone in the WT of small dimensions T-313 (0.6 × 0.6 m) and T-325 (0.2 × 0.2 m) with relatively large models. In this case, the disturbed pressure measurements near model are performed either with a special

plate with array of pressure taps, or using the measuring probe in a dynamic mode (i.e. during continuous movement of the model relative to the probe).

3.2 Implementation and development of experimental and computational method of the sonic boom modeling in TsAGI wind tunnels

TsAGI assumes the introduction and development of the experimental-computational method of the sonic boom modeling in large industrial wind tunnels. At the first phase, this method is expected to be introduced into T-113 WT (similar to ITAM T-313).

For this purpose, the bench for implementation of the experimental and computational method of the sonic boom studying was designed and mounted in T-113 wind tunnel. As stated above, for the static pressure field measuring in similar situation ITAM specialists used the coordinate mechanism with a probe or a special plate with pressure taps rows. In TsAGI the static pressure fields on the measuring plate surface were measured using the fluorescent pressure transducers (pressure-sensitive paints, PSP). Application of the PSP method should in principle ensure the acquisition of continuous pressure distribution on the control surface in the near-field of the model, calculation of Zhilin's integral [7] on its basis and definition of the sonic boom wave profile at large distances using the quasi-linear theory. Yu. L. Zhilin showed that the desired asymptotically-remote solution at the certain assumptions may be associated with the near-field integral over a surface S_2, placed at a relatively small distance from the airplane (about 0.1-0.5 of its length, Figure 1).

The surface enveloping the control volume (Figure 1) is as follows. The surface S_1 is a boundary between the disturbed and undisturbed flow (Mach cone), S_2 plane is parallel to the x axis is located under the tested body without touching it. On this plane the isolated segment AB is outlined with points A and B lying on the intersection of plane S_2 and surface S_1. Planes S_3 and S_4 are the envelopes of the inverse Mach cones emanating from the segment AB.

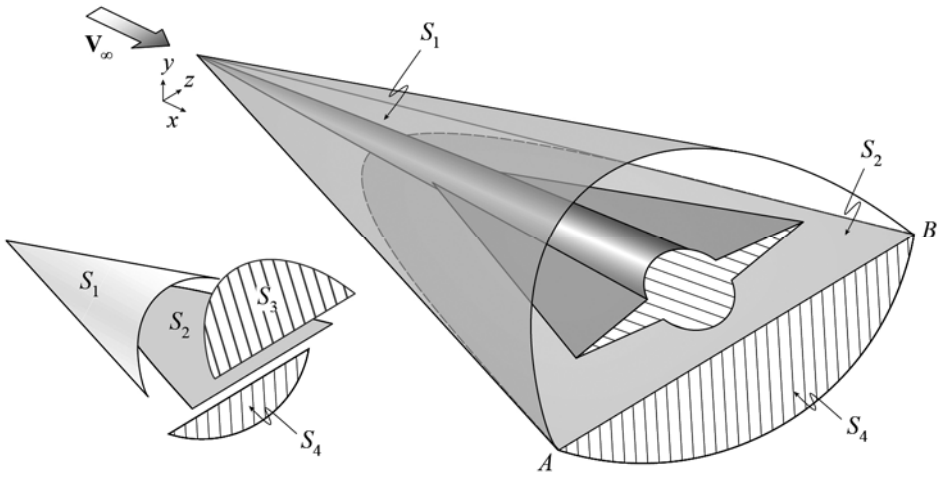


Figure 1: Control volume of the supersonic flow over the aircraft configuration.

In the case when the plane S_2 and the segment AB are perpendicular to the axis of y, for the direction "straight down", Zhilin's integral gives the following expression for calculation of derivative of the cross sectional area of the equivalent body of rotation S_{eq}:

$$\frac{dS_{\text{eq}}(x_0)}{dx_0} = -\frac{1}{V_\infty}\int_{z(A)}^{z(B)}(\beta u + v)dz = -\frac{2\beta}{V_\infty}\int_{z(A)}^{z(B)} u dz = -\frac{2}{V_\infty}\int_{z(A)}^{z(B)} v dz. \quad (1)$$

Here x_0 — coordinate of the point of intersection of the S_3 with the axis x; u and v are components of the vector of the perturbed velocity along the axes x and y, $\beta = \sqrt{M_\infty^2 - 1}$. For an arbitrary direction the coordinate system (x, y, z) may be selected so that the plane S_2 and the segment AB are perpendicular to the axis y.

An important feature of the formula (1) is that at the fixed position of the plane S_3 it is possible to change the distance of the plane S_2 from the x-axis. For example, we can place this plane at a sufficient distance from the body so that the provisions of the linear theory would be valid ($R \sim 0.3...0.5\ L$), but at the same time a lot closer than it is required to establish the asymptotic behavior of the solution ($R \sim 3...5\ L$).

PSP-method is based on the phenomenon of quenching of the organic luminophors luminescence by air oxygen. For implementing the method, the studied surface is covered with special paint which is a thin layer of polymer permeable to oxygen and containing molecules of the luminophor. The luminophor is excited by light of appropriate wavelength and then the luminescence intensity or the life time of excited molecules is measured. The luminescence intensity and lifetime is inversely proportional to the pressure. To control the accuracy of the PSP-method the measuring plate has a row of pressure taps in the central section.

The designing of the test bench in T-113 wind tunnel was carried out with the assistance of modern software tools of computational fluid dynamics. The main goal of mathematical modeling was to predict the flow structure in the near field of the model and calculate the distributions of the overpressure generated on the measuring plate. The simulation was carried out using the software ANSYS CFX [8] (TSAGI license No. 501024), based on solving the Reynolds-averaged Navier-Stokes equations.

The shock waves propagating from the model fall on the plate interact with the boundary layer and generate reflected shock waves. There is a complex three-dimensional flow pattern and interaction of shock waves with the boundary layer on the measuring plate and with each other. In order to adequately resolve the most important (from the physical point of view) phenomena, in the regions of interaction between shock waves and boundary layer on the plate surface the construction of appropriate mesh is required. An example of such a computational mesh is shown in Figure 2, [9].

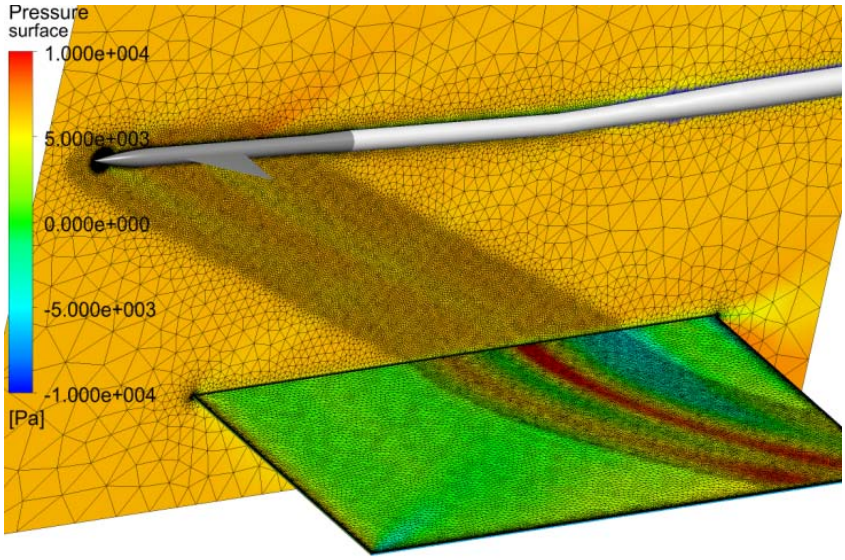


Figure 2: View of the spatial structure of the computational mesh in the symmetry plane and on plate surface for modeling the propagation of shock waves on the plate

The results of mathematical modeling were also used for comparison with the experimental data. Comparison with the experimental pressure distribution on the measuring plate mounted at a certain distance from the model in the WT test section, allows for evaluating the possibility of experiment and numerical simulation in predicting the formation of shock waves in the near zone of disturbed flow.

In the supersonic T-113 wind tunnel selected for testing a new technique there is no optical windows on the top and bottom walls of the test section and the PSP method has never been used here. Existing side-wall widows designed for Toepler device, do not pass ultraviolet radiation required for the excitation of the PSP. It was decided to use PSP, which is excited by soft ultraviolet radiation passing through the side windows of the WT. This sensor was invented and widely used in the USA [10] (in Russia is not patented).

For fluorescence excitation the LED illuminator was used, and the luminescence was registered by two CCD cameras "VIDEOSCAN-V2-285/P1-33" with multiple exposure mode. Two cameras are needed to measure the lifetime of the luminescence. At the time of measurement LED illuminator flashed for 100 μs with a period of 220 μs (4.54 kHz), while one camera has been accumulating the light intensity of the coating during the flash (first 50 μs), and the second after the flash for 110 μs.

The experimental studies were carried out both without models in the test section of WT and with two generating near-field models representing a schematized supersonic delta-wing plane and the axisymmetric cone-cylinder configuration. The tests were conducted at three Mach numbers: M = 2.0, 2.25 and 1.75. The scheme of the model and the measuring plate arrangement in the WT test section is shown in Figure 3.

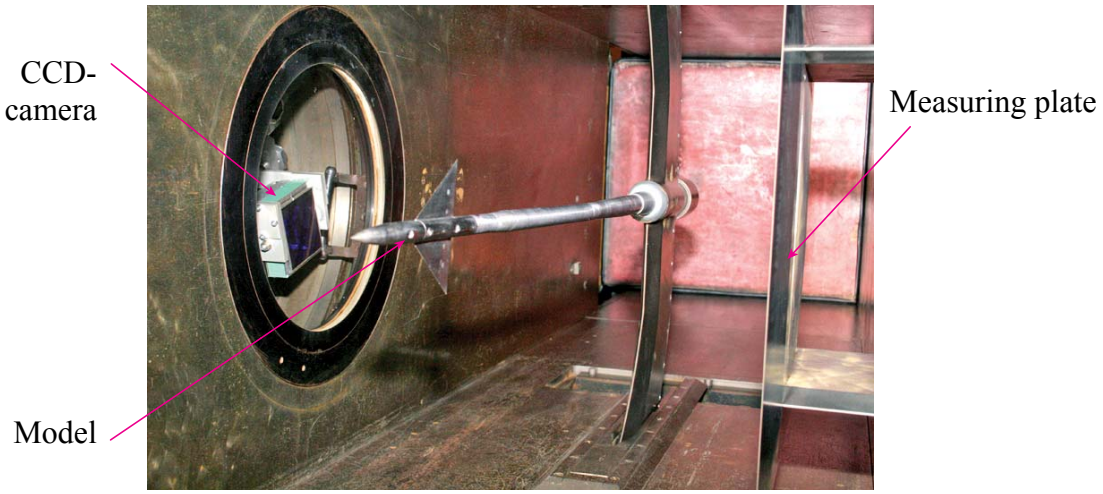


Figure 3: Experimental equipment in TsAGI T-113 WT.

The comparison of calculated and measured (using the PSP-method) pressure coefficient distributions on the measuring plate when the shock waves are generated by the delta-wing plane model at M = 2 are shown in Figure 4.

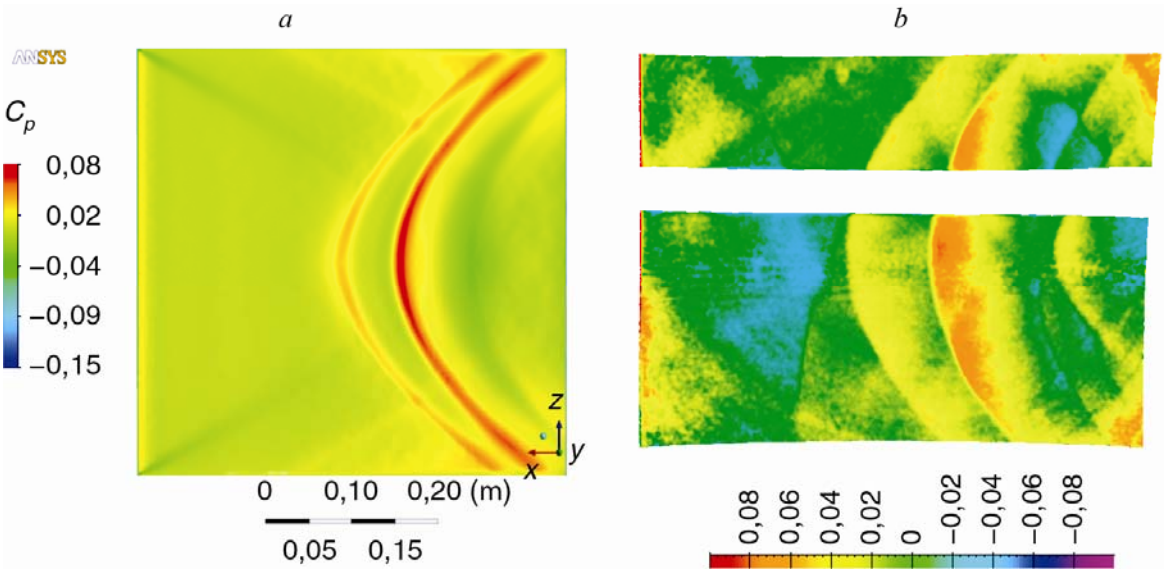


Figure 4: Pressure coefficient distributions over the measuring plate generated by the delta-wing plane model at M = 2: CFD prediction (*a*), PSP-measurements (*b*).

As can be seen from the above data, the position of the shock waves arriving at the measuring plate from the model agrees well with the CFD results.

Pressure perturbations registered by the control plate when testing models contain disturbances of an empty WT. To improve the results of pressure distribution measurements on the control plate, the background distributions obtained in empty WT are subtracted from the initial data. Then, on the basis of the Zhilin's integral [7], the derivative of the cross section area of equivalent body of revolution and pressure coefficient at the middle cross section of the measuring plate is calculated. The procedure of subtraction of "background" values of the pressure coefficient distribution registered by the measuring plate in empty WT, can significantly improve the final pressure distribution and the derivative of the cross section area of equivalent body of revolution. These results are used further to determine the intensity of the sonic boom wave on the ground.

Sonic boom can be calculated using as the derivative of the cross-section area of the equivalent body of revolution dS/dx and the pressure distribution in the initial sound wave, however, the calculation with the use of dS/dx should give a more accurate result. Red color in Figure 5 shows the distribution obtained from the numerical solution of the Euler equations.

Comparison of the results obtained by numerical methods with the experimental data of the first tests allows us to make conclusions about the prospects of the PSP-method in the studies of the near-field of aircraft models in the supersonic wind tunnel. At the same time, the analysis of the experimental results and their processing to obtain the source data in the form of derivative of area of the equivalent body of revolution showed the following:

1) In the WT tests without models, high level of the background disturbances is detected. Sometimes these disturbances are of the same order of magnitude as the values, induced by the shock generators (configurations "wing-body" and "cone-cylinder").

2) Flow disturbances, generated by tested configurations ("wing-body" and "cone-cylinder") may be effectively refined by means of subtraction of the corresponding background pressure coefficient field.

3) In further research on determination of initial data for solution of the sonic boom problem it is necessary to improve the accuracy of the pressure coefficient measurements.

4) For further studies some geometrical parameters should be changed, for example, it is necessary to place the measuring plate closer to the model. This should lead to higher levels of induced perturbations, reduce the impact of side effects on the edges of the plate and to ensure more uniform flow on it.

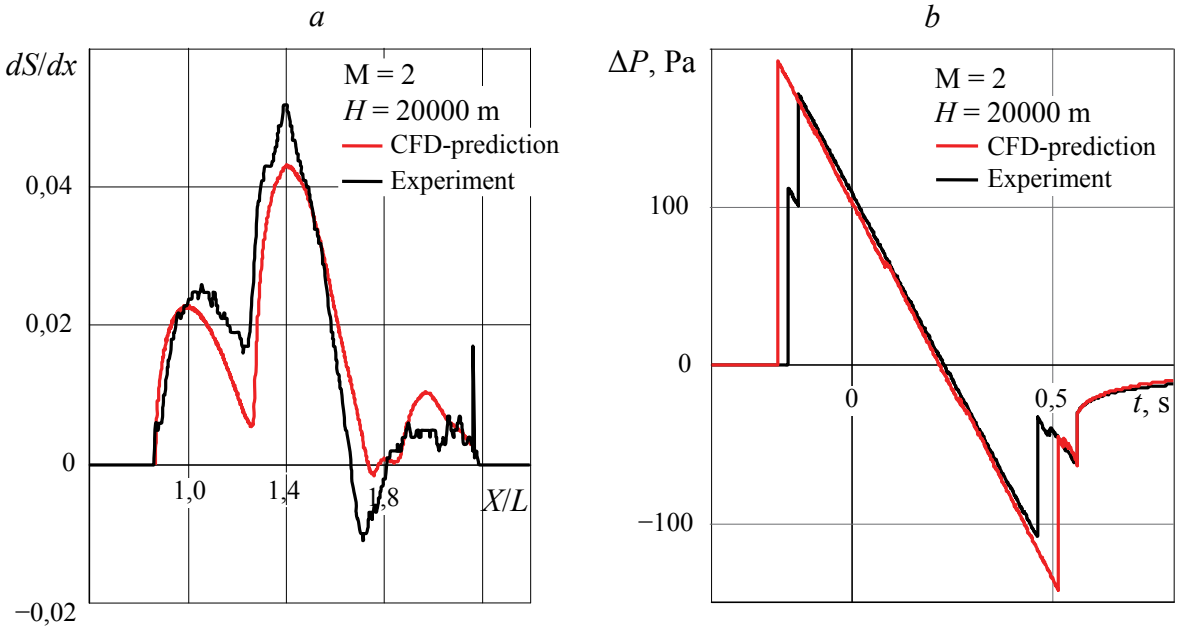


Figure 5: Distribution of the derivative of the cross-section area of equivalent body of revolution used for the sonic boom prediction (*a*), and the shock wave signature (*b*).

4 CONCLUSIONS

- On the basis of numerical simulation results, the test bench for sonic boom studies is designed and assembled in TsAGI T-113 WT. New test bench should allow for realization of experimental and computational method of the sonic boom prediction.
- For measurements of the static pressure distributions in the model near-field (on the measuring plate surface) the pressure-sensitive paints (PSP) method is used. Application of the PSP method should in principle ensure the acquisition of continuous pressure distribution on the control surface in the near-field of the model, calculation of Zhilin's integral on its basis and definition of the sonic boom wave profile at large distances using the quasi-linear theory.
- In the WT tests without models, high level of the background disturbances is detected on the measuring plate surface. Sometimes these disturbances are of the same order of magnitude as the values, induced by the shock generators (configurations "wing-body" and "cone-cylinder").
- Flow disturbances, generated by tested configurations ("wing-body" and "cone-cylinder") may be effectively refined by means of subtraction of the corresponding background pressure coefficient field.
- In further research on determination of initial data for solution of the sonic boom problem it is necessary to improve the accuracy of the pressure coefficient measurements. Recommendations are drawn for changes in some geometric parameters of the test bench.

REFERENCES

[1] S.L. Chernyshev, *The sonic boom*. Moscow: Science, 2011 (*in Russian*).

[2] S.L. Chernyshev, L.G. Ivanteeva, A.Ph. Kiselev, V.V. Kovalenko, P.P. Vorotnikov, A.V. Rodnov, L.L. Teperin, Yu.A. Zavershnev, Sonic boom research in Russia: computational methods improvement, aerodynamic configuration effects, analysis of meteorological variability influence. A. Atchley, V. Sparrow, R. Keolian eds. *Innovations in nonlinear acoustics: Proceedings of 17th International Symposium on Nonlinear Acoustics*, Melville, New-York: American Institute of Physics Conference Proceedings, **838**, 2005.

[3] G.B. Whitham, The flow pattern of a supersonic projectile. *Communications on Pure and Applied Mathematics*, **5**, No. 3, 301-338, 1952.

[4] V.F. Chirkashenko, Yu.N. Yudintsev, Development of measuring technique of the sonic boom parameters in the supersonic wind tunnels. *Preprint No.6 - 83, ITAM SB AS USSR*, Novosibirsk, 1983 (*in Russian*).

[5] V.F. Chirkashenko, Yu.N. Yudintsev, System for the sonic boom parameters measurements in wind tunnels using a computer. *Preprint No.21 - 83, ITAM SB AS USSR*, Novosibirsk, 1983 (*in Russian*).

[6] A.I. Rudakov, Yu.N. Yudintsev. Sonic boom from the bodies with spatial configurations in high supersonic flows. *Uchenye zapiski TsAGI*, **10**, No. 3, 27—36, 1979 (*in Russian*).

[7] Yu. L. Zhilin, V.V. Kovalehko. On connection of the near-field and far field in the sonic boom problem. *Uchenye zapiski TsAGI*, **29**, No. 3-4, 111-122, 1998 (*in Russian*).

[8] F.R. Menter, P.F. Galpin, T. Esch, M. Kuntz, C. Berner, CFD Simulations of aerodynamic flows with a pressure-based method. *Proceedings of 24th International Congress of the Aeronautical Sciences (ICAS 2004*), Yokohama, Japan, 2004.

[9] S.L. Chernyshev, A.I. Ivanov, A.V. Kazakov, A.Ph. Kiselev, V.E. Mosharov, Investigation of flow disturbances in the near field of aircraft in the supersonic wind tunnel. *Proc. of Int. Conf. "Progress in Continuum Mechanics"*, Vladivostok, Russia , 28 Sept.-04 Oct. 2014.

[10] M.P. Gouterman, W.B. Carlson, Acrylic and fluoroacrylic polymers for oxygen pressure sensing and pressure-sensitive paints utilizing these polymers. *Patent US 5965642*. Priority 12.10.1999.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

GLOBAL OPTIMIZATION OF THE SHAPE OF AN AEROSPACE VEHICLE, VIA ITERATIVE OPTIMUM-OPTIMORUM STRATEGY

A. Nstase[1]

[1] Aerodynamics of Flight, RWTH, Aachen University,
Templergraben 55, 52062 Aachen Germany
e-mail: nastase@lafaero.rwth-aachen.de

Keywords: Aerodynamic Global Shape's Oprimization, Evolutionary Iterptive Optimum-Optimorum Strategy, Enlarged Variational Method with Free Boundaries, Supersonic Flow, Catamaran Space Vehicle, Hybrid, Three-Dimensional Full Navier-Stokes Start Solutions

Abstract. *The determination of the global optimized (GO) shape of a flying configuration (FC), leads to an enlarged variational problem with free boundaries. An own developed evolutionary iterative optimum-optimorum (OO) strategy was developed in order to solve this problem. The GO shape of the FC is chosen among a class of elitary FCs. A lower limit hypersurface of the drag functional as function of the similarity parameters of the planforms of elitary FCs of the class is used and the elitary FC, which coresponds to the minimum of this hypersurface is, in the same time, the GO shape of the class. The iterative OO strategy uses, in its first step of iteration, analytical start solutions for the determination of the inviscid GO shape of FC, as surrogate model. The OO strategy and own developed software were used for the determination of inviscid GO shapes of three surrogate models, namely of ADELA (a delta wing alone) and of FADET I and FADET II, (two fully-integrated wing-fuselage configurations). These surrogate GO models were optimized with respect of inviscid minimum drag, respectively, at cruising Mach numbers* $M_\infty = 2,\ 2.2,\ 3.0$ *. Further, the friction drag coefficients of the surrogate models are computed by using own developed hybrid solutions for the Navier-Stokes PDEs and are checked for the structure point of view. The iterative OO strategy uses, up its second step of iteration, the total drag as new functional and the analytical start solutions are replaced with Navier-Stokes solutions. A GO shape of a space vehicle model, in form of CATAMARAN II, which is optimized with respect of minimum drag, at Mach number* $M_\infty = 3.0$ *, is proposed. It can be usefull for the design of performant sub-orbital space vehicles and for UAVs, It flies shock free, without sonic boom interference, has a reinforced structure and presents a high value of* L/D.

1 INTRODUCTION

The exploration of the space has taken a tremenduous development. The new scientifical research and the high technical performances make some dream of humans being a reality. At the same time more business, scientists and tourist travelers are interested to have faster intercontinental aircraft and to observe the earth during aerospace voyages. New cosmodromes are constructed and suborbital flights are planned by Virgin company, in order to respond to this increasing travel interest and to realize cheaper travels by increasing the number of passengers and by building more economic and ecologic acceptable supersonic transport aircraft (STA) and low earth orbit (LEO) space vehicles. The main aim of the design of GO shapes of supersonic FCs is to use them to increase the aerodynamic performances of aircraft and space vehicles.

The classic optimization of the shape of a FC consists in the determination of its surface of FC wuth fixed planform in order to reach a minimum drag at cruise. Such classic optimized shapes of FCs are here called **elitary FCs**. The shape of FC is GO if its camber, twist and thickness distributions and **also** the similarity parameters of its planform are **simultaneously** optimized in order to reach a minimum drag, at a chosen cruising Mach number. The determination of the GO shape of FC leads to an enlarged variational problem with free boundaries, which needs a special mathematical treatment. The own developed optimum-optimorum (OO) and iterative optimum-optimorum (IOO) theories are special strategies for the determination of the GO shape of the FC, inside of a class of elitary FCs, which satisfy some common properties, as in [1-4].

2 DETERMINATION OF THE INVISCID GLOBAL OPTIMIZED SHAPE OF A SURROGATE MODEL

Let us firstly consider an integrated wing-fuselage FC with arbitrary camber, twist and thickness distributions, which is flying at the cruising Mach number M_∞. Dimensionless coordinates are used for the computation of the distributions of velocity's components:

$$\tilde{x}_1 = \frac{x_1}{h_1} \quad , \quad \tilde{x}_2 = \frac{x_2}{\ell_1} \quad , \quad \tilde{x}_3 = \frac{x_3}{h_1} \quad . \qquad (\tilde{y} = \frac{\tilde{x}_2}{\tilde{x}_1}) \tag{1}$$

Further, integrated wing-fuselage FCs are considered, namely, for which the mean surface is continuous and the thickness distributions on the wing and on the fuselage are different and along the junction lines wing-fuselage they have the same tangent planes. The downwashes on the thin and thick-symmetrical components of the integrated FCs, w and w^*, w'^* (on the wing and on the fuselage of FCs) are supposed to be expressed in form of superposition of homogeneous polynoms with arbitrary coefficients, namely:

$$w \equiv \tilde{w} = \sum_{m=1}^{N} \tilde{x}_1^{m-1} \sum_{k=0}^{m-1} \tilde{w}_{m-k-1,k} \left|\tilde{y}\right|^k \quad ,$$

$$w^* \equiv \tilde{w}^* = \sum_{m=1}^{N} \tilde{x}_1^{m-1} \sum_{k=0}^{m-1} \tilde{w}^*_{m-k-1,k} \left|\tilde{y}\right|^k \quad , \qquad w'^* \equiv \overline{w}^* = \sum_{m=1}^{N} \tilde{x}_1^{m-1} \sum_{k=0}^{m-1} \overline{w}^*_{m-k-1,k} \left|\tilde{y}\right|^k \quad . \tag{2a-c}$$

The coefficients of the downwashes and the similarity parameter $\nu = B\,\ell$ ($\ell = \ell_1 / h_1$, $B = \sqrt{M_\infty^2 - 1}$), of the planform of the wing are the free parameters of the optimization and ℓ, ℓ_1, h_1 are the dimensionless span, the half-span and the depth of the planform of the delta wing. The quotient of the similarity parameters of the wing and of the fuselage, which depend on the purpose of the FC, is supposed constant. If the principle of minimal singularities (which fulfill the jumps of the velocity's components) and the hydrodynamic analogy of Carafoli are used, the following expressions for the axial disturbances of the thin and of the thick-

symmetrical components of the integrated wing-fuselage FC with subsonic leading edges are obtained, as in [1]:

$$u \equiv \ell\ \tilde{u} = \ \ell\ \sum_{n=1}^{N}\ \tilde{x}_1^{n-1} \left\{ \sum_{q=0}^{E\left(\frac{n}{2}\right)} \frac{\tilde{A}_{n,2q}\tilde{y}^{2q}}{\sqrt{1-\tilde{y}^2}} \ + \ \sum_{q=1}^{E\left(\frac{n-1}{2}\right)} \tilde{C}_{n,2q}\tilde{y}^{2q}\ \cosh^{-1}\sqrt{\frac{1}{\tilde{y}^2}} \right\} \quad , \tag{3a}$$

$$u^* \equiv \ell\ \tilde{u}^* = \ell\ \sum_{n=1}^{N}\ \tilde{x}_1^{n-1} \left\{ \sum_{q=0}^{n-1}\ \tilde{H}_{nq}^*\tilde{y}^q \left(\cosh^{-1} M_1 + (-1)^q \cosh^{-1} M_2 \right) \right.$$

$$+ \ \sum_{q=0}^{n-1}\ \tilde{G}_{nq}^*\ \tilde{y}^q \left(\cosh^{-1} S_1 + (-1)^q \cosh^{-1} S_2 \right)$$

$$\left. + \ \sum_{q=1}^{E\left(\frac{n-1}{2}\right)} \tilde{C}_{n,2q}^*\ \tilde{y}^{2q} \cosh^{-1}\sqrt{\frac{1}{\nu^2\tilde{y}^2}} + \ \sum_{q=0}^{E\left(\frac{n-2}{2}\right)} \tilde{D}_{n,2q}^*\ \tilde{y}^{2q}\sqrt{1-\nu^2\tilde{y}^2} \quad \right\} \quad , \tag{3b}$$

$$\left(\ M_{1,2} = \sqrt{\frac{(1+\nu)(1\mp\nu\ \tilde{y})}{2\nu(1\mp\tilde{y})}} \quad , \quad S_{1,2} = \sqrt{\frac{\left(1+\bar{\nu}\right)\left(1\mp\nu\ \tilde{y}\right)}{2\ \left(\bar{\nu}\mp\nu\ \tilde{y}\right)}} \ \right)$$

The lift and the pitching moment coefficients of the integrated wing-fuselage FCs, computed with the hyperbolic potential theory, are the following:

$$C_\ell \equiv 8\ell \int_{\tilde{O}\tilde{A}_1\tilde{C}_1} \tilde{u}\ \tilde{x}_1 d\tilde{x}_1 d\tilde{y} \quad , \qquad C_m \equiv -8\ell \int_{\tilde{O}\tilde{A}_1\tilde{C}_1} \tilde{u}\ \tilde{x}_1^2 d\tilde{x}_1 d\tilde{y} \quad . \tag{4a,b}$$

The inviscid drag coefficients of the thin, thick-symmetrical and thick, lifting FCs with retracted flaps are quadratic forms with respect of the downwashes coefficients :

$$C_d \equiv \ell\tilde{C}_d = 8\ell \int_{\tilde{O}\tilde{A}_1\tilde{C}} \tilde{u}\ \tilde{w}\ \tilde{x}_1 d\tilde{x}_1 d\tilde{y} \ ,$$

$$C_d^* \equiv \ell\tilde{C}_d^* = 8\ell\ \left[\int_{\tilde{O}\tilde{C}_1\tilde{C}} \tilde{u}^*\overline{w}^*\tilde{x}_1 d\tilde{x}_1 d\tilde{y} + \int_{\tilde{O}\tilde{A}_1\tilde{C}_1} \tilde{u}^*\tilde{w}^*\tilde{x}_1 d\tilde{x}_1 d\tilde{y} \ \right],$$

$$C_d^{(i)} \equiv \ell\tilde{C}_d^{(i)} = \ \ell\ \left(\ \tilde{C}_d + \tilde{C}_d^* \right). \tag{5a-c}$$

The hyperbolic integrated solutions for the axial disturbance velocity given in (3a,b) are used as start solutions for the determination of the inviscid GO shape of the wing-fuselage FC, which is of minimum drag at cruising Mach number. The free parameters of the optimization are the coefficients of the downwashes w , w^* and w'^* and also the similarity parameters of the planforms of the wing and of the fuselage. Further it is supposed that the quotient of these similarity parameters, which is determined for the purpose of the FC, is considered constant.

The constraints of the inviscid GO shape's design are: the given lift, pitching moment and the Kutta condition along the subsonic leading edges of the thin FC component (in order to cancel the induced drag at cruise and to suppress the transversal contournement of the flow around the leading edges, in order to increase the lift) and the given relative volumes of the wing and of the fuselage zone, the cancellation of thickness along the leading edges and the new introduced integration conditions along the junction lines between the wing and fuselage zone of the thick-symmetrical FC component (in order to avoid the detachment of the flow along these lines). According to the optimum-optimorum strategy, the GO shape of the FC is searched among the elitary FCs with the same area of their planforms which belong to the same class of FCs. The class is defined by the common properties of the elitary FCs, which belong to this class. The similarity parameter ν of the planform of FC is sequentially varied and a lower limit-line of the inviscid drag functional of elitary FCs, as function of this similarity parameter ν, is obtained. For FCs with subsonic leading edges, is: $0 < \nu < 1$. The position of the minimum of this limit-line gives the optimal value of the similarity parameter $\nu = \nu_{opt}$ and the corresponding elitary FC is, at the same time, the GO FC of the class.

The author has used the OO strategy for the determination of the GO shapes of three models, namely, ADELA (a wing alone) and FADET I and FADET II (two fully-integrated wing-fuselage FCs), which are of minimum drag at , respectively, cruising Mach numbers $M_\infty = 2;\ 2.2;\ 3$.

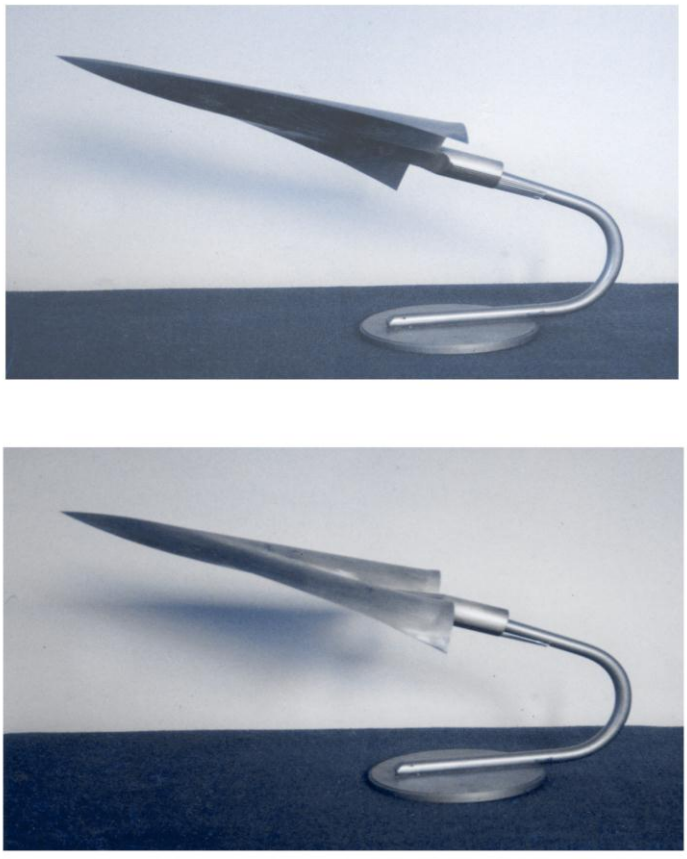


Fig. 1a,b The Global Optimized Model FADET II

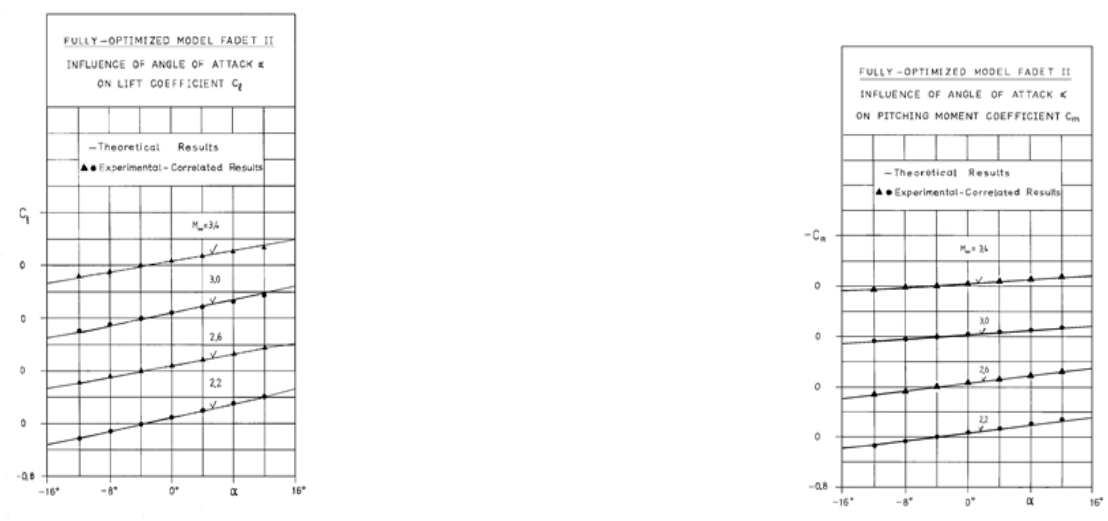


Fig. 2a,b The Agreement of the Theoretical Determined Lift and Pitching Moment Coefficients of the Global Optimized Model FADET II with the Experimental Results

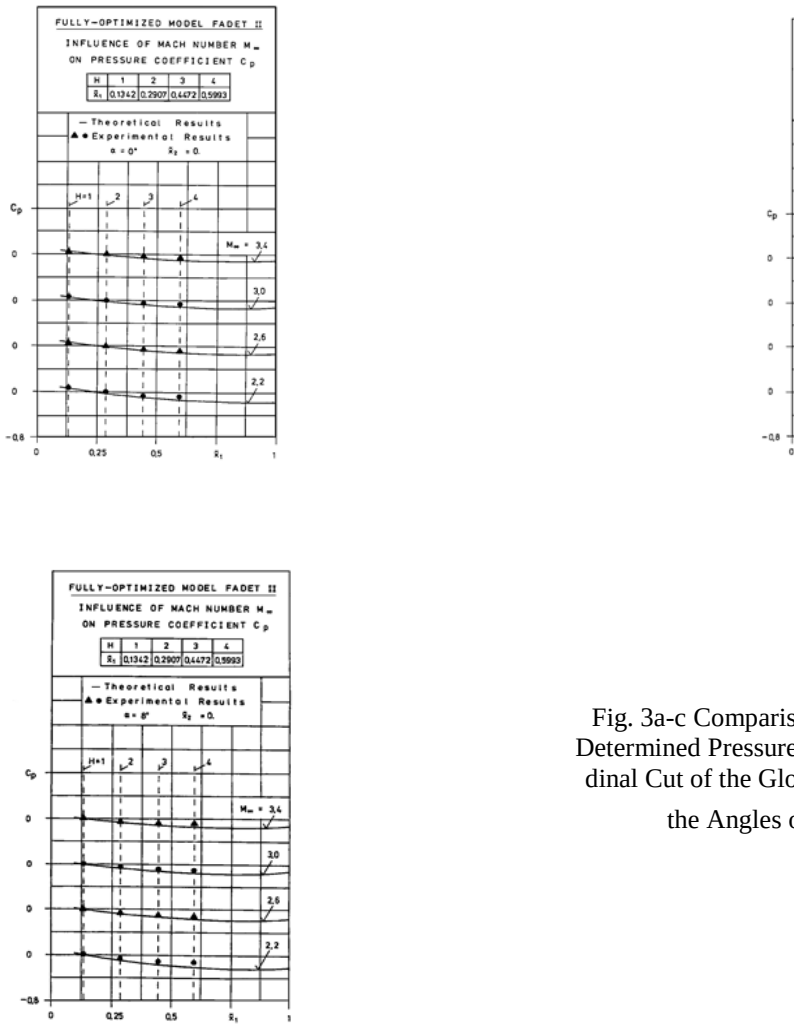


Fig. 3a-c Comparison of Theoretical and Experimental Determined Pressure Coefficients on the Central Longitudinal Cut of the Global Optimized Model FADET II , at the Angles of Attack $\alpha = -8°,\ 0°,\ 8°$

In the (Figs. 1a,b) are presented two views of the GO shape of the fully-integrated Model FADET II. It was measured in the trisonic wind tunnel of DLR-Cologne, in the frame of research projects of the author, sponsored by the DFG. The comparisons of theoretical and experimental-correlated values of the lift and pitching moment coefficients of all these models were in very good agreements with the experimental results. In the (Fig. 2a,b) are presented these agreements for the lift and pitching moment coefficients of the model FADETt II with subsonic leading edges at moderate angles of attack, as exemplification.

In the (Fig3a-c) are presented the agreements between the theoretical predicted and the measured pressure coefficients along the central longitudinal cut of the model FADET II at moderate angles of attack $\alpha = -8°,\ 0°,\ 8°$.

3 HYBRID SOLUTIONS FOR THE THREE-DIMENSIONAL COMPRESSIBLE NAVIER–STOKES LAYER

The new developed, hybrid, meshless solutions for the boundary value problems of the PDEs of the Navier-Stokes layer, proposed here, use the hyperbolic potential solutions of the flow on the same FC twice, namely: at the NSL's edge (instead of parallel flow used by Prandtl in his boundary layer theory) and in the structure of the velocity's components, which are expressed inside the NSL, as products between the corresponding potential velocity's components with polynoms with arbitrary coefficients, versus a spectral variable. These coefficients are used to satisfy the NSL's PDEs, in an arbitrary chosen number of points. Let us firstly introduce a spectral variable:

$$\eta = \frac{x_3 - Z\left(x_1,\ x_2\right)}{\delta\left(x_1,\ x_2\right)} \qquad (0 \le \eta \le 1) \tag{6}$$

The proposed forms for the hybrid numerical solutions of the velocity's components are , as in [1-4], the following:

$$u_\delta = u_e \sum_{i=1}^{N} u_i\ \eta^i \quad , \qquad v_\delta = v_e \sum_{i=1}^{N} v_i\ \eta^i \quad , \qquad w_\delta = w_e \sum_{i=1}^{N} w_i\ \eta^i \ . \tag{7a-c}$$

The here introduced logarithmic density function $R = \ln \rho$ and the absolute temperature T are the following:

$$R = R_w + \left(\ R_e - R_w\right) \sum_{i=1}^{N} r_i\ \eta^i \quad , \qquad T = T_w + \left(\ T_e - T_w\right) \sum_{i=1}^{N} t_i\ \eta^i \quad . \tag{8a,b}$$

The pressure p is computed by using the physical equation of perfect gas and, for the viscosity μ , an exponential law is used

$$p = R_g\ \rho\ T = R_g\ e^R\ T \quad , \qquad \mu = \mu_\infty \left(\frac{T}{T_\infty}\right)^{n_1} \tag{9a,b}$$

The free coefficients $u_i,\ v_i,\ w_i,\ r_i$ and t_i are used to satisfy the NSL's PDEs in some chosen points. If the hybrid forms for the velocity's components (7a-c) are introduced in the continuity's PDE and the collocation method is used, the coefficients r_i are determined only as func-

tions of the coefficients of the velocity's components, by solving a linear algebraic system and the coefficients t_i satisfy the PDE of absolute temperature and are also obtained only as functions of the coefficients of the velocity's components by solving of a transcendental algebraic system. A splitting of the NSL's PDEs is obtained and the physical entities are expressed only as function of the spectral coefficients of the velocities components and can be easy updated in an iterative process. A speed up of computation time is obtained. The coefficients of velocity's components are determined by using the impulse PDEs, which are iteratively solved, as in [1-4].

The hybrid solutions for the NSL presented here are reinforced numerical solutions, which present important analytical properties, namely: they have correct last behaviors, they have correct jumps due to the singularities located only along the singular lines (like the junction lines wing-fuselage and the subsonic leading edges of the wing of the FC) obtained according to the principle of minimal singularities which fulfil the jumps and the singularities are balanced, they are accurate because the partial derivatives of velocity's components can be exactly computed, they are split due to the use of the logarithmic density function and therefore they produce a speed up of the computation time, they fulfil automatically the non-slip condition on the FCs surface, they are matched with the outer potential flow and for moderate perturbations, they are reduced to the potential solutions at the NSL's edge and they do not need interface. Additionally, for hyperbolic PDEs the boundary condition on its characteristic surface is automatically fulfilled.

The hybrid solutions of the NSL's PDEs are useful for the computation of the friction drag coefficient of the FC. The skin friction coefficient at the wall takes the form:

$$\tau_{x_1}^{(w)} \equiv \tau_{x_1}\Big|_{\eta=0} = \mu_f \frac{\partial u_\delta}{\partial \eta}\Big|_{\eta=0} = \mu_f u_1 u_e \quad . \tag{10}$$

The friction drag coefficient and the total drag of the FC, with arbitrary camber, twist and thickness distributions are:

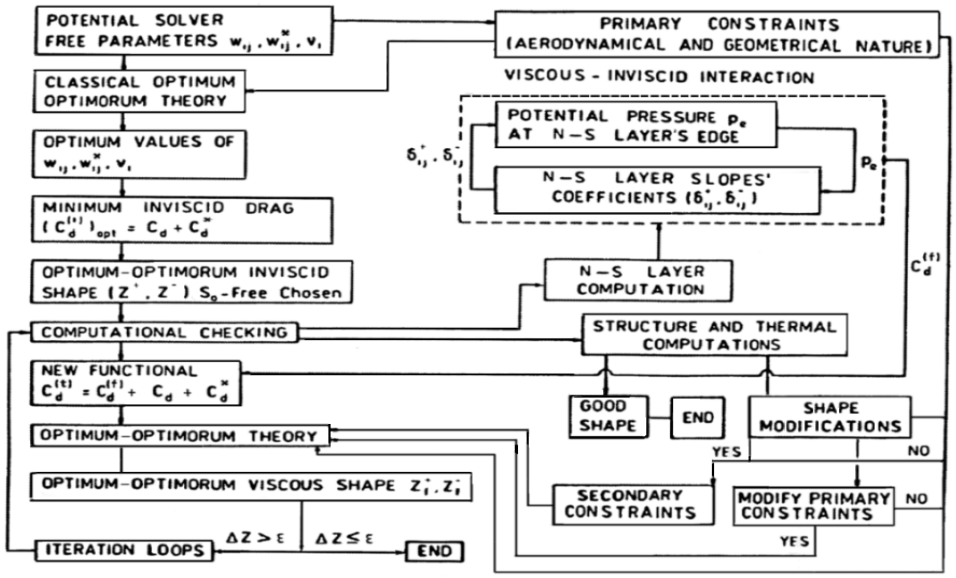
$$C_d^{(f)} = 8\ \nu_f u_1 \int\limits_{\tilde{O}\tilde{A}_1\tilde{C}} u_e x_1 dx_1 dy \ , \qquad C_d^{(t)} = C_d^{(f)} + C_d^{(i)} \quad . \tag{11a,b}$$

These hybrid NSL's solutions are also used for the viscous design of the GO shape of FC.

3 THE ITERATIVE OPTIMUM-OPTIMORUM STRATEGY

The viscous iterative OO theory of the author is proposed for the viscous determination of the GO shape of the FC in order to present a total minimum drag at cruising Mach number. The viscous iterative OO strategy uses the inviscid hyperbolic potential solutions as start solutions and the inviscid GO shape of these FCs as surrogate models, **only** in its first step of iteration. An intermediate computational checking of this inviscid GO shape of the FC is made with own hybrid solvers, for the three-dimensional compressible NSL. The friction drag coefficient $C_d^{(f)}$ of the FC is computed and the inviscid GO shape is checked also for the structure point of view. A weak interaction aerodynamics-structure is proposed. Additional or modified constraints, introduced in order to control the camber, twist and thickness distributions of the GO shape, for structure reasons, are here proposed. In the second step of optimization, the predicted inviscid GO shape of the FC is corrected by including these additional constraints in

the variational problem and of the friction drag coefficient in the drag functional. The chart flow of the iterative OO strategy is given in the (Fig. 4).



4. The Iterative Optimum-Optimorum Strategy

4 THE PROPOSED AEROSPACE VEHICLE MODEL FOR LEO

The main aim of the design of new GO shapes of supersonic FCs is to use them to increase the aerodynamic performances of supersonic transport aircraft (STA) and of low earth orbit (LEO) space vehicles and to build new supersonic FCs , which are more economical and ecological acceptable.

The proposed GO shape of Catamaran II with respect of minimum drag, at cruising Mach number $M_\infty = 3$ is presented in (Fig. 5). It has twin fuselages, partially embedded in the thickness of the GO FCs.

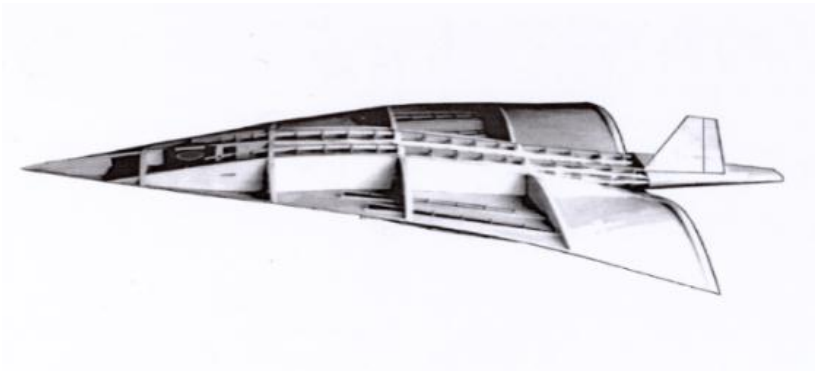


Fig. 5 GO Shape of Catamaran II

The following benefits are expected by using such GO shapes of FCs:

- Due to the global optimization with respect of minimum drag the GO shape are economical, chiper for passengers and produce less polution;

Due to the premises and of judicious choose of constraints it fly shock free, has no sonic boom interference and has more stiffness.

5 CONCLUSIONS

The Catamaran II can be used also as sources of inspirations for the shapes of the future new generation of aerospace vehicles for sub-orbital flights and for UAVS.

The Catamaran II has some advantages, when it is compared with the FCs with one, central, non-integrated fuselage, carrying the same number of passengers, at the same cruising Mach number:

- it flies with a *shock-free surface*;
- it has *no sonic boom interference* because it flies with one *characteristic surface* (the classical FCs with one central non-integrated fuselage flies with two shock surfaces, one produced at the frontal part of the fuselage and the other at the roots of the wing and in their intersection zones, the sonic boom interference occurs);
- it has a *better structural stiffness and increased lateral stability*; because instead of one long fuselage there are two twin fuselages embedded in the wing, with half length;
- it needs *less trim* because the weight is better distributed and the pressure center and the center of gravity points are closer together;
- it has a *higher L/D*, due to global optimization, full integration, of flattened form and due to the fulfilling of the Kutta condition along its subsonic leading edges, which avoids the leading edges contournements, which cancels the induced drag, destroys the leading edge vortices and increases the lift, not only at cruise, but also for large ranges of Mach numbers and angles of attack.

The proposed GO shapes of supersonic FCs look like birds in gliding flight!

- REFERENCES

[1] A. Nastase, *Computation of Supersonic Flow over Flying Configurations,* Elsevier, Oxford, UK, 2008

[2] A. Nastase, *Evolutionary, Iterative Optimum-Optimorum Theory,* INCAS Bulletin, Vol. 2, Number 4, Bucharest, Romania pp. 153-161 (2010)

[3] A. Nastase, *Multipoint and Iterative Aerodynamical Global Shape's Optimization*, Proceedings of Evolutionary and Deterministic Methods for Design, Optimization and Control of Eurogen 2011, Paper 30 , CIRA, Italy (2011).

[4] A. Nastase, *Hybrid Navier-Stokes Solutions for Aerodynamical, Global Optimal Shape's Design,* Proceedings of International Conference EngOpt, Paper 750, Rio de Janeiro, Brazil, (2008).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SHAPE IDENTIFICATION ANALYSIS OF CAVITY IN RESIN STRUCTURE BASED ON THERMAL NONDESTRUCTIVE TESTING METHOD

K. Maruoka[1], **T. Kurahashi**[2], **and T. Iyama**[3]

[1] Graduate School of Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata, Japan
e-mail: s133075@stn.nagaokaut.ac.jp

[2] Nagaoka University of technology
1603-1 Kamitomioka, Nagaoka, Niigata, Japan
e-mail: Kurahashi@mech.nagaokaut.ac.jp

[3] Nagaoka National College of Technology
888 Nishikatakai, Nagaoka, Niigata, Japan
e-mail: iyama@nagaok-ct.ac.jp

Keywords: Shape identification analysis, Thermal nondestructive testing method, Finite element method, Adjoint variable method, Gaussian filter.

Abstract. *The shape identification analysis is carried out to obtain the unknown defects shape in the structure based on the finite element and the adjoint variable methods. In this study, the test piece including the known defect shape is employed to solve the shape identification problem. In addition, we present the shape identification problem of cavity in resin test piece made by 3D printer using the observed temperature on the test piece surface. It is known that the temperature on top of the test piece is not uniformly distributed, if there are cavities in the test piece. Furthermore, according to practical experiment, it has been confirmed that the characteristic of the temperature distribution depends on cavities size. The thermal physical constants, i.e., the thermal conductivity and the convection coefficient, are identified for the model of the test piece including a cavity based on the experimental data, and the shape identification analysis is carried out. In the numerical analysis, the finite element method is applied to simulate the temperature distribution in the test piece, and the adjoint variable method is employed to identify the cavity shape.*

1 INTRODUCTION

Thermal testing method is known as a kind of non-destructive testing for estimating corrosion in structures in fields of civil and mechanical engineering. This testing method is intended to find out existence of defects by thermal image of thermography. In addition, it has the advantage that can give defect inspection through the positon of surfaces. An example of a thermal image of the top surface of test piece having a cavity which is heated from the lower surface is shown in Figure 1. Then outer and inner test piece have different temperatures, being able to estimate that the test piece has a defect. However there are some cases too difficult to find out defects, depending on the size and depth of the defects. On the other hand, by applying a thermal testing method, studies to identify the defect shape by inverse analysis method from time history of temperature [1, 2] is carried out. In previous studies, it is concluded that the corrosion shape of the concrete can be identified if assuming the initial corrosion shape as appropriate. In the fields of optimal control problems, the study about reduction of convergence for performance function [3, 4].

In this study, we use the resin structure created by 3D printer such that the shape of cavity is freely changed. The purpose of this study is to identify the cavity shape based on the inverse analysis, and to carry out the improvement of the treatment in computation of the shape identification.

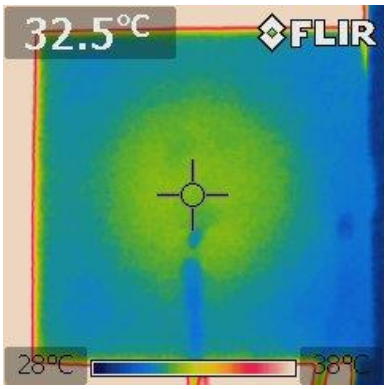


Figure 1: Heat image

2 EXPERIMENT BY THERMAL NONDESTRUCTIVE TESTING METHOD

2.1 Measurement situation

The photo of experiment is shown in Figure 2, and the drawing of test piece used this study is shown in Figure 3 (a). The sample has been made of ABS resin and created by 3D printer. Then a test piece with a cavity having a thickness of 10mm and a test piece without cavity are also created. Pair of test pieces (15mm thickness cavity, no cavity) are set on hot plate, heated up for 1200 seconds, and thermal observation at two observation points on the surface with a thermocouple is carried out every 10 seconds. Observation point is placed on the upper surface of the thick point (Point A) and thin portion (Point B) of the cavity thickness (See Figure 3(b)).

2.2 Observational result

The measured temperature history at Point A and Point B is shown in Figure 4 (a) and (b). As a result, without relation to presence or absence of the cavity, there are no big difference in temperature history observed at Point B. However, in Point A, it is confirmed a tendency that

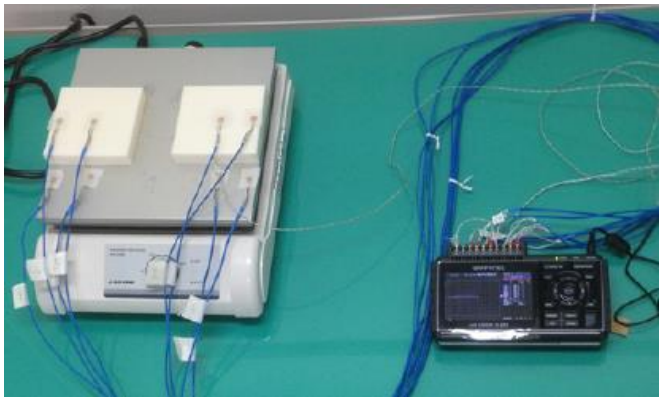
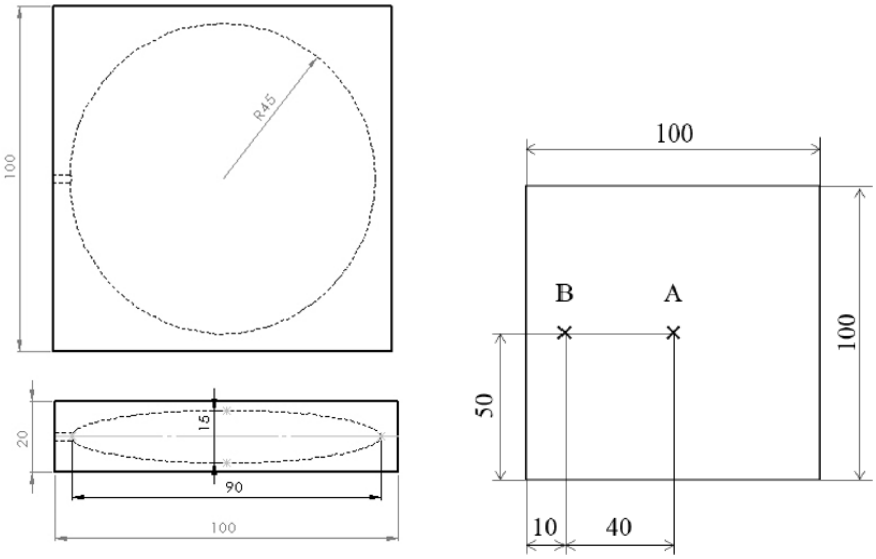


Figure 2: Photo of observation



(a) Drawing of the test piece (b) Observation point

Figure 3: Detail of the test piece

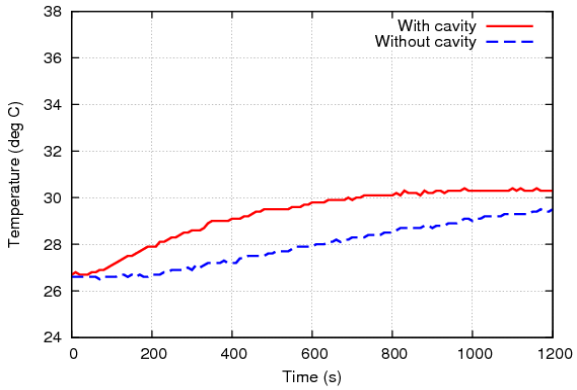
the temperature of the test piece with a cavity increases more than the test piece without a cavity at the same time. This is considered to effect due to the influence of heat transfer from the air which is heated in the cavity. Similar experiment are also carried out using a test piece with a cavity having a thickness of 10mm. The temperature history measured at Point A and Point B is shown in Figure 5 (a) and (b). From the results, it is found that the temperature difference at Point A is small compared the case with a test piece cavity thickness 15mm. Thus, the thickness of the cavity difference, it can be seen that the result is a difference in the temperature history at the test piece surface.

3 SHAPE IDENTIFICATION ANALYSIS

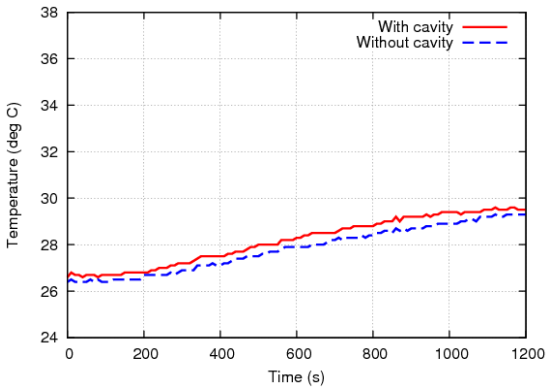
Using thermal properties found out previous chapter, the shape identification analysis is carried out. Initial shape model is given as including the cavity thickness 13mm, we set target shape as cavity thickness 15mm.

3.1 State equation

Formulation in the shape identification analysis is described below. In this study, the whole domain of the test piece is denoted as Ω. Then temperature distribution ϕ satisfies heat transfer

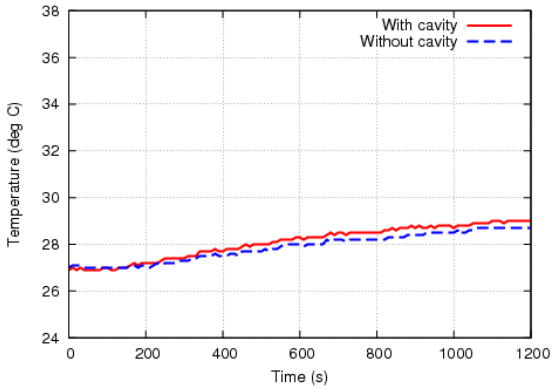


(a)At Point A

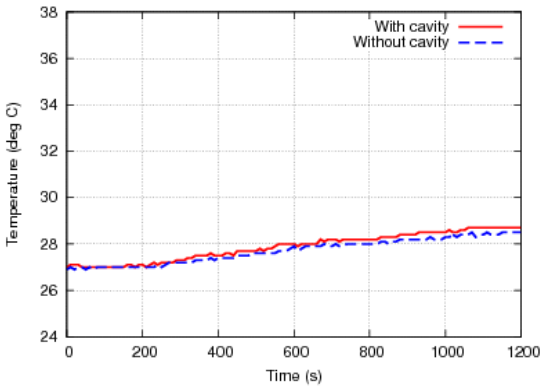


(b) At Point B

Figure 4: Time history of temperature of resin surface (1)



(a)At Point A



(b) At Point B

Figure 5: Time history of temperature of resin surface (2)

equation shown in Equation (1). For the heat transfer equation, initial condition and boundary condition are defined as Equations (2).

$$\rho c\dot{\phi} - \kappa\phi_{,ii} = 0 \tag{1}$$

$$\begin{cases} \phi = \hat{\phi}(t_0) \quad in \quad \Omega \\ \phi = \hat{\phi} \quad on \quad \Gamma_1 \\ -\kappa\phi_{,i}n_i = h_1(\phi - \phi_{\inf}) \quad on \quad \Gamma_2 \\ -\kappa\phi_{,i}n_i = h_2(\phi - \phi_{cav}) \quad on \quad \Gamma_3 \end{cases} \tag{2}$$

where ρ, c, κ, n_i, h_1, h_2, $\phi_{\inf}$ and ϕ_{cav} denote density, specific heat, thermal conductivity, unit normal vector, heat transfer coefficient of inside cavity, heat transfer coefficient of outside surface, surrounding temperature, temperature inside cavity. Γ_1 means lower surface, Γ_2 means outside surfaces, and Γ_3 means surface of cavity (See Figure 6).

3.2 Performance function

To evaluate computed temperature history, following performance function is defined.

$$J = \frac{1}{2}\int_{t_0}^{t_f}\int_{\Omega} R(\phi - \phi_{obs})^2 \, d\Omega dt \tag{3}$$

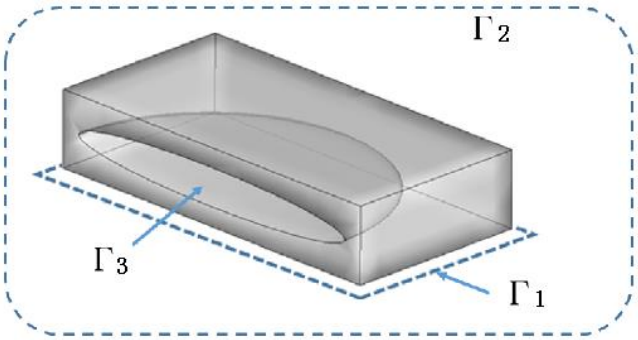


Figure 6: Boundary definition of the test piece

where t_f, R, and ϕ_{obs} mean heating termination time (Total time), weight diagonal matrix, observed temperature.

3.3 Lagrangian function

Applying adjoint variable λ, following Lagrangian function is obtained to minimize performance function [6, 7]. The adjoint variable method is one of the minimization technique of the performance function, and is suitable for the inverse problems such that a lot of unknown parameters should be solved.

$$J^* = J + \int_{t_0}^{t_f} \int_{\Omega} \lambda(\rho c\dot{\phi} - \kappa\phi_{,ii}) d\Omega dt \tag{4}$$

First variation which is necessary condition that Lagrangian function become minimization is shown in equation (5). Stationary condition is given as condition that each terms equal zero.

$$\delta J^* = \frac{\partial J^*}{\partial \lambda}\delta\lambda + \frac{\partial J^*}{\partial \dot{\phi}}\delta\dot{\phi} + \frac{\partial J^*}{\partial \phi}\delta\phi + \frac{\partial J^*}{\partial x_i}\delta x_i \tag{5}$$

where x_i indicates transplantable nodes.

3.4 Finite element equation

Shape function for four-node tetrahedral element is introduced into state equation to discretize the state equation spatially with Galerkin method. In addition, the equation is discretized in time based the Crank-Nicolson method. The finite element equation is shown in equation (6).

$$\rho c[\mathrm{M_e}]\{\dot{\phi}_e\} + \kappa[H_e]\{\phi_e\} = \{T_e\} \quad in \quad \Omega_e \tag{6}$$

In equation (6), [$\mathrm{M_e}$], [$\mathrm{H_e}$], and {$\mathrm{T_e}$} indicate:

$$[M_e] = \int_{\Omega_e} \{\Phi_e\}\{\Phi_e\}^T d\Omega \tag{7}$$

$$[H_e] = \int_{\Omega_e} \{\Phi_{e,i}\}\{\Phi_{e,i}\}^T d\Omega \tag{8}$$

$$\{T_e\} = -\int_{\Gamma_2} q\{\Phi_e\} ds - \int_{\Gamma_3} q\{\Phi_e\} ds \tag{9}$$

where q means thermal flow late. Superposing finite element equation for individual domain, such equation can be represented as equation (10).

$$[A(x_i)]\{\dot{\phi}\}+[B(x_i)]\{\phi\}=\{C(x_i)\} \tag{10}$$

3.5 Adjoint equation

From the first variation of the Lagrangian function, following adjoint equation and conditions are obtained. This equation manifests oneself as adjoint problem to calculate λ.

$$-[\mathrm{A}]^{\mathrm{T}}\{\dot{\lambda}\}+[B]^{T}\{\lambda\}+[R]^{T}\{\phi-\phi_{obs}\}=\{S\} \quad in \quad \Omega_e \tag{11}$$

$$\begin{cases} \lambda(t_f)=0 & in \quad \Omega \\ \lambda=0 & on \quad \Gamma_1 \\ S=0 & on \quad \Gamma_2 \\ S=0 & on \quad \Gamma_3 \end{cases} \tag{12}$$

3.6 Gradient vector of the Lagrangian function

Final term in equation (5) can be calculated with adjoint variable as following equation:

$$\left\{\frac{\partial \mathrm{J}^{*}}{\partial x_i}\right\}=\int_{t_0}^{t_f}\{\lambda\}^{T}\left(\left[\frac{\partial \mathrm{A}}{\partial \mathrm{x_i}}\right]\{\dot{\phi}\}+\left[\frac{\partial \mathrm{B}}{\partial \mathrm{x_i}}\right]\{\phi\}-\left\{\frac{\partial C}{\partial \mathrm{x_i}}\right\}\right)dt \quad in \quad \Omega_e \tag{13}$$

Nodal positions are updated with gradient vector. In this method, the update of the coordinate on cavity surface and the re-evaluation of the temperature difference at observation points are iteratively carried out according to steepest descent method [7].

3.7 Computational condition

The finite element model of the test piece is shown in Figure 7. The model is composed four-node tetrahedral element. Total number of nodes are 1460, and total number of elements are 5552. The computational conditions are referred to experiment and given as Table 1. Thermal properties for the test piece is given as shown in Table 2 considering the thing which the ABS resin is not filled enough as shown in Figure 8. The observational data at Point A in the previous experiment (See Figure 4 (a)) is employed as observed temperature. In addition, the time history of temperature on the hot plate and temperature within the cavity as boundary conditions are shown in Figure 9 (a) and (b).

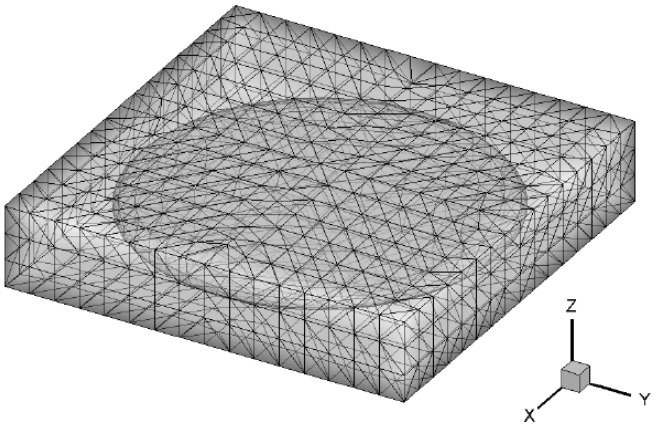


Figure 7: Finite element model

Total number of nodes	1460
Total number of elements	5552
Total time t_f, sec	600
Time increment Δt, sec	10
Time step	60
Convergence criterion ε	10^{-6}

Table 1: Computational condition.

	Density ρ	Specific heat c	Thermal conductivity κ	Heat transfer coefficient h	
	Kg/m^3	J/(kg℃)	W/(m^2℃)	Outside surface W/(m^2℃)	Cavity W/(m^2℃)
ABS	750	1386	0.1354	10.0	7.0

Table 2: Example of the construction of one table.

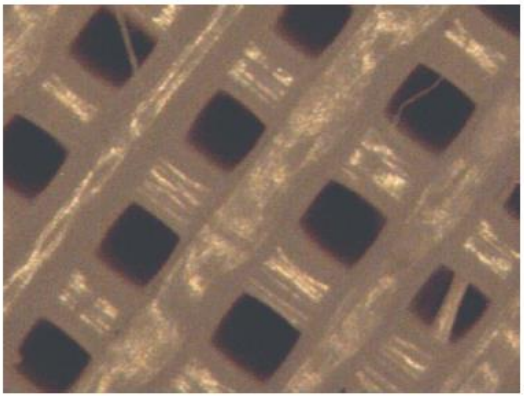
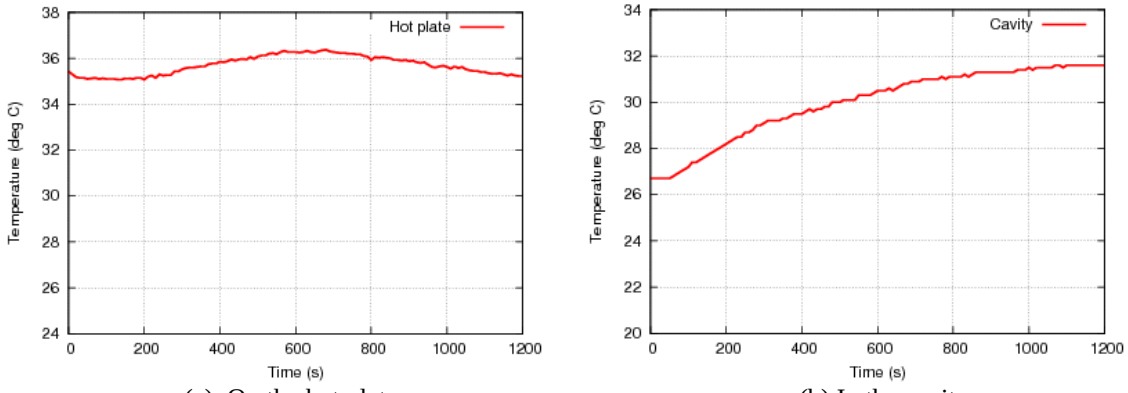


Figure 8: Cross-section of test piece

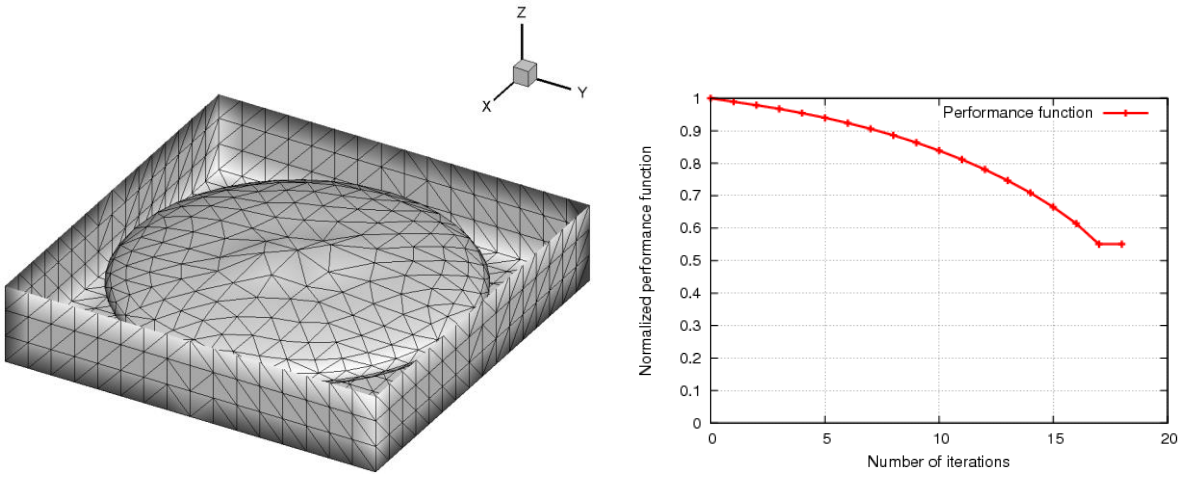


(a) On the hot plate (b) In the cavity

Figure 9: Time history of temperature of resin surface

3.8 Computational result

Figure 10 (a) shows identified shape of the cavity and Figure 10 (b) shows variation of normalized performance function per iteration number. Final iteration number is 18, performance function is about 0.55 when converting initial performance function into 1. However surface of the cavity becomes undulation, targeted shape as 15mm thickness cavity cannot be obtained.



(a) The shape of the cavity at the final iteration (b) Variation of performance function

Figure 10: Computational results (1)

4 SMOOTHING PROCESS

The smoothing process for gradient vector is introduced to modify the oscillation of the gradient vector with respect to coordinate on cavity surface. For the smoothing method, Gaussian filter [8] is employed. Looking at the surface of the cavity as two-dimensional surface, the gradient vector is smoothed.

4.1 Gaussian filter

Gaussian filter is a kind of the smoothing filter for the field of graphics. In smoothing method with Gaussian filter input of weighting parameter, value of the target nodes is determined by value of all nodes on the same plane. And the Gaussian filter has the advantage that the extent of the smoothing process is adjustable with the parameter σ. The weighting parameter in Gaussian filter [9] is defined as following equation:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{14}$$

where x and y is the distance of x-axial distance and y-axial distance between the target node and referenced node. The weighting parameter becomes large value if the distance is short. The parameter σ which is comparable to the variance in Gaussian distribution determines flattening of the distribution. In addition, as the summation of the weighting parameter for the target node becomes 1, we multiply the weighting parameter by the summation. The weighting parameter for node m to node n is shown in following equation:

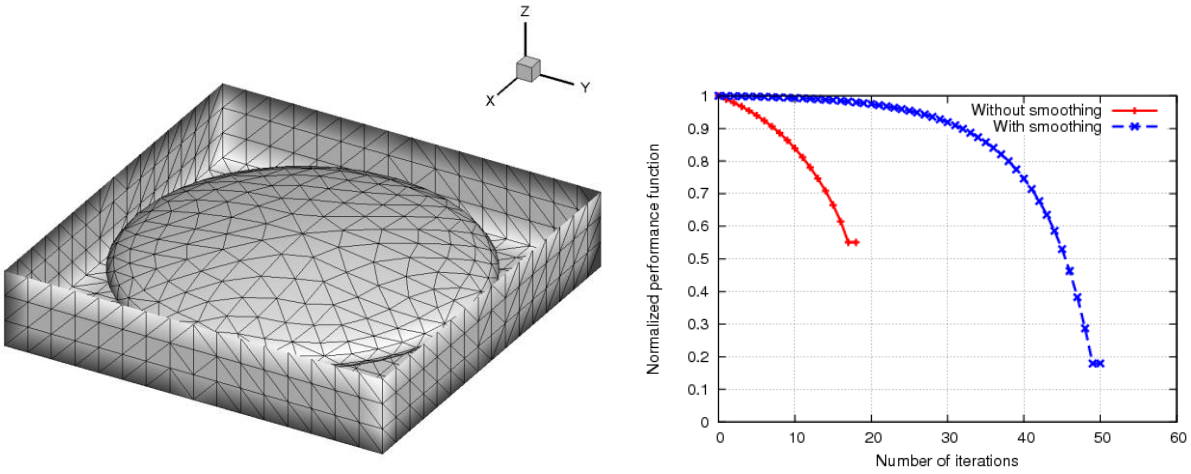
$$G^{(m,n)} = \frac{\frac{1}{2\pi\sigma^2} e^{-\frac{(x^{(m)}-x^{(n)})^2+(y^{(m)}-y^{(n)})^2}{2\sigma^2}}}{\sum_{n=1}^{cx} \frac{1}{2\pi\sigma^2} e^{-\frac{(x^{(m)}-x^{(n)})^2+(y^{(m)}-y^{(n)})^2}{2\sigma^2}}} \tag{15}$$

where cx means the number of nodes on the surface, $x^{(m)}$ and $y^{(m)}$ mean x-coordinate and y-coordinate of the node m. Then smoothed gradient vector is computed, the calculating formula is represented as following matrix form:

$$\begin{Bmatrix} \dfrac{\partial J^*}{\partial x^{(1)}}' \\ \dfrac{\partial J^*}{\partial x^{(2)}}' \\ \vdots \\ \dfrac{\partial J^*}{\partial x^{(cx)}}' \end{Bmatrix} = \begin{bmatrix} G^{(1,1)} & G^{(1,2)} & \cdots & G^{(1,cx)} \\ G^{(2,1)} & G^{(2,2)} & \cdots & G^{(2,cx)} \\ \vdots & \vdots & \ddots & \vdots \\ G^{(cx,1)} & G^{(cx,2)} & \cdots & G^{(cx,cx)} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial J^*}{\partial x^{(1)}} \\ \dfrac{\partial J^*}{\partial x^{(2)}} \\ \vdots \\ \dfrac{\partial J^*}{\partial x^{(cx)}} \end{Bmatrix} \tag{16}$$

4.2 Shape identification introducing Gaussian filter

In the Gaussian filter, the parameter σ is given as 20, and the shape identification analysis is carried out. The identified shape of the cavity is shown in Figure 11 (a). The surface of the cavity is smoother than the result without smoothing process in previous section. As the variation of performance function, in the case using Gaussian filter, number of iteration is larger and performance function is lower (See Figure 11(b)). It is confirmed that the temperature comes closer to observed value by applying smoothing process.



(a) The shape of the cavity at the final iteration

(b) Variation of performance function

Figure 11: Computational results (2)

5 CONCLUSIONS

In this study, the shape identification problem of the cavity in the structure based thermal testing method is carried out. The test piece including a cavity that the shape is variable made by 3D printer is heated from undersurface and the temperature is observed. The shape identification analysis, based the finite element method for four-node tetrahedral element and adjoint variable method, discretizing equations spatially and temporally with Galerkin method and Crank-Nicolson method, is carried out. In the identification process, Gaussian filter is employed to smooth the movement of nodes. The results in this study are shown below.

- When the test piece is heated from lower surface, the temperature of the test piece including a cavity is higher than the one without cavity at the same observation point and time, the temperature at face just above the cavity is especially high.
- The temperature history of the test piece on heating changes depending on the cavity size, the temperature becomes highly as the cavity is larger.

- Using Gaussian filter and smoothing gradient vector, the identified shape of the cavity is smoother than the result without smoothing process, and calculated temperature comes closer to observed temperature.

REFERENCES

[1] T. Kurahashi, H. Oshita, Evaluation of 3D Reinforcement Corrosion Shape in Concrete Based on Observed Temperature on Concrete Surface, *International Journal of Computer Science*, 4:4:340-351, 2010.

[2] T.Kurahashi, H.Oshita, Shape Determination of 3-D Reinforcement corrosion in Concrete Based on Observed Temperature on Concrete Surface, *Computers and Concrete*, 7:1:63-81, 2010.

[3] Z. Wang, I.M. Navon, X. Zou, F.X. Le Dimet, The second order adjoint analysis: theory and applications, *Meteo. and Atmo. Phys*. 50:3-20, 1992.

[4] Z. Wang, I.M. Navon, X. Zou, F.X. Le Dimet, A truncated Newton optimization algorithm in meteorology applications with analytic Hessian/vector products., *Computational Optimization and Applications*, 4:241-262, 1995.

[5] T. Kurahashi, M. Kawahara, Examinations for thermal condition of Lagrange multiplier for heat transfer control problems, *International Journal for Numerical Methods in Engineering*, 73:982-1009, 2008.

[6] T. Kurahashi, Examination for adjoint boundary conditions in initial water elevation estimation problems, *International Journal for Numerical Methods in Fluids*, 63:1270-1295, 2010.

[7] Y. Sakawa, Y. Shindo, On global convergence of an algorithm for optimal control, *IEEE Transactions on Automatic Control*, 25:1149-1153, 1980.

[8] S. Wang, W. Li, Y. Wang, Y. Jiang, S. Jiang, R. Zhao, An improved Difference of Gaussian Filter in Face Recognition, *Journal of Micromechanics and Microengineering*, 7.6.429-433, 2012.

[9] H. Ishiwata, H. Kubota, Y. Shima, Experimental study on Gaussian filter in restoration of error diffused character images, *Information Processing Society of Japan*, 18, 5, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DESIGN AND INSTALLATION OF A PERMANENT MONITORING SYSTEM FOR PALAZZO LOMBARDIA IN MILANO, ITALY

M. Berardengo[2], A.Cigada[1], S.Manzoni[1], M.Vanali[2]

[1] Dipartimento di Meccanica, Politecnico di Milano
Via La Masa,1 Milano Italy
e-mail: stefano.manzoni@polimi.it, alfredo.cigada@polimi.it

[2] Dipartimento di Ingegneria Industriale, Università degli studi di Parma
Parco Area delle Scienze, 181/A, 43124 – PARMA Italy
marcello.vanali@unipr.it, marta.berardengo@unipr.it

Keywords: Structural monitoring, wind excited vibrations, operational modal analysis.

Abstract. *This work describes the design and installation of a permanent system for the structural monitoring of Palazzo Lombardia Building in Milano. This is one of the high rise buildings (162 m tall) built in Milano in the last few years and is actually the seat for the regional government. A dynamic testing of the building was performed during the construction, highlighting the main vibration frequencies and the associated mode shapes. Now that the building is fully operational the public administration asked for the installation of a monitoring system in order to satisfy the following requirements: (i) monitoring the evolution of the dynamic parameters identified during the previous testing phase, (ii)monitoring the vibration to assess comfort and serviceability against the wind, (iii) monitoring the static behavior of Core 1 (tallest tower), (iv) having a continuous monitoring to capture exceptional events*

To this aim a continuous monitoring system was developed relying on high sensitivity piezo-accelerometers, low-noise high stability clinometers and a wind measurement station. The system has been installed and is fully operational since the end of October 2015. The system is described in details in this paper and the first results in terms of vibration and static behavior evolution are presented. Starting from these data a continuous modal parameter extraction will be developed and implemented.

1 INTRODUCTION

In the last few years a number of high rise buildings have been built in the center of the Italian city of Milano completely changing the city skyline. Among them the Palazzo Lombardia was the first to be built and is the current seat for the regional government.

Due to the building strategic relevance, a number of dynamic tests have been performed at the end of the construction phase in order to assess the structure dynamic behavior and update the FE model to seismic resistance purposes [1,2,3]. After these tests the first principal vibration modes and their associated mode shapes are known and set the base for the planned long term monitoring of the structure.

As other important buildings in town [4,5,6], the public administration asked for a continuous monitoring system in order to be able to detect and prevent possible problems in terms of comfort and structural failures, and to be able to assess the structural status after exceptional events such as earthquakes.

In order to fulfill the request, and after all the bureaucratic issues were accomplished, a permanent monitoring system has been designed, installed and is now running in the building. The main aims of the system are:

- Monitoring the evolution of the dynamic parameters identified during the previous testing phase
- Monitoring the vibration to assess comfort and serviceability against the wind
- Monitoring the static behavior of Core 1 (tallest tower)
- Having a continuous monitoring to capture exceptional events

The idea is to exploit the data that will be available to set-up a health monitoring procedure of the core 1 structure [7,8].

In this paper, after a brief description of the building structure is shown, the monitoring system layout is presented in terms of sensors and data acquisition/storage devices. In the last section the first results coming from the monitoring system are presented.

2 STRUCTURE DESCRIPTION

The “Palazzo Lombardia” building is the first in a series of high rise buildings which have been built in Milano in the last years. It is the current seat for the Regional Government and offices and therefore considered of strategic relevance. The complex is made up of five lower buildings (about 40m high, called Cores 2, 3, 4, 5 and 6), surrounding the high-rise Tower (Core 1), which scored, at the time of construction, the new height record in Italy. The complex sinuous interweaving strands recall the mountains, valleys, and rivers of the Lombardia region.

The curvilinear shapes of the buildings, while having a very strong aesthetical impact, are also strongly reflected in the irregular plan-wise configuration of the complex, which, at its center, defines an inner covered public ‘plaza’(Piazza Lombardia), having an area of about 4200 m^2, covered by a steel truss system supporting transparent Texlon ETFE (ethylene-co-tetrafluoroethylene) cushions. In Figure 1, the six buildings, or Cores (1-6), making up the complex are visible.

Figure 1 Palazzo Lombardia: the whole complex

The structural system is entirely made up of reinforced concrete load bearing elements, except for the Auditorium area in Core 4 and the 'Velarium' on top of Core 1, a three-storey high belvedere area to be used for official public purposes and rented for private events. For these two parts structural steel was employed. The monitoring system will be dedicated to the Core 1 structure (tallest tower) in the initial phase with the idea of possibly expanding it to the remaining sub-structures.

3 GENERAL SPECIFICATIONS OF THE MONITORING SYSTEM

As stated in the introduction, the monitoring system should be capable of handling both dynamic vibration signals and static variables, as well as of the wind conditions. To this aim, high reliable low noise sensors have been selected and a well-known already tested and employed solution has been implemented for data acquisition[5].

3.1 Selected sensors

The chosen sensors had to fulfill all the requirements which were proposed to Regione Lombardia, formerly the work committer. In order to be able to monitor vibration comfort levels against wind serviceability [11][12] and perform operational modal analysis identifications [9][10] very high sensitivity low noise accelerometers had to be employed. Moreover, the building first natural frequency is around 0.3 Hz [1][3], thus posing problems both on the sensor choice and the data acquisition hardware.

On the other hand, the selected tilt sensors had to provide long term stability and a certified temperature sensitivity in order to guarantee reliability to the static measurements.

As a cabled solutions have been chosen (synchronized wireless measurements were not considered affordable on such a high building) the sensor types were chosen in order to minimize the needed cabling.

According to all the above stated needs, the following sensors were chosen:

- Accelerometers: PCB 393B31 Piezo units, which have proven to have a very low noise floor level and guarantee good frequency response down to 0.1 Hz, having a 0.5 g full scale value.

- Clinometers: ±5° Singer TS servo clinometers with extended temperature calibration. High reliability sensors with a frequency response up to 3 Hz, which is enough to cover the building first frequencies
- Wind speed/direction: Anemometer NESA ANS-VV1-A + ANS-DVE-A (potentiometric wind direction), with a 50 m/s full scale value.

Al the installed sensors are connected through custom built connection boxes and multipolar shielded cables to the data acquisition device. Sensor positions are described in the next section.

3.2 Sensor positioning

To the aim of identifying at least the first three vibration modes of the structure (2 bending modes and the 3rd torsional one [1][3]) and according to the commitment request, a total number of 24 piezo accelerometers have been installed at five different levels of the Core 1 tower.

Figure 2 shows the instrumented floors and the different sensors that are installed.

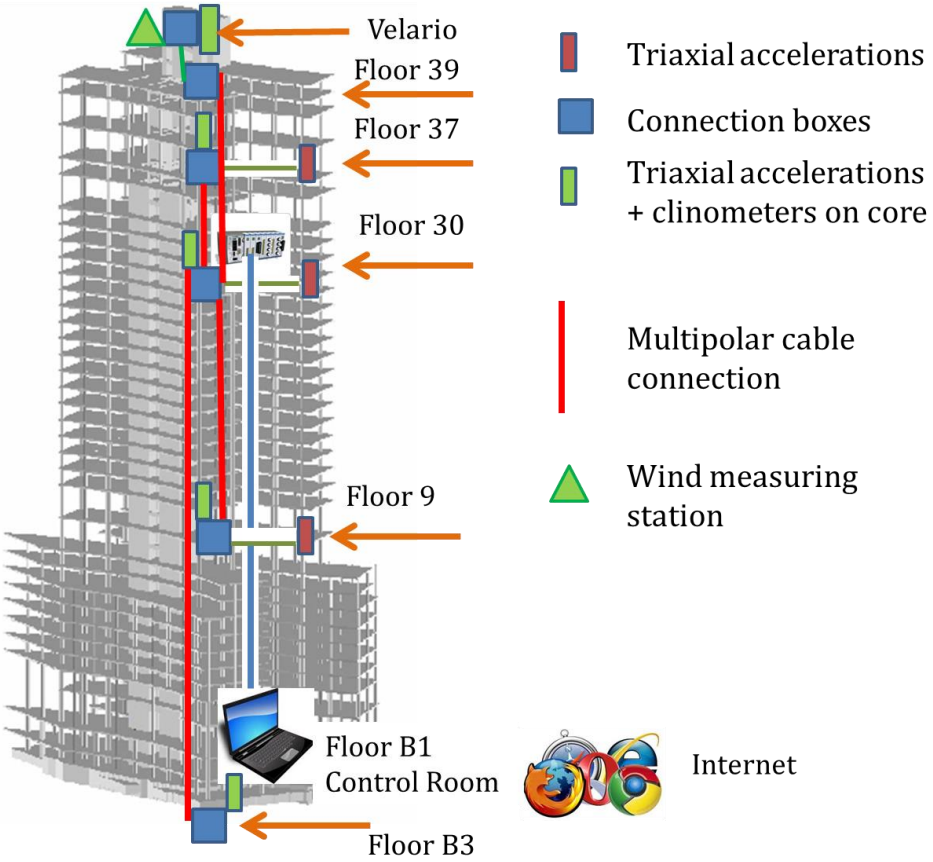


Figure 2 Sensor positions in the building

As can be noticed in Figure 2, sensors are placed starting from the top steel structure "Velario", to the last underground basement (B3 floor). Floor 39 has no sensors and only a couple of connection boxes are installed there in order to facilitate system maintenance.

Sensor position at each flor is depicted in Figure 3 for floors 37/30/09. The B3 and Velario floors are instrumented with sensors in the core area.

Clinometers are placed at all instrumented levels on the concrete core in order to detect rotations around the X and Y axes.

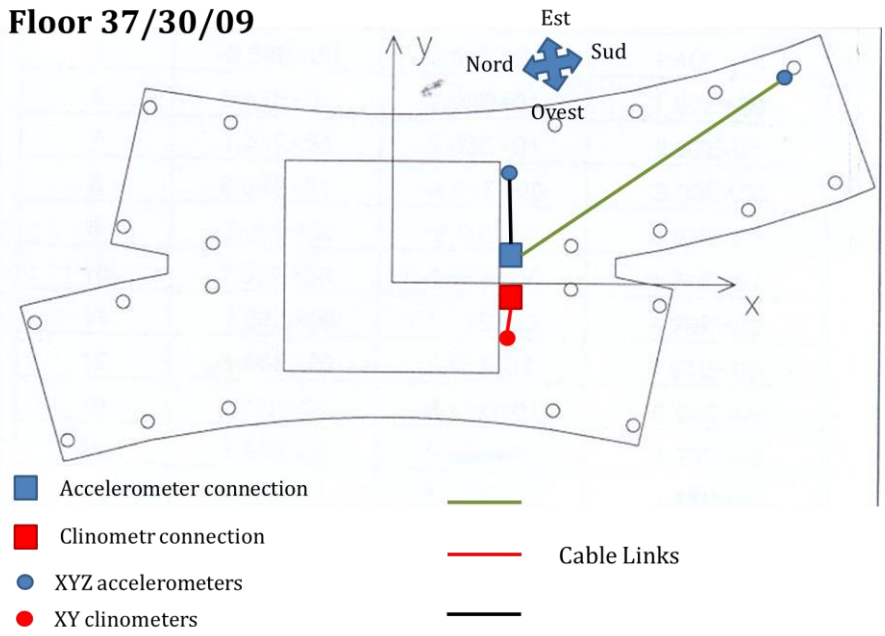


Figure 3 Sensor positions on floor 37/30/09, on floor B3 and on the "Velario" only the Core sensors are installed.

According to the sensor positions illustrated in Figure 3, it is possible to identify and distinguish the two principal bending modes and the first torsional one. It will also be possible to study the higher modes and compare them with what was identified during the validation testing at the end of the construction phase [1][3].

Accelerometers at level B3 will provide information in case of seismic events on the earthquake strength at the building foundations and therefore to evaluate the overall structural amplification. Clinometers at the same level will be used to track deformations on the building foundations.

All sensors are directly cabled to the data acquisition system described in the next section.

3.3 Data acquisition and storage

The data acquisition and storage system was designed in order to guarantee continuous functioning 24hours/7days-per-week . In order to fulfill the design needs the system is split into 2 independent units:

- Data acquisition unit with a local storage capability on floor 30
- Data storage and visualization on floor B1

The data acquisition device is a National Instrument C-RIO unit, in charge of measuring the acceleration and inclination signals together with the wind information and assemble them in 10-minutes-long data files which are then stored locally. The accelerometers are powered and conditioned using PCB ICP power units able to guarantee a low cut-off frequency below 0.1 Hz.

All accelerometers signals are converted using 24 bits simultaneous conversion and built-in anti-aliasing filters. Clinometers and wind data are sampled using 16 bit converters. The analog to digital sampling frequency is 2000 Hz, which is then digitally down-sampled 8 times to 250 Hz.

In Figure 4 it is possible to see some pictures of the data acquisition cabinet placed at the 30th floor.

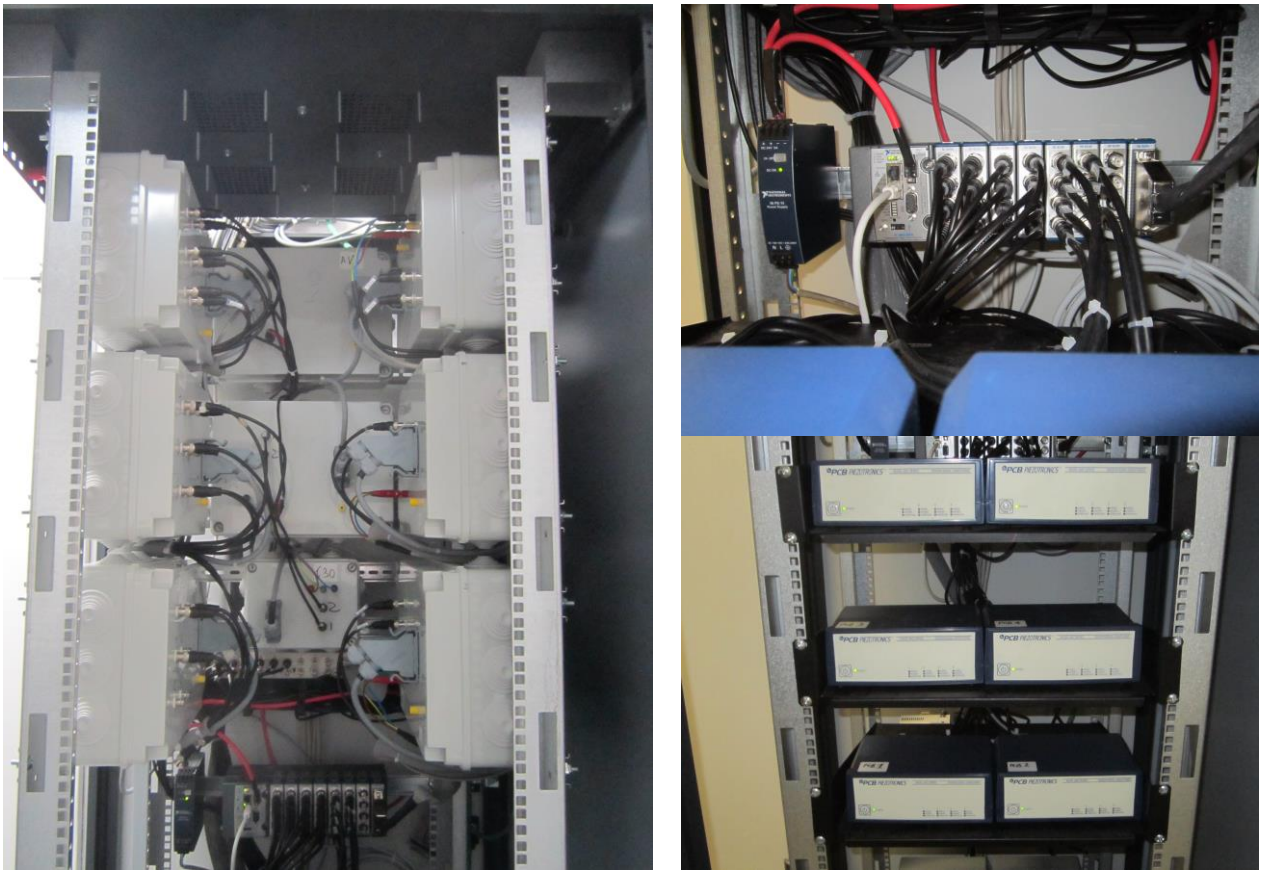
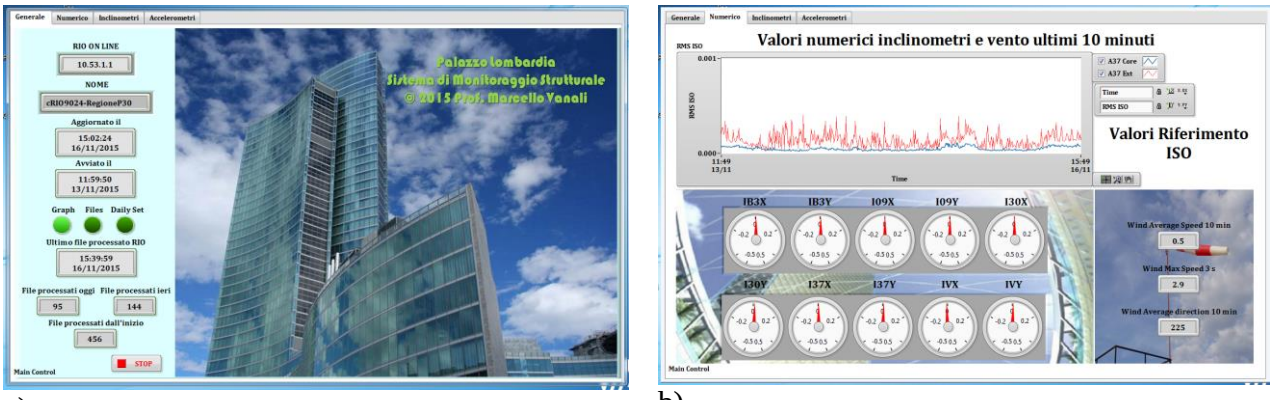


Figure 4 Data acquisition cabinet at floor 30: whole view (left), CRIO and IEPE conditioner (right)

Every time a new data file is stored on the local unit, the system automatically transfers it via Ethernet to the storage unit at floor B3. In case of an Ethernet failure, the local system can store up to two days of data in order not to lose any critical event. System power is assured by the building preferential line that guarantees continuous power supply to all critical systems.

The storage unit at floor B1 is a common server grade PC with raid data storage and a software that periodically check system functionality and, every time a new data file is found, process it to extract some synthetic data. At the moment, all raw data are stored on the B1 floor unit, but in the future a selection of the raw data to be stored is planned based on the acceleration RMS values and wind speed values.



a) b)

Figure 5 Screenshot of the data storage and visualization Software. a) Main Screen, b) Resume of the last ten minutes and wind induced vibrations

Among the synthetic information given by the software, it is worth citing a the vibration RMS magnitude of floor 37 in the frequency range of wind induced vibrations [11] that gives an immediate feedback of possible wind serviceability problems, and the average value of all clinometers sensors in the last ten minutes and their evolution in the last 48 hours.

The system is running since 19 October 2015 and the first data together with some analysis are presented in the next section.

4 FIRST DATA

As previously stated, the system was started on the 19th of October 2015 and is running continuously since then. The first data coming from the system are shown here.

Figure 6 shows the average value evolution of all clinometers since October 19th (the zero value was set at the first valid file that was measured).

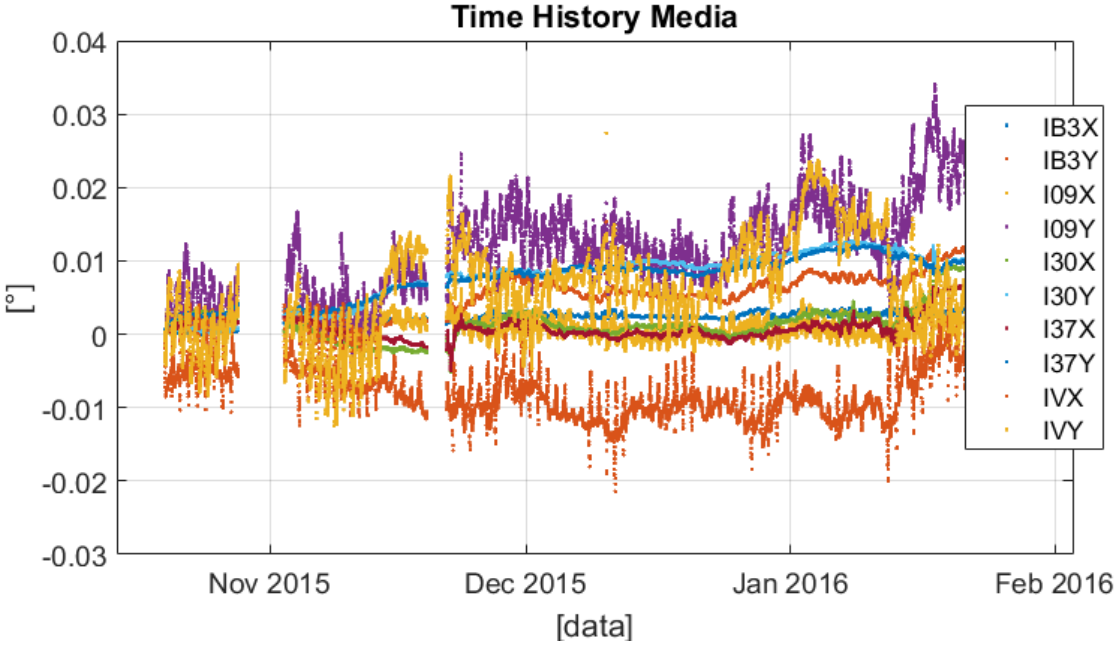


Figure 6 Clinometers average values up to 31/01/2016

As can be noticed in Figure 6 two main signal family are present. The first one (IB3,I30, I37) which shows few oscillations and a constant trend, the second one (I09, IV) that shows wide intraday oscillations.

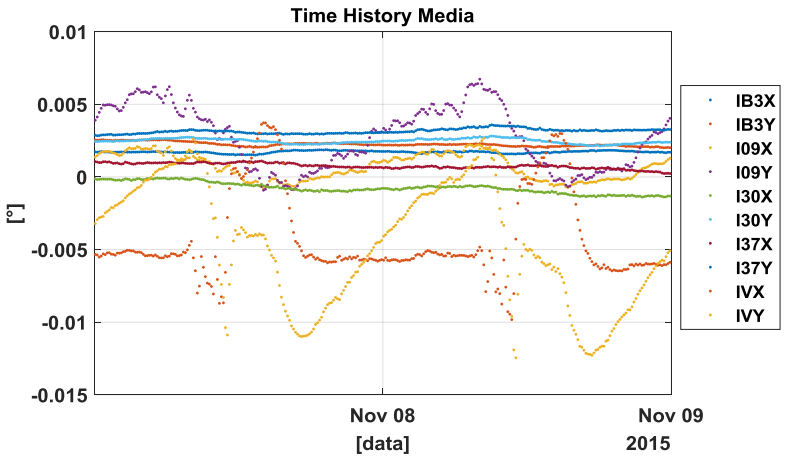


Figure 7 Two-days zoom of clinometers average value

The difference is highlighted in Figure 7 where the clinometers average value evolution over two days is shown. In order to verify and understand this behavior, a temperature sensor has been installed at floors 09 and 30, close to the clinometers. Measurements showed that the higher oscillations experienced at floor 09 are strictly linked to the local temperature and the same applies for clinometers at the "Velario" level, which are outside the building.

The overall variability is around 0.01 ° and the trend will be kept under observation in order to quantify the one-year thermal cycle of the building.

According to what was stated at the beginning, acceleration measurement are taken to the aim of identifying continuously the building modal parameters and to provide information on the vibration serviceability against the wind induced vibrations.

Figure 8 shows the power spectral densities (PSD) measured during a windy day at level 37. The three main vibration modes (1 weak axis flexural, 2 Strong axis flexural, 3 Torsion) are put into evidence by the blue arrows.

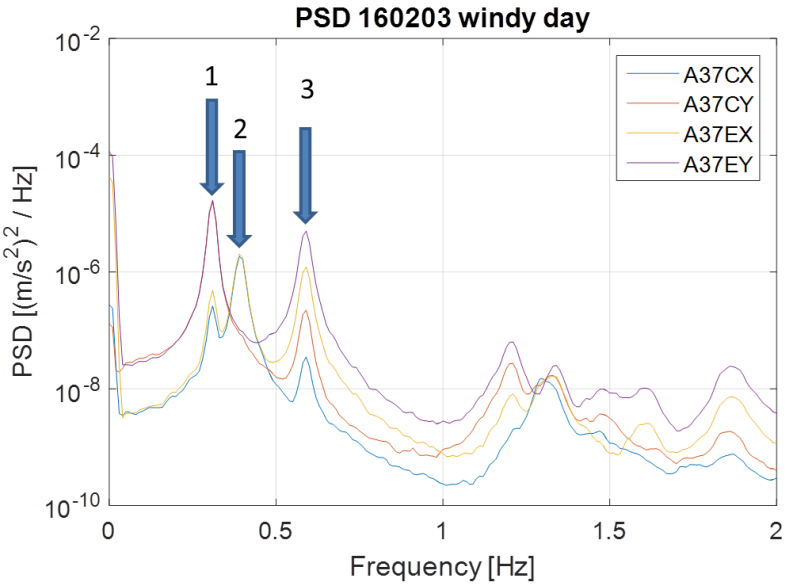


Figure 8 Power Spectral densities at floor 37, 37CXY core accelerations, 37EXY extremity accelerations

The main frequency values and the associated mode shapes are in a very good agreement to the ones found during forced testing [1][3]. Figure 9 shows the same day PSDs in Y direction at different levels on the core.

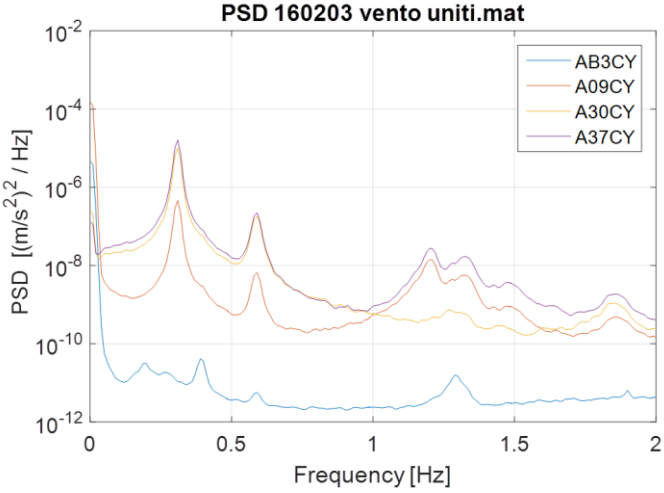


Figure 9 PSD core accelerations at different levels

Vibration levels shown in Figure 9 are increasing from level B3 (bottom) to level 37 (top) of the building in accordance to the expected mode shapes. This measurement will be the input data to an automated modal extraction procedure which will provide synthetic information on the building dynamics [13].

As for the wind induced vibrations, the most critical levels are the highest ones. In Figure 10 the RMS accelerations measured at floor 37 and B3 on the core are shown together with the wind speed. It is noticed that the measured accelerations increase as the wind speed raises even if the reached levels are far below the comfort limits [11][12].

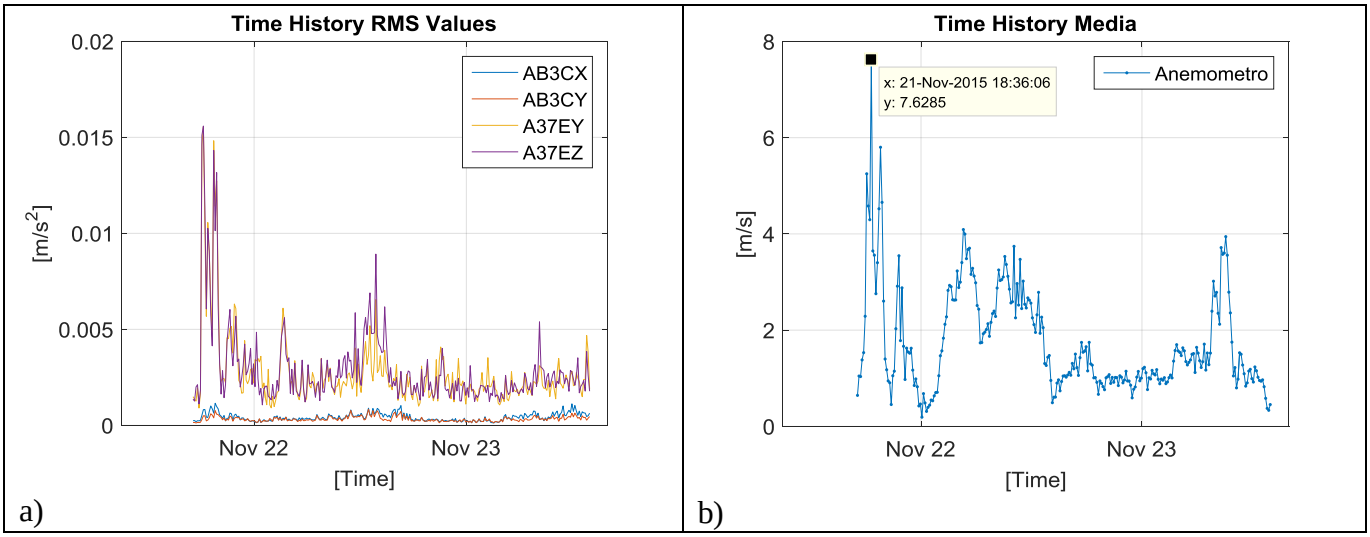


Figure 10 Windy day, a) RMS accelerations on floor 37, b) "Anemometro", measuring the wind speed

The filtered RMS value, averaged on 10 minutes and in the 0.1-1 Hz frequency band, at floor 37 is continuously computed and displayed as one of the synthetic data from the program running on the PC at floor B1 (Figure 11), so that any increase in this level is immediately evident.

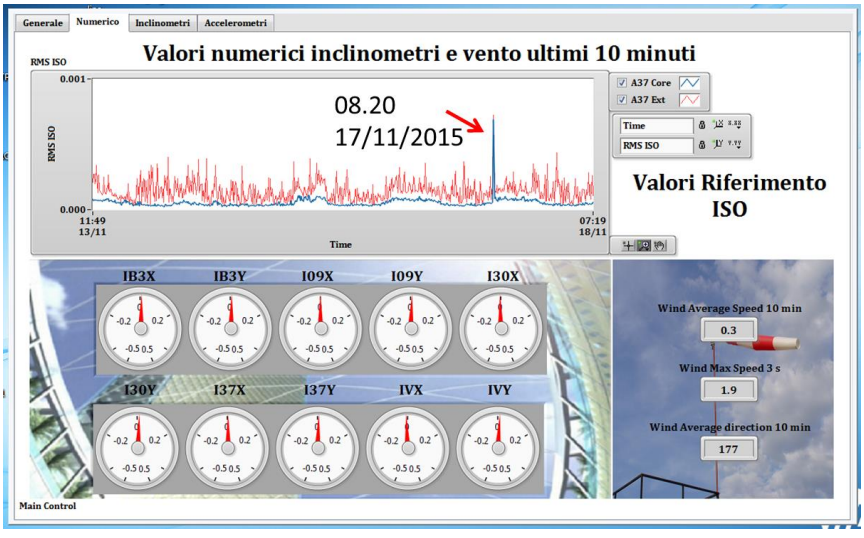


Figure 11 Singular event on November 17th 2015

Figure 11 also shows an anomalous increment in the structural vibration measured at floor 37, occurred on November 17th 2015 around 08.20am hours. This event was not related to the wind or other specific local episodes (e.g. works and/or maintenance operations at the floor).

It was then decided to investigate better this RMS peak. At first it was found that the same peak could be experienced in the structural frequency range at all building levels, including the B3 (basement level). The only remarkable event that happened in that time frame was the earthquake in the Greek Ionic coast as can be seen in Figure 12, where the web page of the Istituto Nazionale di Geofisica e Vulcanologia (INGV) reports the event (green arrow).

Figure 12 Web page of INGV reporting the Greek Earthquake on Novembre 17th.

We contacted the INGV in the person of Dr. Paolo Augliera and he confirmed that very low frequency shock waves travel very far during earthquakes and they are used to seismic prediction purposes [14].

We computed the acceleration PSD during the event and compared it to the usual one; the results are shown in Figure 13. A noticeable increase in the vibration levels below 1 Hz is experienced during the earthquake (terremoto).

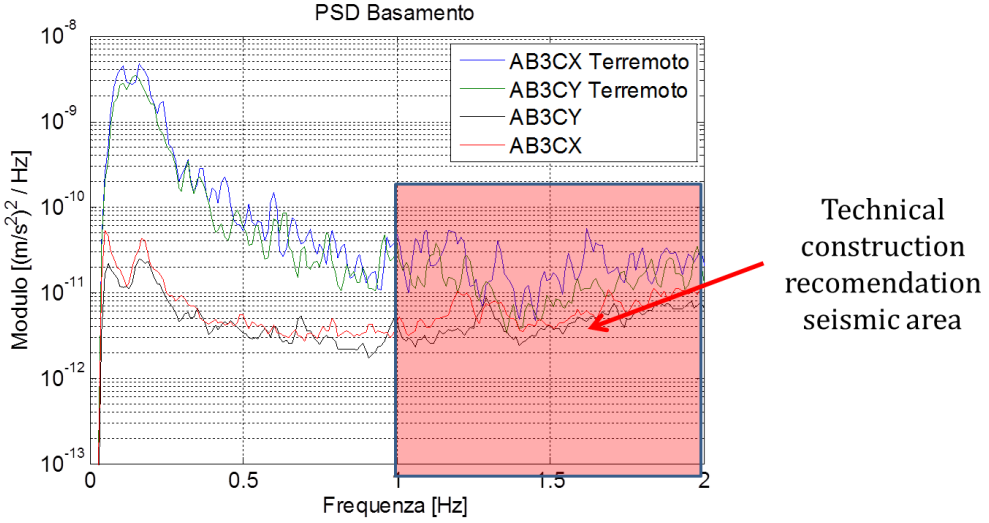


Figure 13 PSD at level B3, normal activity VS Terremoto (Earthquake) of November 17th.

It is noticed that the frequency range involved in this activity is completely out of those considered in NTC08 (technical construction recommendation) [15] for the Milano area which is highlighted in red in Figure 13.

Other similar events have been experienced and recorded since then, proving the system capability to evidence any peculiar situation occurring to the building.

5 CONCLUSION

In this paper the permanent monitoring system for "Palazzo Lombardia" design an installation has been described. The system aims have been illustrated and explained and the instruments chosen to fulfill them have been described. The data acquisition set-up that was designed to guarantee continuous measurements and analysis of the data was presented and finally a first glance to measured data has been given.

First results showed the system capability to follow the building static and dynamic evolution and therefore to provide a valid database for future health monitoring. Moreover, the system proved to be capable to put into evidence a number of unusual events linked to earthquakes which occurred far away from the building location.

ACKNOWLEDGMENTS

The authors wish to acknowledge the Lombardia Government authority and the people at "CarboThermo" group (facility manager) for the support given.

REFERENCES

[1] Cigada, E. Mola, F. Mola, G. Stella, and M. Vanali, "Modal analysis of the Palazzo Lombardia Tower in Milano," *Conf. Proc. Soc. Exp. Mech. Ser.*, vol. 1, pp. 401–413, 2012.

[2] G. Busca, A. Cigada, E. Mola, F. Mola, and M. Vanali, "Dynamic testing of a helicopter landing pad: Comparison between operational and experimental approach," *J. Civ. Struct. Heal. Monit.*, vol. 4, no. 2, pp. 133–147, 2014.

[3] Cigada, E. Mola, F. Mola, G. Stella, and M. Vanali, "Dynamic Behavior of the Palazzo Lombardia Tower : Comparison of Numerical Models and Experimental Results," J. Perform. Constr. Facil., vol. 28(3), no. JUNE, pp. 491–501, 2014.

[4] Cigada, A. Caprioli, M. Redaelli, and M. Vanali, "Vibration Testing at Meazza Stadium: Reliability of Operational Modal Analysis to Health Monitoring Purposes," *J. Perform. Constr. Facil.*, vol. 22, no. 4, pp. 228–237, Aug. 2008.

[5] Cigada, G. Moschioni, M. Vanali, and A. Caprioli, "The measurement network of the san siro meazza stadium in milan: Origin and implementation of a new data acquisition strategy for structural health monitoring: Dynamic testing of civil engineering structures series," *Exp. Tech.*, vol. 34, no. 1, pp. 70–81, 2010.

[6]]Busca, G., Cappellini, A., Cigada, A., Scaccabarozzi, M., & Vanali, M. (2011). Dynamic properties of the Guglia Maggiore of the Duomo in Milano via Operational Modal Analisys. EVACES - Experimental Vibration Analysis for Civil Engineering Structures, (p. 1-8). Varenna (LC) - Italy.

[7] J. M. W. Brownjohn, "Structural health monitoring of civil infrastructure," Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 365, no. 1851, pp. 589–622, Feb. 2007.

[8] C. R. Farrar and K. Worden, "An introduction to structural health monitoring," Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 365, no. 1851, pp. 303–315, Feb. 2007.
[9] C. R. Farrar and G. H. James III, "SYSTEM IDENTIFICATION FROM AMBIENT VIBRATION MEASUREMENTS ON A BRIDGE," J. Sound Vib., vol. 205, no. 1, pp. 1–18, Aug. 1997.
[10] P. Mohanty and D. J. Rixen, "Operational modal analysis in the presence of harmonic excitation," J. Sound Vib., vol. 270, no. 1–2, pp. 93–109, Feb. 2004.
[11] L. G. Griffis, "Serviceability limit states under wind load," Eng. J., vol. 30, no. 1, pp. 1–16, 1993.
[12] P. Mendis, T. Ngo, N. Haritos, A. Hira, B. Samali, and J. Cheung, "Wind loading on tall buildings," EJSE Spec. Issue Load. Struct., vol. 3, pp. 41–54, 2007.
[13] Cattaneo A, Manzoni S, Vanali M, Measurement uncertainty in operational modal analysis of a civil structure, in Proceedings of the International Conference on Uncertainty in Structural Dynamics – USD 2010, p. ,5103-5116, Leuven (Blegium), 20-22 September 2010.
[14] M. Massa and P. Augliera, "Teleseisms as Estimators of Experimental Long-Period Site Amplification: Application to the Po Plain (Italy) for the 2011 Mw 9.0 Tohoku-Oki (Japan) Earthquake," Bull. Seismol. Soc. Am., vol. 103, no. 5, pp. 2541–2556, Oct. 2013.
[15] Consiglio Superiore dei Lavori Pubblici. (2008). Nuove Norme Tecniche per le Costruzioni, DM 14-01-08, Rome (in Italian).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

AN INVERSE PROBLEM STRATEGY BASED ON FORWARD MODEL EVALUATIONS: GRADIENT-BASED OPTIMIZATION WITHOUT ADJOINT SOLVES

Miguel A. Aguiló

Computational Solid Mechanics and Structural Dynamics
Sandia National Laboratories, PO Box 5800, MS 0380 Albuquerque, NM 87185-0380, USA
e-mail: maguilo@sandia.gov

Keywords: Inverse problem, Compliance error minimization, Gradient-based optimization, Newton's method, Nonlinear conjugate gradient, Heat transfer.

Abstract. *A new inverse problem formulation based on the compliance error functional is presented. The proposed formulation permits the derivation of an explicit expression for the Lagrange multipliers. Thus, the computation of the Lagrange multipliers does not require the solution of the computationally intensive adjoint system of equations. This enables the computation of the first order derivative information at the expense of just one model evaluation. Leading to significant speedups for large-scale, gradient-based inverse problems.*

Second, if second order optimization algorithms are available, Newton's method can be applied to the first order necessary optimality conditions. Newton's method relies on second order derivative information to compute descent directions during optimization. This paper presents two Hessian formulations based on the compliance error functional. The first formulation relies on the mathematical properties of the compliance error functional to compute the application of the trial step to the analytical Hessian at the expense of one model evaluation per Newton iteration. The second formulation is based on a Gauss-Newton approximation to the Hessian. The Gauss-Newton approximation further speedups analysis since the second order derivative information is computed without performing additional model evaluations during the Newton iterations.

Third, examples in heat transfer are presented to demonstrate the effectiveness of the compliance error functional. The compliance error minimization formulation is compared against the data misfit minimization formulation. Results will show that the compliance error minimization formulation outperforms the data misfit formulation.

1 INTRODUCTION

This paper investigates the problem of large-scale parameter estimation, i.e. inverse problem, in the context of a particular problem: determining the thermal conductivity properties of a region of interest given an 'observed' temperature field. Specifically, a new inverse problem formulation based on the compliance error functional is presented. A general inverse problem framework for the application of first and second order optimization algorithms is presented, The proposed framework enables the implementation of the compliance error minimization formulation to many applications of interests in science and engineering.

Multiple inverse problem techniques have been proposed in the literature for the solution of inverse problems. For instance, many researchers have applied data misfit error functionals for the solution of inverse problems in many physics settings [1, 2, 3, 4]. The objective of data misfit functional is to characterize the parameter of interest by minimizing the discrepancy between the simulation and the experimental data (target/observed data). An alternative to data misfit functionals is the modified error in the constitutive equation (MECE) functional. The objective of the MECE functional is to characterize the parameter of interests by minimizing the discrepancy in the constitutive equations [5, 6, 7, 8, 9, 10, 11, 12]. This formulation leads to a coupled system of equations that is solved for any new set of materiel parameters.

Allix *et al.* [13] performed several numerical studies that demonstrated that the MECE functional improved the convexity of the objective function. Furthermore, Gockenbach *et al.* [14, 15, 16] showed that the energy norm term of the MECE functional is convex for elliptic boundary value problems when full field measurements are available. However, in a subsequent study, Gockenbach [17] showed that inverse problem formulations based on energy norm minimization can lead to inaccurate estimates if the 'observed' data is corrupted. Moreover, he showed that the data misfit functional was less sensitive to data corruption than the energy norm functional. Lastly, the virtual field method (VFM) is a recent inverse problem technique developed for extracting constitutive parameters from full-field measurements [18]. The VFM combines the principle of virtual work with kinematically admissible virtual fields, which are computed *a-priori*, to characterize the constitutive parameters. The choice of the kinematically admissible virtual field is key in order to improve the performance of the VFM. Thus, most research is focus on improving the predictive reliability of the kinematically admissible virtual fields [19].

In this work, a new inverse problem formulation based on the compliance error functional is presented. The main advantage of the compliance error functional is that it enables the computation of the Lagrange multipliers, and thus the first order derivative information, at the expense of just one model evaluation. Thus, the calculation of the Lagrange multipliers does not require the solution of the computationally intensive adjoint system of equations. This leads to significant speedups for large-scale, gradient-based inverse problems since the gradient computation is only based on forward model evaluations. The proposed formulation also simplifies implementation in production software since only forward model evaluations are needed during optimization.

If second order optimization algorithms are available, Newton's method can be applied to the first order necessary optimality conditions. Newton's method relies on second order derivative information to compute descent directions during optimization. If the data misfit functional is used, two model evaluations are required per Newton iteration to compute the second order derivative information. This paper presents two Hessian formulations based on the compliance error functional. The first formulation relies on the mathematical properties of the compliance

error functional to reduce the number of model evaluations from two to one without sacrificing accuracy. Consequently, substantial speedups are gained. The second formulation relies on a Gauss-Newton Hessian approximation based on the compliance error functional to further speedup the Newton solver. By using a Gauss-Newton approximation, the second order derivative information is computed without performing additional model evaluations during the Newton iterations. Results will demonstrate that the compliance error functional leads to substantial speedups over the the data misfit functional.

This paper is organized as follows: Section 2 presents the general first and second order inverse problem formulations for the compliance error and data misfit functionals. Furthermore, the respective general inverse problem solution framework is presented for both formulations. Section 3 presents an inverse problem example in heat transfer to showcase the performance of the compliance error functional. Finally, Section 4 provides concluding remarks and future research directions.

2 FORMULATION

2.1 Inverse problem

Let $\Omega \subseteq \mathbb{R}^d$, $d \in \{1, 2, 3\}$ denote the computational domain with boundary $\partial\Omega$. Lets now define the Lebesgue space $\mathbb{H} = L^2(\Omega; \mathbb{R}^{\mathsf{n}})$ of measurable and square intregrable functions endowed with inner product $\langle \phi, \psi \rangle_{\mathbb{H}} = \int_\Omega \phi\, \psi$ for $\phi, \psi \in \mathbb{H}$ and norm $\|\phi\|_{\mathbb{H}} = \langle \phi, \phi \rangle_{\mathbb{H}}^{1/2}$. Lets also define finite dimensional spaces $\mathbb{U} = \{\text{span}\{\phi^a\}_{a=1}^A \,|\, \phi \in \mathbb{H}\} \subset U$, $\mathbb{Z} = \{\text{span}\{\psi^b\}_{b=1}^B \,|\, \psi \in \mathbb{H}\} \subset Z$, and $\mathbb{Y} = \{\text{span}\{\chi^c\}_{c=1}^C \,|\, \chi \in \mathbb{H}\} \subset Y$. This enables then the following finite dimensional approximations for the state, control, and Lagrange multipliers $\mathbf{u} = \sum_a^A \tilde{u}^a \phi^a \,|\, \tilde{u} \in \mathbb{R}$, $\mathbf{z} = \sum_b^B \tilde{z}^b \psi^b \,|\, \tilde{z} \in \mathbb{R}$, and $\mathbf{v} = \sum_c^C \tilde{v}^c \chi^c \,|\, \tilde{v} \in \mathbb{R}$, respectively.

Lets now define a general parameter estimation (inverse) problem as

$$\begin{aligned} \min_{(\mathbf{u},\mathbf{z}) \in \mathbb{U} \times \mathbb{Z}} \quad & \mathsf{J}(\mathbf{u}, \mathbf{z}) \\ \text{s.t.} \quad & \\ & \mathsf{g}(\mathbf{u}, \mathbf{z}) = 0, \end{aligned} \tag{1}$$

where $\mathbf{u}$ and $\mathbf{z}$ respectively denote the state and control variables, $\mathsf{J}(\mathbf{u}, \mathbf{z}) \colon \mathbb{U} \times \mathbb{Z} \to \mathbb{R}$ denotes the objective function and $\mathsf{g}(\mathbf{u}, \mathbf{z}) \colon \mathbb{U} \times \mathbb{Z} \to \mathbb{Y}$ denotes the equality constraint (physics model).

The implicit function theorem admits the definition of a solution operator $\mathbf{u} \colon \mathbb{Z} \to \mathbb{U}$ such that $\{(\mathbf{u}(\mathbf{z}), \mathbf{z}) \mid \mathbf{z} \in \mathbb{Z}\} = \{(\mathbf{u}, \mathbf{z}) \in \mathbb{U} \times \mathbb{Z} \mid \mathsf{J}(\mathbf{u}, \mathbf{z}) = 0\}$. This enables the redefinition of Equation 1 as

$$\min_{\mathbf{z} \in \mathbb{Z}} \quad \mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z}), \tag{2}$$

where the solution operator $\mathbf{u}(\mathbf{z})$ is obtained by solving $\mathsf{g}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = 0$. This formulation is commonly known as the reduced-space formulation for partial differential equation (PDE) constrained optimization.

Finally, lets assume that the objective function and equality constraint are given by

$$\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = \frac{\beta}{2} \| \langle \mathbf{u}(\mathbf{z}), \mathbf{A}(\mathbf{z})\mathbf{u}(\mathbf{z}) \rangle_{\mathbb{H}} - \langle \hat{\mathbf{u}}, \mathbf{A}(\mathbf{z})\hat{\mathbf{u}} \rangle_{\mathbb{H}} \|_{\mathbb{H}}^2 + \mathsf{R}(\mathbf{z}), \tag{3}$$

$$\mathsf{g}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = \mathbf{A}(\mathbf{z})\mathbf{u} - \mathbf{f} = 0, \tag{4}$$

where $\mathbf{A}(\mathbf{z})\colon \mathbb{Z} \to \mathbb{U} \times \mathbb{U}$ is a non-singular, self-adjoint linear operator, $\mathsf{R}(\mathbf{z})\colon \mathbb{Z} \to \mathbb{R}$ is a regularization functional, $\mathbf{f} \in \mathbb{Y}$ is an external force and $\hat{\mathbf{u}} \in \hat{\Omega} \subset \Omega$ denotes measured data. This formulation will be referred as the compliance error minimization (CEM) formulation.

2.2 First order formulation

2.2.1 Data misfit functional

Lets define the Lagrangian functional $\mathcal{L}\colon \mathbb{U} \times \mathbb{Z} \times \mathbb{Y} \to \mathbb{R}$ for Equation 2 as

$$\mathcal{L}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = \mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z}) + \langle \mathbf{v}, \mathsf{g}(\mathbf{u}(\mathbf{z}), \mathbf{z}) \rangle_{\mathbb{Y}^*, \mathbb{Y}}, \tag{5}$$

where $\mathbf{v}$ denotes the Lagrange multipliers and $\mathbb{Y}^*$ is the dual space of $\mathbb{Y}$. The data misfit objective function is given by

$$\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = \frac{1}{2} \|\mathbf{u}(\mathbf{z}) - \hat{\mathbf{u}}\|^2_{\mathbb{H}}. \tag{6}$$

The equality constraint $\mathsf{g}(\mathbf{u}(\mathbf{z}), \mathbf{z})$ is given by Equation 4. If $\mathbf{z}^\star \in \mathbb{Z}$ is a local solution of Equation 2; then, there exists a set of Lagrange multipliers $\mathbf{v}^\star \in \mathbb{Y}$ such that the first order necessary optimality conditions are satisfied at $\mathbf{z}^\star$.

The first order necessary optimality conditions are given by

$$\mathcal{L}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = \mathsf{J}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z}) + \mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})^* \mathbf{v} = 0 \tag{7}$$

$$\mathcal{L}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = \mathsf{J}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z}) + \mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z})^* \mathbf{v} = 0, \tag{8}$$

where the subscripts $\mathbf{u}$ and $\mathbf{z}$ respectively denote derivatives with respect to the state and control variables. The Lagrange multipliers are computed by solving

$$\mathbf{v} = -(\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})^*)^{-1} \mathsf{J}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z}). \tag{9}$$

Substituting Equation 9 into Equation 8 yields a reduced gradient operator of the form

$$\nabla \mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = \mathsf{J}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z}) + \mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z})^* [-(\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})^*)^{-1} \mathsf{J}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})]. \tag{10}$$

At each optimization iteration, the following sequence of steps are performed to compute the reduced gradient operator and minimize the objective function:

1. Solve equality $\mathsf{g}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = 0$ for $\mathbf{u} \in \mathbb{U}$;
2. Solve $\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})^* \mathbf{v} = -\mathsf{J}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})$ for $\mathbf{v} \in \mathbb{Y}$;
3. Compute the reduced gradient operator defined in Equation 10;
4. Compute descent direction $\mathbf{s} \in \mathbb{Z}$ and set $\mathbf{z}_{k+1} = \mathbf{z}_k + \gamma \mathbf{s}_k$, $\gamma \in \mathbb{R}$.

This sequence of steps are necessary to solve an inverse problem based on the data misfit functional [20].

2.2.2 Compliance error functional

Assume that the objective function and equality constraint for the parameter estimation problem defined in Equation 2 are given by Equations 3 and 4, respectively. Then, the first order derivative operators $\mathsf{g}_{\mathbf{u}}(\mathbf{u},\mathbf{z})^*$ and $\mathsf{J}_{\mathbf{u}}(\mathbf{u},\mathbf{z})$ are given by

$$\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^* = \mathbf{A}(\mathbf{z})^* \tag{11}$$

and

$$\mathsf{J}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) = \alpha\mathbf{A}(\mathbf{z})\mathbf{u}, \tag{12}$$

where

$$\alpha = 2(\langle\mathbf{u}(\mathbf{z}),\mathbf{A}(\mathbf{z})\mathbf{u}(\mathbf{z})\rangle_{\mathbb{H}} - \langle\hat{\mathbf{u}},\mathbf{A}(\mathbf{z})\hat{\mathbf{u}}\rangle_{\mathbb{H}}). \tag{13}$$

Recall that $\mathbf{A}(\mathbf{z})$ is assumed to be a non-singular, self-adjoint linear operator. Thus, an explicit expression for the Lagrange multipliers can be derived by substituting Equations 11 and 12 into Equation 9. After substitution, the Lagrange multipliers are given by

$$\mathbf{v} = -\alpha\beta\mathbf{A}(\mathbf{z})^{-1}(\mathbf{A}(\mathbf{z})\mathbf{u}(\mathbf{z})) = -\alpha\beta\mathbf{I}\mathbf{u}(\mathbf{z}) = -\alpha\beta\mathbf{u}(\mathbf{z}), \tag{14}$$

where $\mathbf{I}$ denotes the identity linear operator. Substituting Equation 14 into Equation 10 gives a reduced gradient operator of the form

$$\nabla_{\mathbf{z}}\mathsf{J}(\mathbf{u}(\mathbf{z}),\mathbf{z}) = \mathsf{J}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) - \alpha\beta\langle\mathbf{u}(\mathbf{z}),\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\rangle. \tag{15}$$

Notice that the adjoint problem defined in Equation 9 is not solved to compute the Lagrange multipliers, which should speedup the optimization problem.

At each optimization iteration, the following sequence of steps are performed to compute the reduced gradient operator and minimize the compliance error objective function:

1. Solve equality $\mathsf{g}(\mathbf{u}(\mathbf{z}),\mathbf{z}) = 0$ for $\mathbf{u} \in \mathbb{U}$;
2. Compute Lagrange multipliers $\mathbf{v} = -\alpha\beta\mathbf{u}(\mathbf{z})$;
3. Compute the reduced gradient operator given by Equation 15.
4. Compute descent direction $\mathbf{s} \in \mathbb{Z}$ and set $\mathbf{z}_{k+1} = \mathbf{z}_k + \gamma\mathbf{s}_k$, $\gamma \in \mathbb{R}$.

In addition to the speedups gained due to the omission of the adjoint model evaluations during optimization, the CEM formulation facilitates implementation since it only relies on forward model evaluations to compute the reduced gradient operator.

2.3 Second order formulation

2.3.1 Data misfit functional

If second order derivative optimization algorithms are available, Newton's method can be applied to the first order necessary optimality conditions. Let $\kappa \in \mathbb{R}^*_+$ and $\delta\mathbf{z} \in \mathbb{Z}$, if $\mathbf{z}^\star \in \mathbb{Z}$ satisfy the first order necessary optimality conditions and

$$\langle\delta\mathbf{z},\nabla^2\mathsf{J}(\mathbf{u}(\mathbf{z}^\star)\mathbf{z}^\star)\delta\mathbf{z}\rangle \geq \kappa\|\delta\mathbf{z}\|^2_{\mathbb{H}} \;\forall\; \delta\mathbf{z} \in \ \ker\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}^\star),\mathbf{z}^\star),$$

the second-order sufficient condition is satisfied at $\mathbf{z}^\star$. Furthermore, $\mathbf{z}^\star$ is a strict local minimum of Equation 2.

The application of the trial step $\delta\mathbf{z}$ to the Hessian operator is given by

$$\nabla^2 \mathsf{J}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \mathcal{L}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{u} + \mathcal{L}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{z} + \mathcal{L}_{\mathbf{zv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{v}, \tag{16}$$

for $\delta\mathbf{u} \in \mathbb{U}$ and $\delta\mathbf{v} \in \mathbb{Y}$. Notice that $\delta\mathbf{u}$ and $\delta\mathbf{v}$ are necessary to compute the application of the trial step to the Hessian operator. Thus, explicit expressions for $\delta\mathbf{u}$ and $\delta\mathbf{v}$ are necessary to calculate Equation 16.

Let $\mathsf{g}(\mathbf{u}(\mathbf{z}),\mathbf{z}) = 0 \; \forall \; \mathbf{z} \in \mathbb{Z}$. Then, $\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = 0 \; \forall \; (\mathbf{z},\delta\mathbf{z}) \in \mathbb{Z} \times \mathbb{Z}$, where

$$\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} + \mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = 0 \tag{17}$$

and $\delta\mathbf{u} \equiv \mathbf{u}_{\mathbf{z}}(\mathbf{z})\delta\mathbf{z}$. Solving Equation 17 for $\delta\mathbf{u}$ gives

$$\delta\mathbf{u} = -\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^{-1}\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z}. \tag{18}$$

Next, an explicit expression is derived for $\delta\mathbf{v}$. By definition, $\mathcal{L}_{\mathbf{u}}(\hat{\mathbf{u}}(\mathbf{z}),\mathbf{z},\mathbf{v}) = 0 \; \forall \; (\mathbf{u},\mathbf{z},\mathbf{v}) \in \mathbb{U} \times \mathbb{Z} \times \mathbb{Y}$; thus, the derivative of $\mathcal{L}_{\mathbf{u}}(\hat{\mathbf{u}}(\mathbf{z}),\mathbf{z},\mathbf{v})$ in the direction of $\delta\mathbf{z}$ gives

$$\mathcal{L}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{u} + \mathcal{L}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{z} + \mathcal{L}_{\mathbf{uv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{v} = 0, \tag{19}$$

$\forall \quad (\mathbf{z},\delta\mathbf{u},\delta\mathbf{z},\delta\mathbf{v}) \in \mathbb{Z} \times \mathbb{U} \times \mathbb{Z} \times \mathbb{Y}$. Solving Equation 19 for $\delta\mathbf{v}$ gives

$$\delta\mathbf{v} = -\mathcal{L}_{\mathbf{uv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})^{-1}[\mathcal{L}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{u} + \mathcal{L}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{z}], \tag{20}$$

where

$$\mathcal{L}_{\mathbf{uv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v}) = \mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^* \tag{21}$$

$$\mathcal{L}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v}) = \mathsf{J}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) + \mathsf{g}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\mathbf{v} \tag{22}$$

$$\mathcal{L}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v}) = \mathsf{J}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) + \mathsf{g}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\mathbf{v}. \tag{23}$$

The following sequence of steps are performed to compute the application of the trial step $\delta\mathbf{z}$ to the analytical Hessian operator at each Newton iteration:

1. Solve $\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} = -\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z}$ for $\delta\mathbf{u} \in \mathbb{U}$
2. Solve $\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{v} = -[\mathcal{L}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{u} + \mathcal{L}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{z}]$ for $\delta\mathbf{v} \in \mathbb{Y}$
3. Compute the application of the trial step to the reduced Hessian operator

$$\nabla^2 \mathsf{J}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \mathcal{L}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{u} + \mathcal{L}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{z} + \mathcal{L}_{\mathbf{zv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{v}, \tag{24}$$

where $\mathcal{L}_{\mathbf{zv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v}) = \mathsf{g}_{\mathbf{z}}(\hat{\mathbf{u}}(\mathbf{z}),\mathbf{z})^*$.

Notice that the data misfit functional requires two model evaluations per Newton iteration to compute the application of the trial step to the analytical Hessian operator.

2.3.2 Compliance error functional: analytical Hessian

Assume that the objective function and equality constraint are given by Equations 3 and 4, respectively. Then, the second order derivative operators $\mathcal{L}_{\mathbf{uv}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v})$, $\mathcal{L}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v})$, and $\mathcal{L}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v})$ are given by

$$\mathcal{L}_{\mathbf{uv}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = \mathbf{A}(\mathbf{z})^* \tag{25}$$

$$\mathcal{L}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = 2\beta[\alpha\mathbf{A}(\mathbf{z})\delta\mathbf{u} + 2\mathbf{A}(\mathbf{z})\mathbf{u}(\mathbf{z})\langle\mathbf{u}(\mathbf{z}), \mathbf{A}(\mathbf{z})\delta\mathbf{u}\rangle_{\mathbb{H}}] \tag{26}$$

$$\mathcal{L}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = 2\beta[\alpha(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\delta\mathbf{z})\mathbf{u}(\mathbf{z}) + \gamma\mathbf{A}(\mathbf{z})\mathbf{u}(\mathbf{z})] + (\mathbf{A}(\mathbf{z})^*\mathbf{v})\delta\mathbf{u}, \tag{27}$$

where

$$\gamma = \langle\mathbf{u}(\mathbf{z}), (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\delta\mathbf{z})\mathbf{u}(\mathbf{z})\rangle_{\mathbb{H}} - \langle\hat{\mathbf{u}}, (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\delta\mathbf{z})\hat{\mathbf{u}}\rangle_{\mathbb{H}} \tag{28}$$

Recall that $\mathbf{A}(\mathbf{z})$ is a non-singular, self-adjoint linear operator. This enables the derivation of an explicit expression for $\delta\mathbf{v}$ by substituting Equations 14, 25, 26 and 27 into Equation 20. After simplification, $\delta\mathbf{v}$ is given by

$$\delta\mathbf{v} = -2\beta[\alpha\delta\mathbf{u} + 2\mathbf{u}(\mathbf{z})\langle\mathbf{u}(\mathbf{z}), \mathbf{A}(\mathbf{z})\delta\mathbf{u})\rangle_{\mathbb{H}} + \gamma\mathbf{u}(\mathbf{z})], \tag{29}$$

where $\delta\mathbf{u}$ is given by Equation 18.

To compute the application of the trial step to the analytical Hessian operator, the second order derivative operators $\mathcal{L}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v})$, $\mathcal{L}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v})$ and $\mathcal{L}_{\mathbf{zv}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v})$ are required. These second order derivative operators are explicitly given by

$$\mathcal{L}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = 2\beta[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\mathbf{u}(\mathbf{z}) - (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] \tag{30}$$

$$\mathcal{L}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = \beta\gamma[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\mathbf{u}(\mathbf{z}) - (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] + \mathsf{R}_{\mathbf{zz}}(\mathbf{z}) \tag{31}$$

$$\begin{aligned}\mathcal{L}_{\mathbf{zv}}(\mathbf{u}(\mathbf{z}), \mathbf{z}, \mathbf{v}) = -2\beta[&\alpha(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\delta\mathbf{u} + 2(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\mathbf{u}(\mathbf{z})\langle\mathbf{u}(\mathbf{z}), \mathbf{A}(\mathbf{z})\delta\mathbf{u}\rangle_{\mathbb{H}} \\ &+ \gamma(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\mathbf{u}(\mathbf{z})],\end{aligned} \tag{32}$$

where Equations 14 and 29 were used to derived and simplified Equations 30 and 32.

An explicit expression for the application of the trial step to the analytical Hessian operator is obtained by substituting Equations 30-32 into Equation 24. After some simplifications, the application of the trial step to the analytical Hessian operator is given by

$$\begin{aligned}\nabla^2\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z})\delta\mathbf{z} = &-2\beta[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\mathbf{u}(\mathbf{z}) + (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}]\langle\mathbf{u}(\mathbf{z}), \mathbf{A}(\mathbf{z})\delta\mathbf{u}\rangle \\ &- \beta\gamma[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\mathbf{u}(\mathbf{z}) + (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] + \mathsf{R}_{\mathbf{zz}}(\mathbf{z})\delta\mathbf{z} \\ &- 2\beta\alpha(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u}(\mathbf{z}))\delta\mathbf{u}.\end{aligned} \tag{33}$$

Therefore, the following sequence of steps are necessary to compute the application of the trial step $\delta\mathbf{z}$ to the analytical Hessian operator per Newton iteration:

1. Solve $\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}), \mathbf{z})\delta\mathbf{u} = -\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z})\delta\mathbf{z}$ for $\delta\mathbf{u} \in \mathbb{U}$
2. Compute $\nabla^2\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z})\delta\mathbf{z}$ as defined by Equation 33.

The proposed compliance error minimization formulation enables the calculation of the second order derivative information at the expense of one model evaluation per Newton iteration. Contrary, the data misfit formulation presented in Section 2.3.1 requires two model evaluations per Newton iteration. Thus, significant speedups are possible with the CEM formulation by omitting one model evaluation per Newton iteration.

2.3.3 Compliance error functional: Gauss-Newton Hessian

To circumvent the additional computational demands associated with computing Equation 33, a Gauss-Newton Hessian formulation based on the compliance error functional is considered. Lets define the application of the trial step to the Gauss-Newton Hessian as

$$\nabla^2 J(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} \equiv \mathcal{L}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{z} = [J_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) + g_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\mathbf{v}]\delta\mathbf{z}. \tag{34}$$

Notice that the derivative operators $\mathcal{L}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{u}$ and $\mathcal{L}_{\mathbf{zv}}(\mathbf{u}(\mathbf{z}),\mathbf{z},\mathbf{v})\delta\mathbf{v}$ are omitted in Equation 34 since these nonlinear terms vanish. Thus, the solution of Equation 18 is not necessary to compute the application of the trail step to the Gauss-Newton Hessian during the Newton iterations. Results will demonstrate that the Gauss-Newton Hessian approximation based on the compliance error functional can lead to significant speedups without hindering solution accuracy.

3 EXAMPLE IN HEAT TRANSFER

The Intrepid PDE discretization package from Trilinos [21] was used to build the finite element models. The direct solver routine from MATLAB [22] scientific package were used to solve the linear system of equations in this numerical study. The optimization algorithms in this work were implemented in C++ and used to generate the results presented herein [23]. Readers are encourage to explore other optimization library of their preference [22, 24, 25, 26, 27]. Finally, all calculations were performed on a Linux workstation with a 2.93 GHz Intel(R) Core Xeon(R) processor and 24 GB of RAM.

To synthesize the 'observed' temperature field, a finer grid with 80,000 triangles was used to generate the experimental temperature field. The experimental temperature field was then projected onto a computational grid 20,000. This was done to avoid using the same computational mesh used to generate the 'observed' temperature field during optimization. Different levels of random Gaussian noise were also considered, $\Delta \sim N(0,\widehat{\sigma})$ in order to test the tolerance of the proposed formulation to corrupt data. A Gaussian distributed set of random numbers has 65%, 95%, and 99.7% certainty of respectively being within one, two, and three standard deviations from the mean. Lets thus assume that the Gaussian random noise generated to test the compliance error minimization formulation is 95% certain of being within $\theta\%$ of the actual data. Then, the perturbation applied to the 'observed' temperature field is scaled from interval $(-2\widehat{\sigma}, 2\widehat{\sigma})$ to $(-\theta\%, \theta\%)$. This produces a perturbation parameter of the form $\varepsilon^\theta = \frac{1}{2}(\frac{\theta}{100})$ [28]. Therefore, the corrupt data was generated as follows

$$\hat{\mathbf{u}}_i^\theta = \mathbf{u}_i(1+\varepsilon_i^\theta), \text{ for } i = 1,\ldots,n_{\mathbf{u}} \tag{35}$$

where $\theta \in \Theta = \{1\%, 3\%, 5\%\}$ and $n_{\mathbf{u}}$ denotes the number of states.

Finally, the compliance error minimization formulation was compared against data misfit formulation. The objective was to highlight the effectiveness of the CEM formulation versus a common inverse problem formulation strategy. For completeness, the corresponding first and second order derivative operators for the data misfit functional are defined herein

$$\mathsf{J}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) = \beta(\mathbf{u}(\mathbf{z}) - \hat{\mathbf{u}}) \tag{36}$$

$$\mathsf{J}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} = \beta\delta\mathbf{u} \tag{37}$$

$$\mathsf{J}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z}) = \mathsf{J}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} = \mathsf{J}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \mathsf{J}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \mathbf{0} \tag{38}$$

3.1 Problem formulation

Lets consider the following parameter estimation (inverse) problem in heat transfer

$$\begin{aligned} \min_{\mathbf{z}\in\hat{\mathbb{Z}}} \quad & \frac{\beta}{2}\|\langle \mathbf{u}, \mathbf{A}(\mathbf{z})\mathbf{u}\rangle_{\mathbb{H}} - \langle \hat{\mathbf{u}}, \mathbf{A}(\mathbf{z})\hat{\mathbf{u}}\rangle_{\mathbb{H}}\|^2_{\mathbb{H}} + \mathsf{R}(\mathbf{z}) \\ \text{s.t.} \quad & \\ & \mathbf{A}(\mathbf{z})\mathbf{u} = \mathbf{f}(\mathbf{x}) \quad \text{in} \quad \Omega \\ & \mathbf{u} = 0 \qquad \text{on} \quad \partial\Omega, \end{aligned} \tag{39}$$

where $\hat{\mathbb{Z}} = \{\mathbf{z} \in \mathbb{Z} : \mathscr{L} \leq \mathbf{z} \leq \mathscr{U}\}$. Here, $\mathscr{L}$ denotes the control lower bounds and $\mathscr{U}$ denotes the control upper bounds. For a steady-state heat equation, $\mathbf{z}$ is the coefficient of thermal conductivity, $\mathbf{u}$ is the temperature field, $\hat{\mathbf{u}} \in \Omega^m \subseteq \Omega$ are the temperature measurements. $\mathbf{f}(\mathbf{x})$ is a heat source given by

$$\mathbf{f}(\mathbf{x}) = A * \sin(\omega\mathbf{x})\cos(\omega\mathbf{x}), \tag{40}$$

where $A \in \mathbb{R}$ is a given amplitude, $\omega \in \mathbb{R}$ denotes angular frequency and $\mathbf{x} \in \Omega$ denotes a position is space. $\mathbf{A}(\mathbf{z})\colon \mathbb{Z} \to \mathbb{U} \times \mathbb{U}$ is a linear operator that depends on the coefficient of thermal conductivity and $\beta \in \mathbb{R}_+$ denotes a penalty parameter. The finite dimensional approximations for the state, control, and Lagrange multipliers were previously defined in Section 2.1.

3.1.1 Regularization

The regularization functional $\mathsf{R}(\mathbf{z})\colon \mathbb{Z} \to \mathbb{R}$ in Equation 39 is given by

$$\mathsf{R}(\mathbf{z}) = \frac{\zeta}{2\theta}(\langle \nabla\mathbf{z}, \nabla\mathbf{z}\rangle_{\mathbb{H}} + \nu)^{\theta}, \tag{41}$$

where $\theta = -1/2$, $0 < \nu \leq 1$ and $0 < \zeta \leq 1$. In this work, the regularization functional in Equation 41 was preferred over Tikhonov regularization due to its ability to capture sharp discontinuities in inverse problems settings. The interested reader is referred to [29] and explore other regularization methodologies that can be applied for inverse problems.

3.1.2 Helmholtz filter

The aim of this section is to present the Helmholtz filter as an alternative to regularization methods, e.g. Equation 41. Regularization functionals are often (if not always) explicitly incorporated into the objective function to solve ill-posed inverse problems. The purpose of regularization functionals is to penalize the objective function and enhance the smoothness of the control field. Thus, bounding the objective function and preventing undesired data overfitting.

Instead of just applying a regularization functional, e.g. Equation 41, to solve the inverse problem in Equation 39, a Helmholtz PDE filter is also employed to filter the optimal control computed using the CEM formulation. Helmholtz PDE filters have been successfully used in topology optimization to avoid numerical artifacts and enhance the design's smoothness and boundary description [30]. The filtered control is computed by applying a convolution operator to the optimal control. However, instead of explicitly defining the convolution integral, the filtered control can be defined implicitly as the solution of the following Helmholtz PDE

$$-r^2\nabla^2\tilde{\mathbf{z}} + \tilde{\mathbf{z}} = \mathbf{z}, \tag{42}$$

with Neumann boundary conditions

$$\frac{\partial \tilde{\mathbf{z}}}{\partial \mathbf{n}} = 0. \tag{43}$$

The parameter r denotes a given filter length scale and $\tilde{\mathbf{z}}$ is the filtered control, which was set to $\frac{1}{3\times 10^3}$ for this study. As $\mathbf{x}/r \to \infty$, the filtering effect on the optimal control is reduced. Thus, larger values of r minimize the filtering effect on the optimal control.

In this work, the discretized Helmholtz PDE is solved after optimization. Contrary, we could have eliminated the regularization functional from Equation 39 and solved the discretized Helmholtz PDE every time a new set of trial controls was computed by the optimization algorithm. However, this would have increased the computational demands of the problem. Furthermore, preliminary results performed as part of this study did not justify this approach. The quality of the solution obtained by applying the filter during optimization did not improve greatly compared to the quality obtained by applying the filter after optimization.

3.2 Optimality conditions

The Lagrangian functional $\mathcal{L}\colon \mathbb{U} \times \hat{\mathbb{Z}} \times \mathbb{Y} \to \mathbb{R}$ for the inverse problem defined in Equation 39 is given by

$$\mathcal{L}(\mathbf{u}, \mathbf{z}, \mathbf{v}) = \frac{\beta}{2} \| \langle \mathbf{u}, \mathbf{A}(\mathbf{z})\mathbf{u} \rangle_{\mathbb{H}} - \langle \hat{\mathbf{u}}, \mathbf{A}(\mathbf{z})\hat{\mathbf{u}} \rangle_{\mathbb{H}} \|^2_{\mathbb{H}} + \mathsf{R}(\mathbf{z}) + \langle \mathbf{v}, \mathbf{A}(\mathbf{z})\mathbf{u} - \mathbf{f} \rangle_{\mathbb{Y}^*, \mathbb{Y}}. \tag{44}$$

The first order necessary optimality conditions for Equation 44 are given by

$$\mathcal{L}_{\mathbf{u}}(\mathbf{u}, \mathbf{z}, \mathbf{v}) = \alpha\beta \mathbf{A}(\mathbf{z})\mathbf{u} + \mathbf{A}(\mathbf{z})\mathbf{v} = 0 \tag{45}$$

$$\mathcal{L}_{\mathbf{z}}(\mathbf{u}, \mathbf{z}, \mathbf{v}) = \alpha\beta[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\mathbf{u} + (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] + \mathsf{R}_{\mathbf{z}}(\mathbf{z}) + (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\mathbf{v} = 0, \tag{46}$$

where α is given by Equation 13. The first order derivative operator $R_{\mathbf{z}}(\mathbf{z})\colon \hat{\mathbb{Z}} \to \mathbb{Z}$ is given by

$$R_{\mathbf{z}}(\mathbf{z}) = \frac{\zeta}{2}(\theta - 1)(\langle \nabla \mathbf{z}, \nabla \mathbf{z} \rangle_{\mathbb{H}} + \nu^2)^{\theta - 1} \mathbf{B}\, \mathbf{z}, \tag{47}$$

where

$$\mathbf{B} = \int_{\Omega} \nabla\psi \nabla\psi \, d\Omega. \tag{48}$$

If $\mathbf{A}(\mathbf{z})$ is a non-singular, self-adjoint linear operator, the Lagrange multipliers for an inverse problem in heat transfer are given by Equation 14. Substituting Equation 14 into Equation 46 yields the following gradient operator

$$\nabla \mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z}) = -\alpha\beta[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\mathbf{u} + (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] + \mathsf{R}_{\mathbf{z}}(\mathbf{z}). \tag{49}$$

Notice, as previously demonstrated in Section 2, that the computation of Equation 49 does not require the solution of the adjoint system of equations.

If second order optimization algorithms are available, Newton's method can be applied to the first order necessary optimality conditions. The following derivative operators are then required to compute the application of the trial step to the analytical Hessian operator

$$\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}), \mathbf{z})\delta\mathbf{z} = (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\delta\mathbf{z})\mathbf{u} \tag{50}$$

$$\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} = \mathbf{A}(\mathbf{z})\delta\mathbf{u} \tag{51}$$

$$\mathsf{g}_{\mathbf{z}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{z} = (\mathbf{A}_{\mathbf{z}}(\mathbf{z})^*\mathbf{v})\mathbf{u} \tag{52}$$

$$\mathsf{g}_{\mathbf{u}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{u} = \mathbf{A}(\mathbf{z})^*\mathbf{v} \tag{53}$$

$$\mathsf{g}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{u} = \mathbf{0} \tag{54}$$

$$\mathsf{g}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{z} = (\mathbf{A}_{\mathbf{z}}(\mathbf{z})^*\delta\mathbf{z})\mathbf{v} \tag{55}$$

$$\mathsf{g}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{u} = \mathbf{0} \tag{56}$$

$$\mathsf{g}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})^*\delta\mathbf{u} = (\mathbf{A}_{\mathbf{z}}(\mathbf{z})^*\mathbf{v})\delta\mathbf{u} \tag{57}$$

$$\mathsf{J}_{\mathbf{uu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} = 2\beta[\alpha\mathbf{A}(\mathbf{z})\delta\mathbf{u} + 2\langle\mathbf{u},\mathbf{A}(\mathbf{z})\delta\mathbf{u}\rangle_{\mathbb{H}}\mathbf{A}(\mathbf{z})\mathbf{u}] \tag{58}$$

$$\mathsf{J}_{\mathbf{uz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = 2\beta[\alpha(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\delta\mathbf{z})\mathbf{u} + \gamma\mathbf{A}(\mathbf{z})\mathbf{u}] \tag{59}$$

$$\mathsf{J}_{\mathbf{zz}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \gamma\beta[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\mathbf{u} - (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] + \mathsf{R}_{\mathbf{zz}}(\mathbf{z})\delta\mathbf{z} \tag{60}$$

$$\mathsf{J}_{\mathbf{zu}}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{u} = 2\beta[\alpha(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\delta\mathbf{u} + ((\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\mathbf{u} - (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}})\langle\mathbf{u},\mathbf{A}(\mathbf{z})\delta\mathbf{u}\rangle_{\mathbb{H}}], \tag{61}$$

where γ is given by Equation 28 and the Lagrange multipliers $\mathbf{v}$ are given by Equation 14. Finally, the second order derivative operator $\mathsf{R}_{\mathbf{zz}}(\mathbf{z})$ is given by

$$\mathsf{R}_{\mathbf{zz}}(\mathbf{z}) = \frac{\zeta}{2}((\theta-1)(\langle\nabla\mathbf{z},\nabla\mathbf{z}\rangle_{\mathbb{H}} + \nu^2)^{\theta-2})\langle\nabla\mathbf{z},\nabla\mathbf{z}\rangle_{\mathbb{H}}\mathbf{B} + (\langle\nabla\mathbf{z},\nabla\mathbf{z}\rangle_{\mathbb{H}} + \nu^2)^{\theta-1}\mathbf{B}. \tag{62}$$

Substituting Equations 50-61 into Equation 24 yields the application of the trial step to the analytical Hessian, as defined in Equation 33. However, if the Gauss-Newton Hessian is applied, the application of the trial step to the Gauss-Newton Hessian operator is then given by

$$\nabla^2\mathsf{J}(\mathbf{u}(\mathbf{z}),\mathbf{z})\delta\mathbf{z} = \gamma\beta[(\mathbf{A}_{\mathbf{z}}(\mathbf{z})\mathbf{u})\mathbf{u} - (\mathbf{A}_{\mathbf{z}}(\mathbf{z})\hat{\mathbf{u}})\hat{\mathbf{u}}] + \mathsf{R}_{\mathbf{zz}}(\mathbf{z})\delta\mathbf{z}, \tag{63}$$

where $\mathsf{R}_{\mathbf{zz}}(\mathbf{z})$ is given by Equation 62.

Finally, the discretized Helmholtz equation is given by

$$r\mathbf{B}\hat{\mathbf{z}} + \mathbf{M}\hat{\mathbf{z}} = \mathbf{z}, \tag{64}$$

where

$$\mathbf{M} = \int_{\Omega} \psi\psi \, d\Omega. \tag{65}$$

and $\mathbf{B}$ is defined by Equation 48.

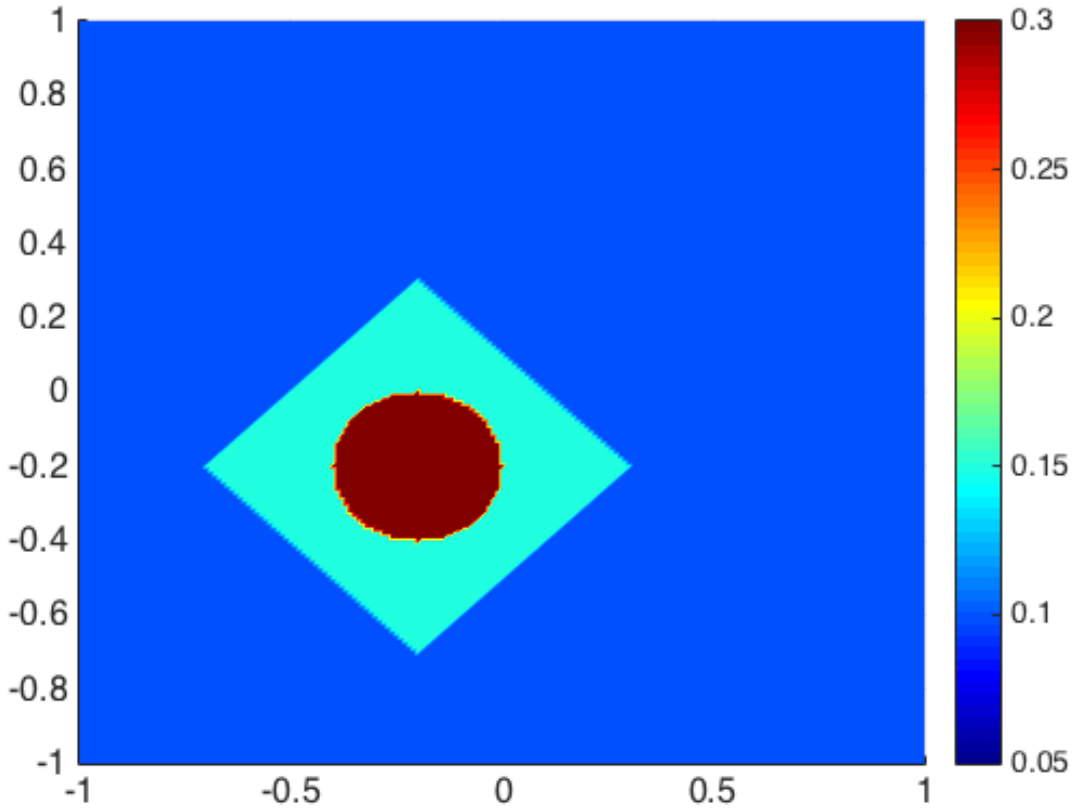


Figure 1: Target thermal conductivity field.

3.3 Results: full-field temperature field

The compliance error minimization and the data misfit formulations were applied to the inverse problem defined in Equation 39. The amplitude for the heat source was set to 1×10^2 and the angular frequency was set to 4π and 16π for the numerical studies performed with the data misfit and compliance error functionals, respectively. The target thermal conductivity field is shown in Figure 1. The thermal conductivity coefficients lower and upper bounds were respectively set to 0.01 and 1.0 during optimization. The regularization parameter ζ was set to 1.0 for all the numerical studies performed herein. To quantify the computational efficacy of the compliance error functional, the corresponding speedups are computed, where

$$S = \frac{CPU_{DMF}}{CPU_{CEM}}. \tag{66}$$

In Equation 66, CPU denotes central processing unit time, DMF denotes data misfit functional and CEM denotes compliance error minimization.

3.3.1 First order formulation

The Perry-Shanno nonlinear conjugate gradient was used to solve the inverse problem in heat transfer [31, 32]. The optimization algorithm stopped when $\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z})_k$, $\|\nabla_{\mathbf{z}}\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z})_k\|$ or $\|\mathbf{s}_k\|$ was below a predefined tolerance of 1×10^{-4}. A backtracking line search with cubic

Noise	DMF	CEM	S
0%	132.7	1.08	122.87
1%	132.25	1.15	119.35
3%	103.18	1.07	96.43
5%	63.2	1.19	53.11

Table 1: CPU time (seconds) and corresponding speedups obtained using the Perry-Shanno nonlinear conjugate gradient algorithm.

step interpolation was applied to enhance the global convergence capability of the nonlinear conjugate gradient algorithm. The line search contraction parameter was set to 0.5 and the step lower bound was set to 1×10^{-5}. Furthermore, the maximum number of line search iterations was set to 5.

The regularization parameters used for the numerical studies based on the data misfit functional were $(\zeta, \nu) = (5 \times 10^{-4}, 1 \times 10^{-4})$ for $\theta = \{0\%, 1\%\}$ and $(\zeta, \nu) = (5 \times 10^{-3}, 1 \times 10^{-4})$ for $\theta = \{3\%, 5\%\}$. Contrary, the regularization parameters used for all the numerical studies based on the CEM formulation were $(\zeta, \nu) = (1 \times 10^{-8}, 1 \times 10^{-8})$.

Table 1 shows the central processing unit times obtained for the numerical studies that are based on first order derivative information. Regardless of the noise level, the compliance error formulation produced noticeable speedups over the data misfit formulation. Clearly, the CEM formulation outperforms the data misfit formulation. Figure 2 shows that the CEM strategy required less than 10 optimization iterations to converged to an optimal and feasible solution in all the numerical studies. Contrary, the data misfit formulation strategy needed over 1000 iterations to meet the convergence criteria. Figure 2 also shows the objective function values produced by Perry-Shanno nonlinear conjugate gradient algorithm. The reader can clearly appreciate that the compliance error functional produced faster convergence rates than the data misfit functional, regardless of the level of corruption in the data.

Figure 3 shows the optimal thermal conductivity field computed using the data misfit functional and the Perry-Shanno nonlinear conjugate gradient algorithm. The thermal conductivity field was accurately approximated for all noise levels. However, the solution obtained with the data misfit functional was sensitive to higher noise levels. Figure 4 shows the optimal thermal conductivity field computed using the compliance error functional and the Perry-Shanno nonlinear conjugate gradient algorithm. The compliance error minimization formulation yielded accurate thermal conductivity fields regardless of the noise levels. However, the CEM strategy produced non-optimal thermal conductivity fields were the Dirichlet and Neumann boundary conditions were applied. More research is needed to understand why these inaccuracies are obtained around areas were Dirichlet or Neumann boundary conditions are applied.

3.3.2 Second order formulation

A dogleg trust region inexact Newton algorithm [33] was used to solve the inverse problem defined in Equation 39 when second order derivative information was available. The optimization algorithm once more stopped when one of the following stopping criterion was satisfied: $\mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z})_k < 1 \times 10^{-4}$, $\|\nabla_{\mathbf{z}} \mathsf{J}(\mathbf{u}(\mathbf{z}), \mathbf{z})_k\| < 1 \times 10^{-4}$ or $\|\mathbf{s}_k\| < 1 \times 10^{-4}$. The trust region contraction and expansion parameters were respectively set to 0.5 and 2. The maximum number of trust region sub-problem iterations was set to 5 and the minimum ratio between the actual

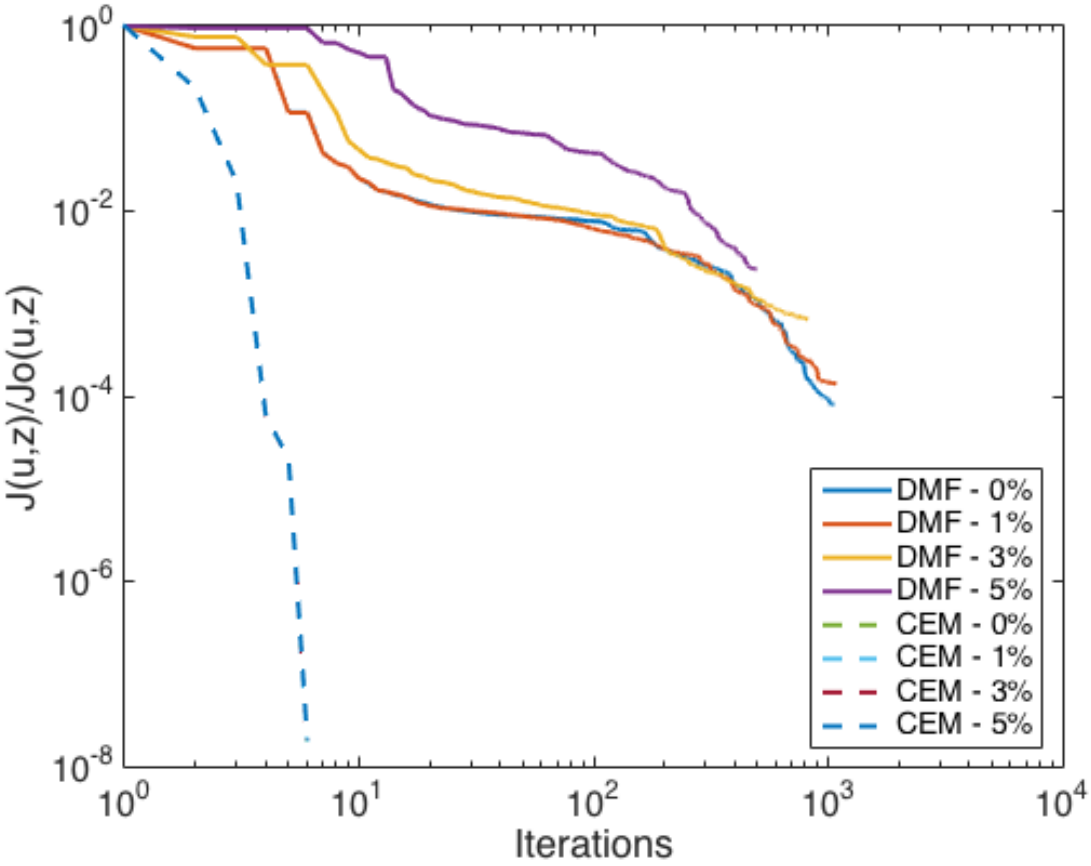
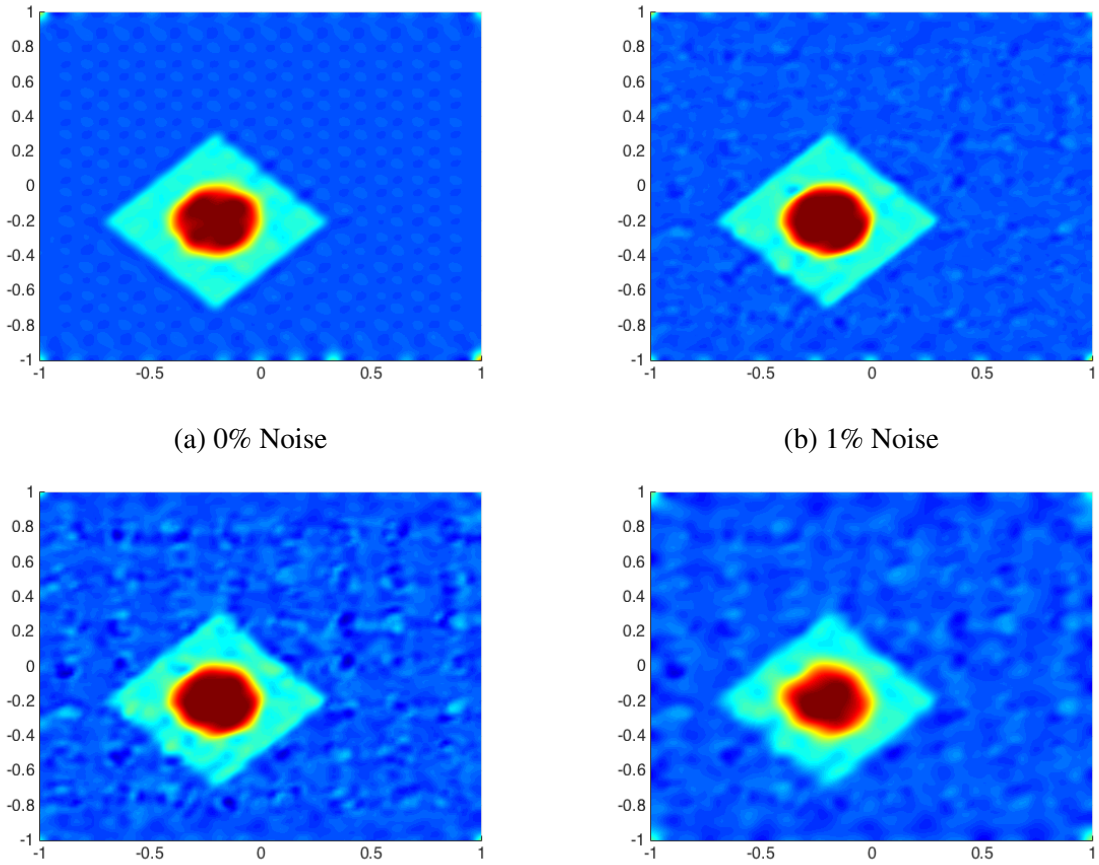


Figure 2: Objective function values computed by the Perry-Shanno nonlinear conjugate gradient algorithm.



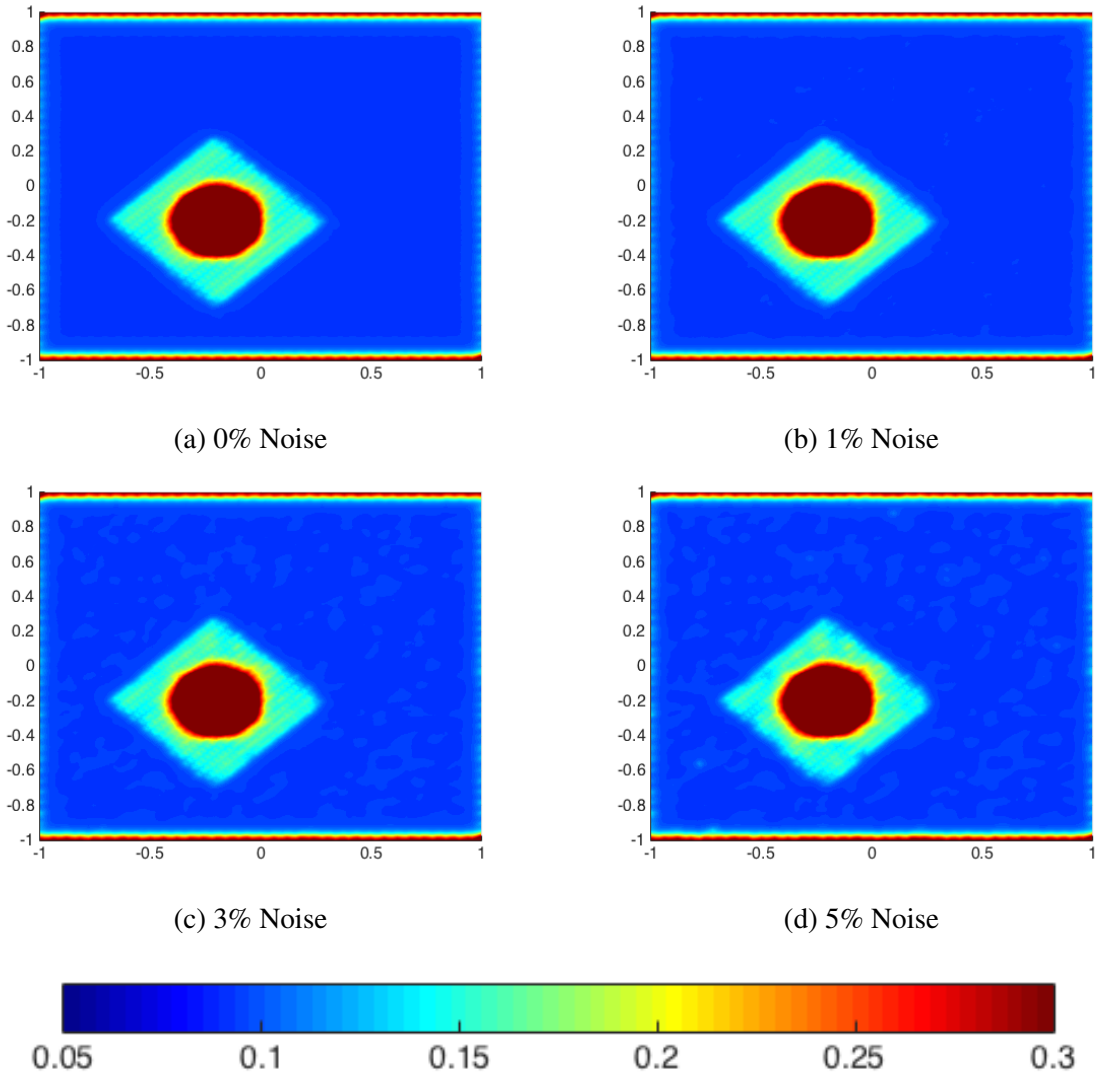
(a) 0% Noise

(b) 1% Noise

(c) 3% Noise

(d) 5% Noise

Figure 3: Optimal thermal conductivity field computed using the first order data misfit formulation strategy and the Perry-Shanno nonlinear conjugate gradient algorithm.



(a) 0% Noise

(b) 1% Noise

(c) 3% Noise

(d) 5% Noise

Figure 4: Optimal thermal conductivity field computed using the first order compliance error minimization formulation and the Perry-Shanno nonlinear conjugate gradient algorithm.

Noise	DMF	CEM-AH	CEM-GN	S-AH	S-GN
0%	123.64	8.18	5.37	15.11	23.02
1%	192.42	7.98	4.61	24.11	41.74
3%	211.29	8.06	6.07	26.21	34.81
5%	91.76	8.08	5.06	11.36	18.13

Table 2: CPU time (seconds) and corresponding speedups obtained with the dogleg trust region inexact Newton algorithm. Here, AH and GN respectively denote analytical and Gauss-Newton Hessians.

and predicted reduction was set to 0.2.

The regularization parameters used for the numerical studies based on the data misfit functional were $(\zeta, \nu) = (1\times10^{-3}, 1\times10^{-3})$ for $\theta = \{0\%, 1\%, 3\%\}$ and $(\zeta, \nu) = (5\times10^{-3}, 1\times10^{-3})$ for $\theta = 5\%$. Once more, the regularization parameters used for all the numerical studies based on the compliance error functional were $(\zeta, \nu) = (1 \times 10^{-8}, 1 \times 10^{-8})$.

Table 2 shows the CPU times gathered for the numerical studies done with the dogleg trust region inexact Newton algorithm. The results on Table 2 once more show that the compliance error minimization formulation produced significant speedups over the data misfit formulation. However, the CPU times obtained with the CEM formulation and the second order optimization algorithm were higher than those obtained with the first order optimization algorithm. Therefore, the second order derivative information did not enable additional speedups. However, these results are specific to the parameter estimation problem in heat transfer and more studies are necessary to further understand the possible benefits of the second order compliance error formulation. Lets recall that an effective preconditioning strategy could be applied to improve the performance of the Newton algorithm and reduce the number of Newton iterations. However, this was outside the scope of this study.

Figure 5 displays the objective function values computed using the dogleg trust region inexact Newton algorithm. Once more, the compliance error minimization formulations required less than 10 optimization iterations to converged to an optimal solution. Contrary, the data misfit formulation required over 50 optimization iterations (in some cases over 100 iterations) to meet one of the convergence criterion. However, why the CPU times produced by the second order optimization algorithm using the data misfit functional are greater than those produced by the first order optimization algorithm? These results can be counterintuitive to the reader since the numerical studies based on the first order data misfit formulation needed more iterations to converge to an optimal solution. The reader should recall that second order optimization algorithms require at least 4 model evaluations (2 for the gradient and 2 for the Hessian calculation) per optimization iteration. Furthermore, every Newton iteration requires 2 model evaluations to compute trial descent direction. Hence, regardless of the fact that the trust region Newton algorithm needed less optimization iterations to converge, the number of model evaluations will always dominate CPU time. This further motivates future research to enable an effective preconditioning strategy for the compliance error minimization formulation.

The same argument can be applied to the second order CEM formulation based on the full Hessian since at least 2 model evaluations (1 for the gradient and 1 for the Hessian calculation) are performed per iteration to compute a new control field. Hence, the higher CPU times seen on Table 2. Next, lets consider the CEM formulation based on the Gauss-Newton Hessian. This formulation only needs one model evaluation per optimization iteration. Thus, why the CPU

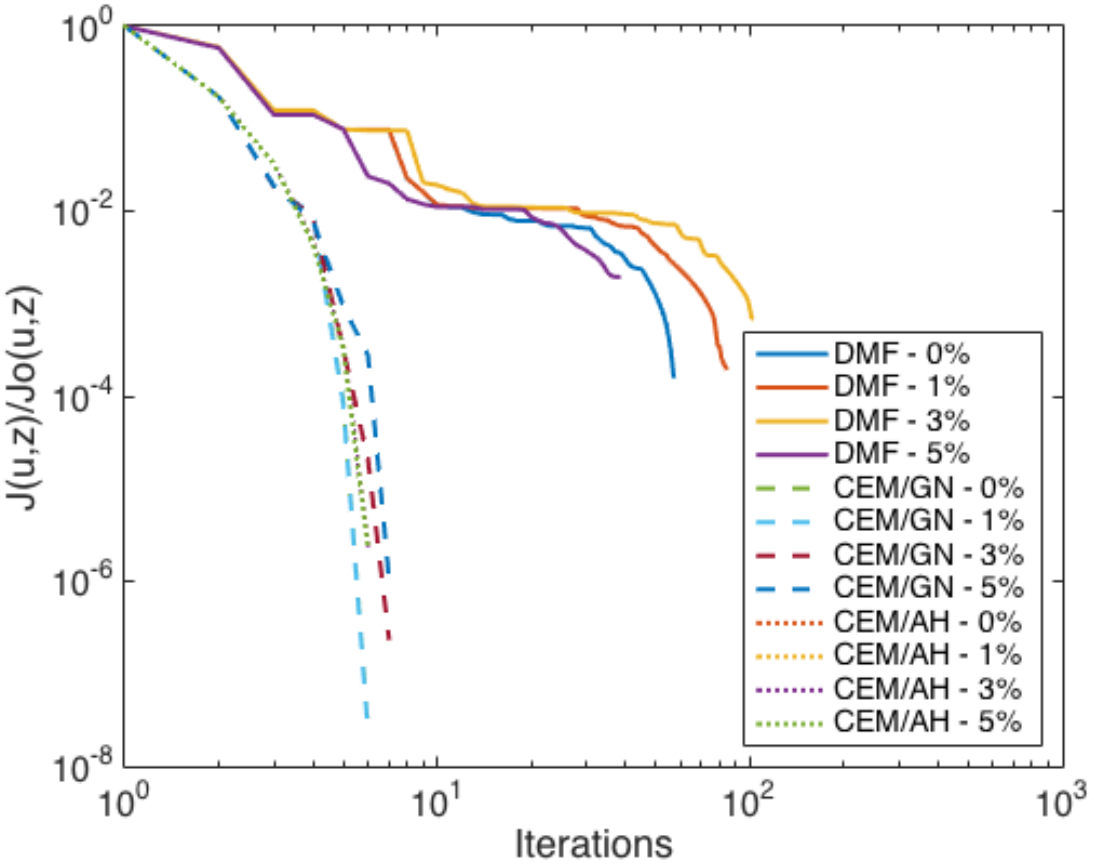
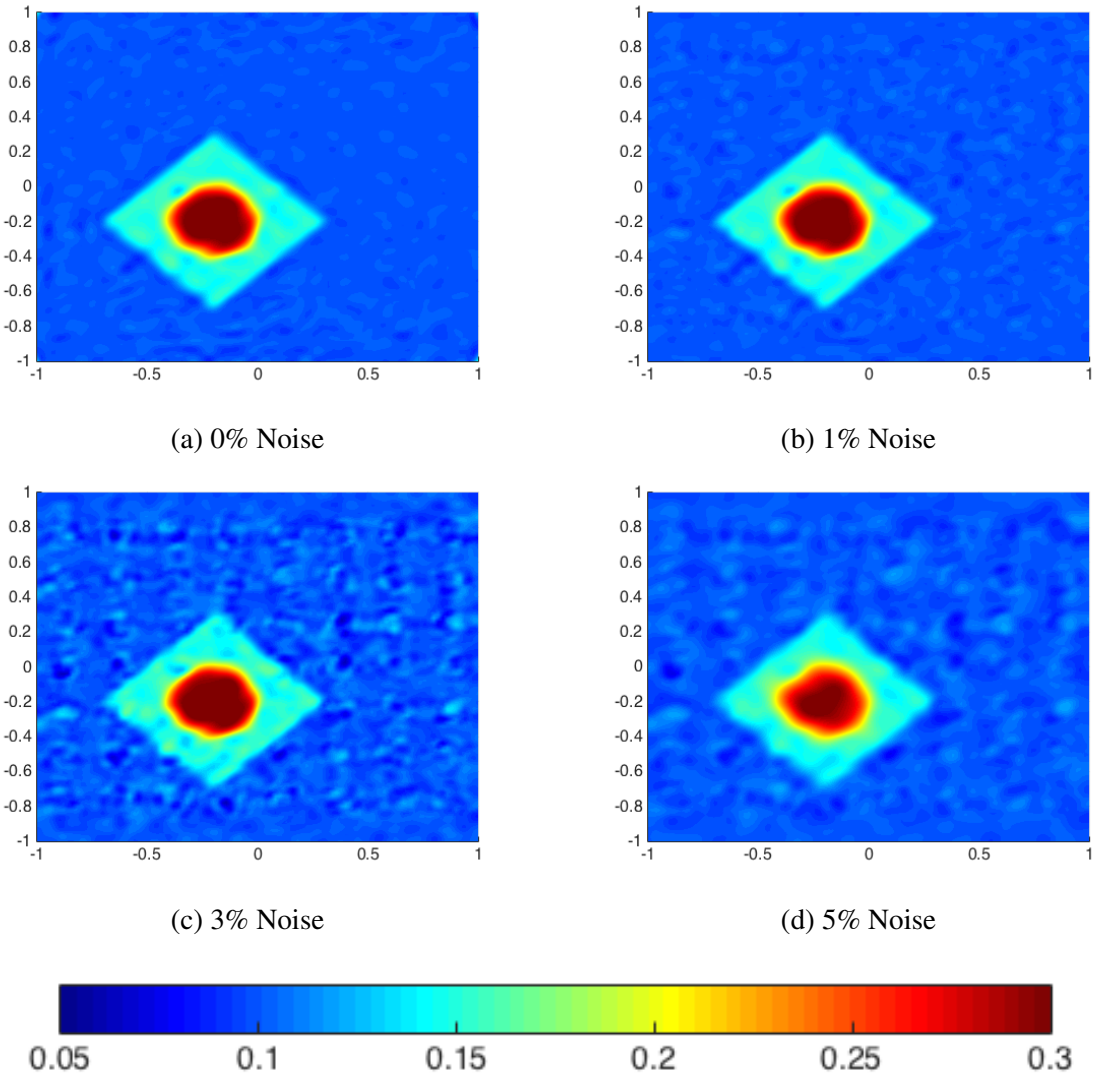


Figure 5: Objective function values computed using the dogleg trust region inexact Newton algorithm.

times are not closer/similar to the CPU times observed for the first order CEM formulation? In all our numerical studies, the optimization algorithm required several trust region sub-problem iterations to compute a trial control that met the predefined ratio between the actual and the predicted reduction. Contrary, the first order CEM formulation did not need additional line search iterations to compute a descent direction that produced an optimal/feasible trial control. Therefore, less model evaluations were performed during optimization; thus, reducing computational demands.

Figure 5 also shows the convergence rates for the second order formulations. Notice that the CEM formulations displayed faster convergence rates than the second order data misfit formulation, similar to the results shown in Section 3.3.1. However, the convergence rates for the second order data misfit formulation displayed faster convergence rates near the optimal/feasible point. This was expected since Newton algorithms are expected to converge quadratically near the optimal/feasible point.

Finally, Figure 6 shows the optimal thermal conductivity field computed using the second order data misfit formulation for all noise levels. Results demonstrate that the second order data misfit formulation produced optimal thermal conductivity fields. However, once more the results show that the optimal solutions were sensitive to corruption in the data. Figures 7 and 8 show the optimal thermal conductivity field computed using the second order CEM formulations based on the analytical and Gauss-Newton Hessians, respectively. Both approaches



(a) 0% Noise

(b) 1% Noise

(c) 3% Noise

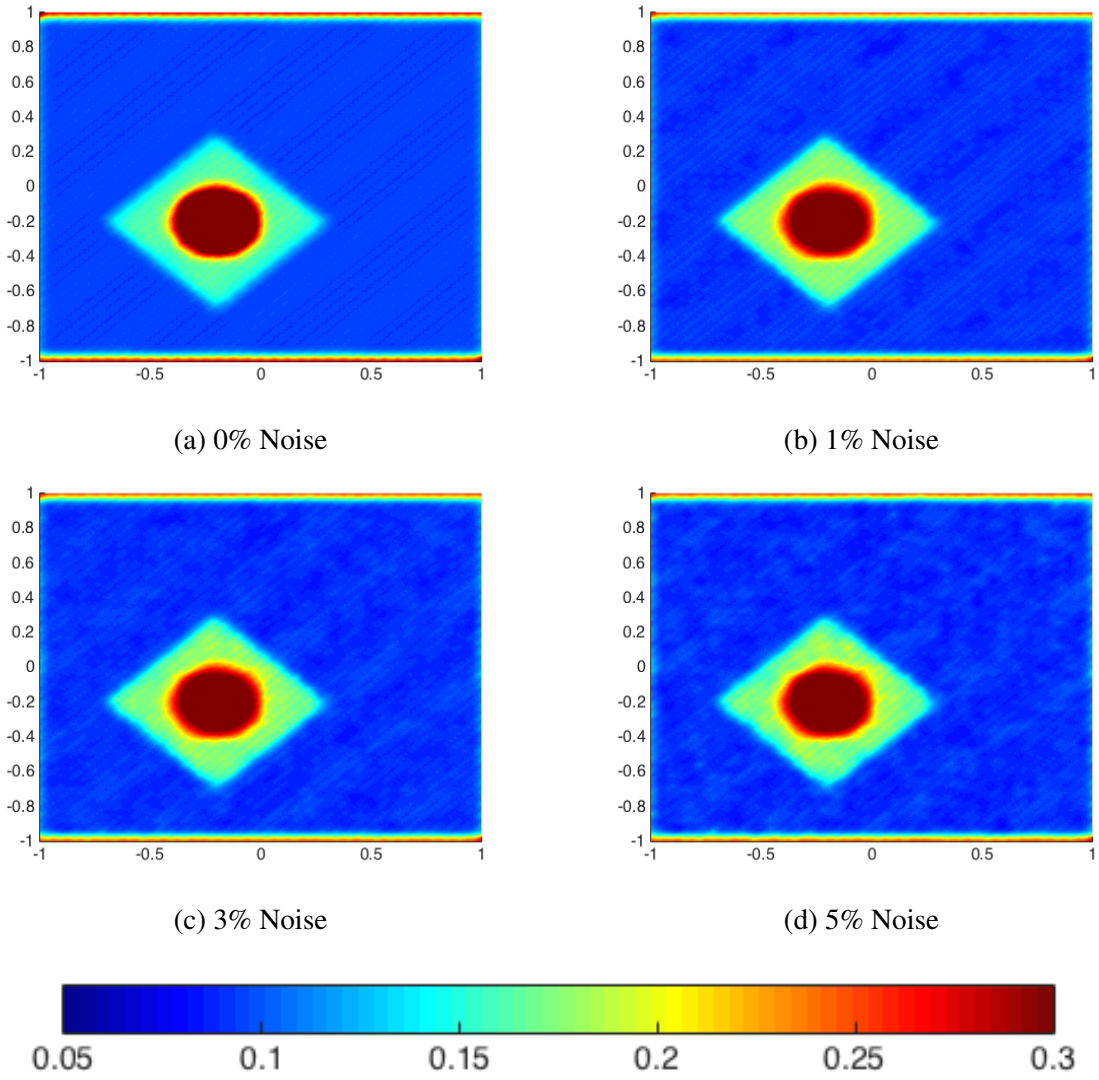
(d) 5% Noise

Figure 6: Optimal thermal conductivity field computed using the second order data misfit formulation and the inexact dogleg trust region Newton algorithm.

accurately characterized the location and the shape of the thermal conductivity field of interest. However, the magnitude of the thermal conductivity field inside the heterogeneous conductivity field was overestimated. Furthermore, the compliance error minimization formulation once more produced non-optimal thermal conductivity values around the regions were Dirichlet and Neumann boundary conditions were applied. These results further highlight the need for more research to eliminate the numerical artifacts computed around these regions. However, the significant speedups observed with the compliance error minimization formulations motivate future research to continue improving the proposed inverse problem formulation.

4 CONCLUSIONS

This paper presented a new formulation for inverse problems based on the compliance error functional. The compliance error functional enabled the computation of the Lagrange multipliers at the expense of just one model evaluation. Thus, the calculation of the Lagrange multipliers does not require the solution of the computationally intensive adjoint problem. This, leads to



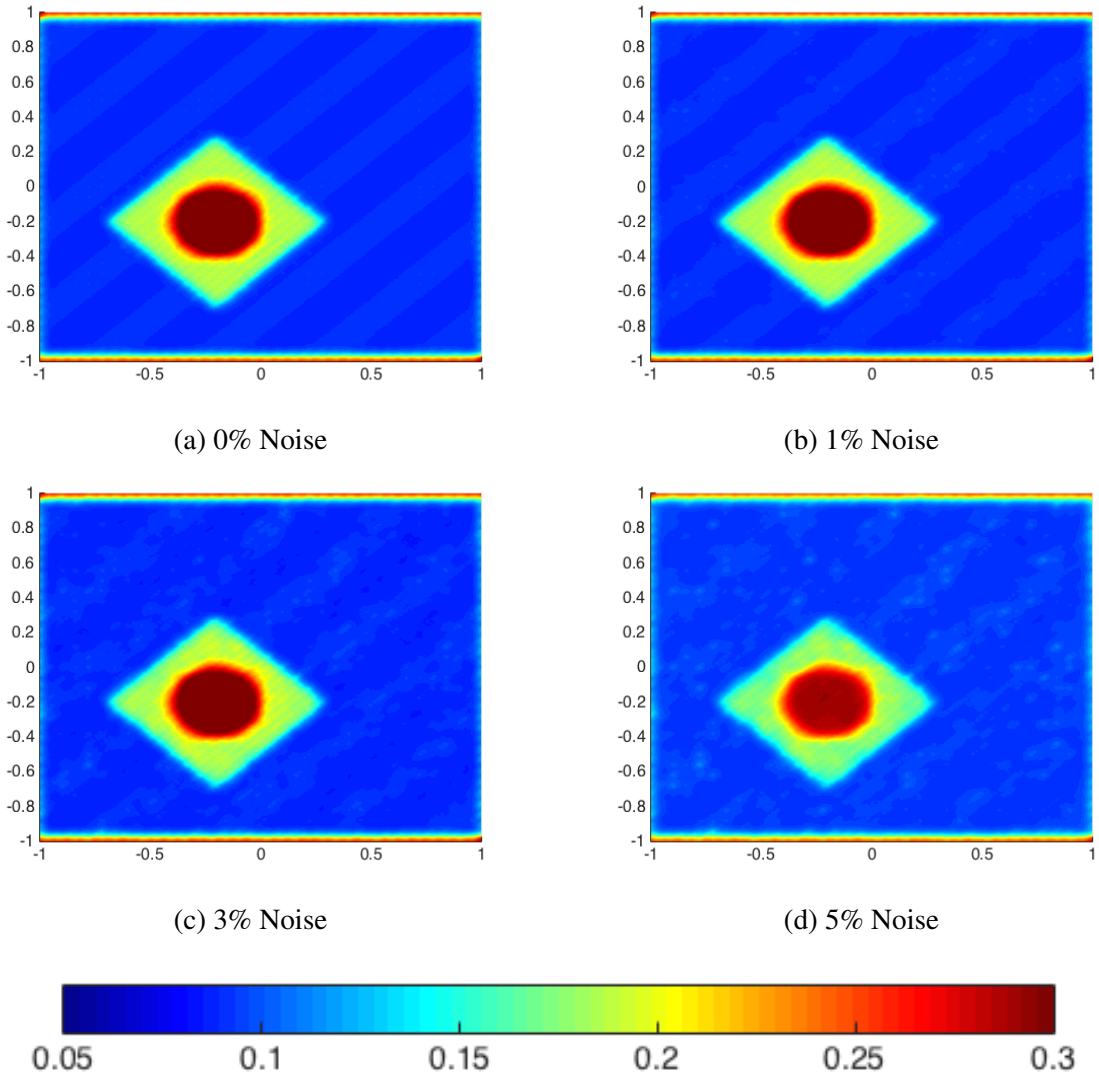
(a) 0% Noise

(b) 1% Noise

(c) 3% Noise

(d) 5% Noise

Figure 7: Optimal thermal conductivity field computed using the second order compliance error minimization formulation based on the analytical Hessian and the inexact dogleg trust region Newton algorithm.



(a) 0% Noise

(b) 1% Noise

(c) 3% Noise

(d) 5% Noise

Figure 8: Optimal thermal conductivity field computed using the second order compliance error minimization formulation based on the Gauss-Newton Hessian and the inexact dogleg trust region Newton algorithm.

significant speedups since the gradient operator can be computed at the expense of one model evaluation per optimization iteration. Furthermore, the implementation of the CEM formulation in production finite element software is greatly simplified since the propose formulation only requires forward model evaluations to compute the gradient. Therefore, the implementation of the adjoint system of equations is not necessary to solve the inverse problem. Likewise, computing the application of the trial step to the analytical Hessian operator is greatly simplified since only one additional model evaluation is requires. This also leads to significant speedups when second order optimization algorithms and the CEM formulation are applied to solve the inverse problem.

This paper also presented a Gauss-Newton Hessian formulation based on the compliance error functional that effectively approximates the analytical Hessian during optimization. By applying the Gauss-Newton Hessian, optimal results were recovered, while reducing the computational cost associated with evaluating the analytical Hessian. Furthermore, the Gauss-Newton Hessian approximation further reduced the computational demands of the inverse problem. However, results suggested that the first order CEM formulation was computationally more effective than the second order CEM formulation. Therefore, the second order CEM formulations did not provide additional computational benefits over the first order CEM formulation. However, more research is needed since the second order CEM formulations can be combined with an effective preconditioning strategy to improve accuracy and computational efficacy.

Results also showed that the data misfit formulation produced better results around the regions were Dirichlet and Neumann boundary conditions were applied. The compliance error functional produce perceivable numerical artifacts near these regions. However, the interior thermal conductivity fields computed with the first and second order CEM formulations were accurate. It was also observed that the second order CEM formulations were inclined to overestimate the thermal conductivity field around the heterogeneous thermal conductivity region. Future research will focus on exploring alternate inverse problem formulations to improve the CEM strategy, without losing its computational advantages. Finally, it was observed in this study that compliance error minimization formulation was less sensitive to corruption in the experimental data.

Overall, the results obtained with the compliance error minimization formulation are encouraging due to the significant speedups produced during optimization. Furthermore, the first and second order CEM formulation yield accurate thermal conductivity fields. Although their is room for improvements, the gains in speed are substantial and thus motivate future research to continue improving the efficacy of the proposed inverse problem formulation.

REFERENCES

[1] M. Aguiló, W. Aquino, J. Brigham, and M. Fatemi, "An inverse problem approach for elasticity imaging through vibroacoustics," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 4, pp. 1012–1021, 2010.

[2] M. Gockenbach, B. Jadamba, and A. Khan, "Numerical estimation of discontinuous coefficients by the method of equation error," *Mathematics and Mechanics of Solids*, vol. 1, pp. 343–359, 2006.

[3] H. L. Liew and P. M. Pinsky, "Recovery of shear modulus in elastography using an adjoint method with b-spline representation," *Finite Elements in Analysis and Design*, vol. 41, no. 7-8, pp. 778–799, 2005.

[4] R. Sampath and N. Zabaras, "An object-oriented framework for the implementation of adjoint techniques in the design and control of complex continuum systems," *International Journal for Numerical Methods in Engineering*, vol. 48, no. 2, pp. 239–266, 2000.

[5] A. Deraemaeker, P. Ladevèze, and P. Leconte, "Reduced bases for model updating in structural dynamics based on the constitutive relation error," *Computer Method in Applied Mechanics and Engineering*, vol. 191, no. 21-22, pp. 2427–2444, 2002.

[6] P. Feissel and O. Allix, "Modified constitutive relation error identification strategy for transient dynamics with corrupted data: The elastic case," *Computer Method in Applied Mechanics and Engineering*, vol. 196, no. 13-16, pp. 1968–1983, 2007.

[7] R. V. Kohn and B. D. Lowe, "A variational method for parameter identification," *Mathematical Modeling and Numerical Analysis*, vol. 22, no. 1, pp. 119–158, 1998.

[8] R. Kohn and A. McKenney, "Numerical implementation of a variational method for electrical impedance tomography," *Inverse Problems*, vol. 6, no. 3, pp. 389–414, 1990.

[9] R. Kohn and M. Vogelius, "Determining conductivity by boundary measurements," *Communications on Pure and Applied Mathematics*, vol. 37, no. 3, pp. 289–298, 1984.

[10] R. Kohn and M. Vogelius, "Determining conductivity by boundary measurements II. interior results," *Communications on Pure and Applied Mathematics*, vol. 38, no. 5, pp. 643–667, 1985.

[11] P. Ladevèze and C. A., "Updating structural dynamic models with emphasis on the damping properties," *American Institute of Aeronautics and Astronautics Journal*, vol. 36, no. 6, pp. 1094–1099, 1998.

[12] P. Ladevèze and D. Leguillon, "Error estimates procedures in the finite element method and applications," *SIAM Journal on Numerical Analysis*, vol. 20, no. 3, pp. 485–509, 1983.

[13] O. Allix, P. Feissel, and H. Nguyen, "Identification strategy in the presence of corrupted measurements," *Engineering computations*, vol. 22, no. 5-6, pp. 487–504, 2005.

[14] M. S. Gockenbach and A. A. Khan, *Mathematical Models and Methods for Real World Systems*, ch. A convex objective functional for elliptic inverse problems. Chapman & Hall CRC Taylor & Francis Group, 2006.

[15] M. S. Gockenbach and A. A. Khan, "An abstract framework for elliptic inverse problems: Part 1. An output least-squares approach," *Mathematics and Mechanics of Solids*, vol. 12, no. 3, pp. 259–276, 2007.

[16] M. S. Gockenbach and A. A. Khan, "An abstract framework for elliptic inverse problems: Part 2. An augmented Lagrangian approach," *Mathematics and Mechanics of Solids*, vol. 14, no. 6, pp. 517–539, 2009.

[17] M. Gockenbach, "Numerical analysis of elliptic inverse problems with interior data," *Journal of Physics: Conference Series*, vol. 124, no. 1, pp. 1–12, 2008.

[18] F. Pierron and M. Grédiac, *The Virtual Fields Method: Extracting Constitutive Mechanical Parameters from Full-field Deformation Measurements*. Springer New York, 2012.

[19] M. Grdiac, F. Pierron, S. Avril, and E. Toussaint, "The virtual fields method for extracting constitutive parameters from full-field measurements: a review," *Strain*, vol. 42, no. 4, pp. 233–253, 2006.

[20] M. Aguiló, L. Swiler, and A. Urbina, "First-order formulations for large-scale stochastic parameter estimation within the frameworks of steady state dynamics: the elastic and viscoelastic case," *Inverse Problems*, vol. 28, no. 7, p. 075003, 2012.

[21] M. Heroux, R. Bartlett, V. Howle, R. Hoekstra, J. Hu, T. Kolda, R. Lehoucq, K. Long, R. Pawlowski, E. Phipps, A. Salinger, H. Thornquist, R. Tuminaro, J. Willenbring, and A. Williams, "An Overview of Trilinos," SAND2003-2927, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-1110, 2003.

[22] "MATLAB 8.0 and Statistics Toolbox 8.1, The MathWorks, Inc., Natick, Massachusetts, United States."

[23] M. A. Aguiló, "Design Optimization Toolkit (DOTk): Users' Manual," SAND2014-16272, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-0845, 2014.

[24] B. Adams, W. Bohnhoff, K. Dalbey, J. Eddy, M. Eldred, D. Gay, K. Haskell, P. Hough, and L. Swiler, "Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: version 5.0 users manual," *Sandia National Laboratories, Tech. Rep. SAND2010-2183*, 2009.

[25] D. Kouri, D. Ridzal, B. van Bloemen Waanders, and G. van Winckel, "Rapid optimization library (ROL)." `https://trilinos.org/packages/rol/`, 2014.

[26] P. Gill, W. Murray, and M. Saunders, "Snopt: An sqp algorithm for large-scale constrained optimization," *SIAM journal on optimization*, vol. 12, no. 4, pp. 979–1006, 2002.

[27] A. Wächter and L. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, no. 1, pp. 25–57, 2006.

[28] H. Banks, M. Joyner, B. Wincheski, and W. Winfree, "Nondestructive evaluation using reduced-order computational methodology," *Inverse Problems*, vol. 16, no. 4, pp. 929–945, 2000.

[29] C. Vogel, *Computational methods for inverse problems*. Society for Industrial and Applied Mathematics, 2002.

[30] B. Lazarov and O. Sigmund, "Filters in topology optimization based on helmholtz-type differential equations," *International Journal for Numerical Methods in Engineering*, vol. 86, no. 6, pp. 765–781, 2011.

[31] J. Perry, "A class of conjugate gradient algorithms with a two-step variable-metric memory," *Discussion Papers*, vol. 269, 1977.

[32] D. Shanno, "On the convergence of a new conjugate gradient algorithm," *SIAM Journal on Numerical Analysis*, vol. 15, no. 6, pp. 1247–1257, 1978.

[33] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust region methods*, vol. 1. SIAM, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
R. Ly, J. Jumel, M.E.R. Shanahan, F. Lavelle
Crete Island, Greece, 5–10 June 2016

COHESIVE ZONE MODEL IDENTIFICATION WITH DCB TEST PARAMETER ESTIMATION SENSITIVITY

Racine LY[1, 2], Julien JUMEL[1], Martin E.R. SHANAHAN[1] and Florian LAVELLE[2]

[1] University of Bordeaux, I2M, UMR 5295
351 Cours de la Libération, F-33400 Talence, France
e-mail: {racine.ly, julien.jumel, martin.shanahan}@u-bordeaux.fr

[2] Centre National d'Etudes Spatiales, DLA
52 Rue Jacques Hillairet, 75612 Paris, France
florian.lavelle@cnes.fr

Keywords: Cohesive zone model, DCB, interface bonding, identification, confidence interval

Abstract. *Crack initiation and propagation conditions during a double cantilever beam test under constant rate loading are investigated. Elastic-plastic cohesive zone model is considered to represent the interface separation law. A finite difference model is derived to evaluate the applied force versus specimen opening evolution. Cohesive zone model parameters effect on energy release rate is investigated to evaluate the ability of measuring protocol based on J-integral approach for cohesive zone model identification. Confidence intervals of these estimated parameters are evaluated rigorously considering reasonable assumption on numerical simulation conditions regarding experimental one, with nonlinear optimization algorithm.*

1 INTRODUCTION

Cohesive zone models are now extensively used for evaluating the failure conditions of interfaces in laminates or bonded joints. They allow evaluating the development of cohesive stresses along the crack propagation path and are relevant for analysis of both crack nucleation and propagation conditions under monotonous loading. More advanced models are now proposed to deal with fatigue, creep or ageing damage of the interface so that more complex loading configuration can be analysed for the structure itself. Today, these models have been studied extensively from a theoretical and numerical point of view and several contributions have been proposed to compare the influence of the chosen CZM on failure load prediction for example.

Indeed, the CZM are still generally identified rather than measured so that the interface separation law shape is still chosen an empirical manner based on expected global behaviour. Bilinear shape is chosen for brittle like behaviour and trapezoidal shape when ductile behaviour is expected. This approach is clearly not appropriate and some efforts have been devoted to the development of specialized experimental techniques for measuring a precise manner the interface separation law (viz. cohesive stress versus displacement jump across the interface). These characterization protocols make use of new specimen geometry or loading system or using standard specimens utilize new analysis techniques. In the following we will concentrate on such new development and focus on the identification of CZM under mode I loading configuration with the Double Cantilever Beam (DCB) test. The traditional technique for "measuring" interface separation law from DCB crack propagation experiment generally consists in identifying the parameters of the CZM using numerical fitting procedure using the global applied force versus opening displacement evolution. Such protocol generally results in poor sensitivity on the global shape of the interface separation but allows precise evaluation of the interface fracture energy. Recently several attempts have been made to implement the digital image correlation (DIC) technique for more precise evaluation of the CZM. The side of the specimen is observed to measure the surface displacement of both adherents in the vicinity of the crack tip. It is expected that more numerous experiment data will improve the global sensitivity of the CZM and crack tip position identification. More original and sophisticated protocols have also been proposed such as the $J(\theta)$ method which is based on clever derivation of the J integral and direct measurement of the crack tip opening displacement and specimen end rotation. With this protocol no assumption concerning the CZM shape is required and the J (θ) method effectively provide direct estimate of the interface separation law which can be implemented also under mixed mode condition. Similar evaluations have been also obtained by using the effective crack length method leading to a more straightforward evaluation of the CZM. Finally, the use of strain gages for monitoring the development of damage along an interface in bonded joint or laminate have been recommended for a long time since such measurement seems to provide easy implementation and high sensitivity to damage detection. However, only recently, a direct CZM reconstruction scheme from backface strain monitoring during crack propagation experimental have been proposed while classical CZM identification procedure are generally used. However, direct CZM reconstruction methods are generally limited to simple monotonous loading configurations, direct evaluation of interface separation law under stationary, fatigue or environmental coupling still appears difficult at this step.

However, some works have been conducted for cohesive zone model parameters sensitivity and identification. Alfano and Fugiuele [1] have compared experimental data from mode-I crack test of adhesively bonded joints, with those obtained by modeling adhesive layer with exponential intrinsic cohesive zone model, bi-linear and trapezoidal. A good agreement be-

tween experiments and simulation has been obtained and show that energy release rate is the only influencing parameter during propagation phase. They also show that bi-linear and exponential laws well reproduce the primer phase behavior, while trapezoidal law overestimates peak force at which propagation appears. Although critical energy release rate and maximum cohesive laws stress are often seen as most important parameters towards model outputs, Waas and Gustafson [2] concluded with kriging method that, maximum cohesive law stress does not significantly influence model outputs for crack tests like double cantilever beam and end notched flexure type. In addition, several authors have proposed cohesive zone model identification methods. Two approaches are worth noting: the first known as direct which makes no assumptions about law's shape, the second called indirect, imposing law's shape whose parameters are calibrated by minimizing a cost function that measure metric between numerical and experimental observations. In the first case, Sorensen and Jacobsen [3] have adopted Li and Ward [4] approach to successfully measure cohesive zone law by deriving energy release rate J with respect to the CTOD. For the second approach, Valoroso and Sessa [5] identify a posteriori an exponential law whose parameters are re-calibrated compared to experimental data. It should be noted that formulation of cohesive zone model mentioned above is made such that damage is embedded in that it corresponds to stiffness variations. They are used for materials with elastic/elastic-plastic time independent behavior. Here we focus on such material type.

In this article the crack initiation and propagation during double cantilever beam test under monotonous loading is simulated considering elastic perfectly plastic interface separation law. The classical force versus opening displacement curves are evaluated and compared with a theoretically derived form. Additionally the specimen end rotation and crack tip opening displacement as required by the $J(\theta)$ technique is conducted too. The sensitivity of $J(\theta)$ experimental techniques to the CZM parameters is evaluated and reconstruction of the CZM with the various techniques is conducted.

2 ELASTIC PLASTIC COHESIVE ZONE MODEL

Cohesive zone model consist of an interface phenomenological law represented by a relationship between cohesive stress and displacements jump at the bonding line. First models were derived by Barenblatt [6] and Dudgale [7] in 60's and since then, many models were proposed for fatigue, ageing and complex loads. Here we focus on an elastic-plastic cohesive zone model used as interface law for initiation and crack propagation during a mode-I fracture between to substrates assembled by structural adhesive e.g. for the double cantilever beam test.

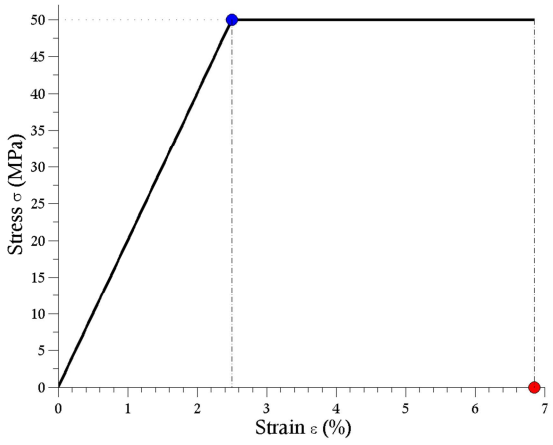


Figure 1 Elastic Perfectly Plastic cohesive zone model

In figure 1 is illustrated a perfect elastic-plastic cohesive zone model. This model, as its name suggests, can be divided in two parts: elastic and plastic. The elastic part is characterized by two parameters σ_{max} and ε_e linked by the initial stiffness k_a, and the second part depends on σ_{max} and strain at break ε_{max}. According to strain partition principle, total strain can be written as the sum of an elastic reversible instantaneous part, and a second one where all dissipation processes are embedded. Here, we consider that this dissipation is provided by plastic phenomena. In addition, in order to take into account damage evolution, Kachanov damage evolution is used in term of mechanical property deterioration. Then, damage is represented here, by the secant stiffness decreasing at each load step. Concerning energy aspects, cohesive zone model formulations are made such that, the critical energy release rate G_c is equal to the area under the law curve.

$$\mathrm{G_c} = \int_0^{\varepsilon_{\max}} \sigma(\varepsilon)\mathrm{d}\varepsilon$$

In our case, the energy can be partitioned into elastic and plastic energy such that $G_c = G_e + G_p$. In table 1 is represented quantities defined above and their formulation.

3 DCB MODELING USING ELASTIC-PLASTIC COHESIVE LAW

Double cantilever beam test is the common one to measure fracture energy in adhesively bonded joints. It sample consist of two adherents assembled by adhesive and loaded in order to initiate and propagate cracks. Standards like ASTM D5528 and ISO 25217, provides different data reduction techniques to compute fracture energy.

Parameters	Formulation	Definition
w	-	Width
$\mathrm{t_a}$	-	Thickness
$\mathrm{E_a}$	-	Young modulus
$\mathrm{k_a}$	$\frac{\mathrm{wE_a}}{\mathrm{t_a}}$	Initial stiffness
$\mathrm{G_e}$	$\frac{1}{2}\sigma_{\max}\varepsilon_{\mathrm{e}}$	Elastic energy
$\mathrm{G_p}$	$\sigma_{\max}(\varepsilon_{\mathrm{m}} - \varepsilon_{\mathrm{e}})$	Plastic energy
$\mathrm{G_c}$	$\mathrm{G_e} + \mathrm{G_p}$	Critical energy

Table 1 Parameters definition for elastic-plastic cohesive zone model

Many works have been proposed to model the double cantilever beam test with cohesive zone model as interface law and most of them use the finite element method. In this present, sample symmetry allows to consider only a half width and Winkler based modeling is used to derive constitutive equations. For this purpose, double cantilever beam sample is considered as beam lying on non-linear springs.

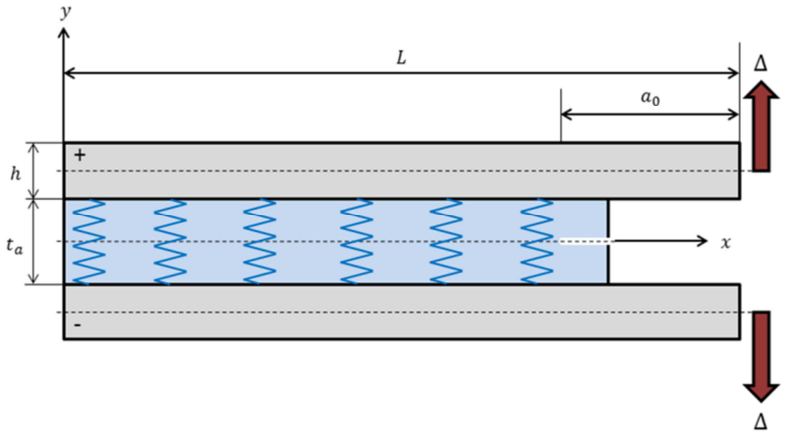


Figure 2 Double Cantilever Beam test specimen illustration

Timoshenko beam kinematics is used and springs follow a perfect elastic-plastic cohesive law. Then, cohesive stress can be seen as foundation reaction force due to opening load at the end-beam load point. In addition, the length of the bonded part is $L - a_0$ and the two slabs located at specimen right end are left unbounded over a distance equal to a_0 and considered as the initial crack length.

3.1 Constitutive equations

Specimen deformation is evaluated at each step of the experiment by solving local beam equilibrium and constitutive equations as presented below:

$$\frac{dM}{dx} + T = 0 \qquad (1)$$

$$\frac{dT}{dx} - w\sigma = 0 \qquad (2)$$

$$M = EI\frac{d\varphi}{dx} \qquad (3)$$

$$T = \kappa GS\left[\frac{dv}{dx} - \varphi\right] \qquad (4)$$

Where M, T, φ and v are bending moment, shear force, rotation and displacement respectively. Here σ is the local stress defined by cohesive law presented in Section 2. Its expressions depend on local secant stiffness χ^* expand as below:

$$\chi^* = \begin{cases} \frac{E_a}{t_a} & \text{if } \sigma < \sigma_{max} \\ \frac{\sigma}{v} & \text{if } t_a\varepsilon \le v \le t_a\varepsilon_{max} \\ 0 & \text{if } v \ge t_a\varepsilon_{max} \end{cases} \qquad (5)$$

Combining equations (1) and (2) on one hand, and equations (2), (3) and (4) in the other, the driven equations for double cantilever beam test modeling obtained are expressed as follows :

$$\frac{d^2M}{dx^2} + w\chi^* v = 0 \qquad (6)$$

$$\frac{d^2v}{dx^2} - \frac{M}{EI} - \frac{w\chi^*}{\kappa GS} v = 0 \qquad (7)$$

3.2 Numerical resolution and algorithmic aspects

Normal displacements along beam and opening force at beam right end is conducted by solving equations (6) and (7) which are a PDE system with bending moment and displacements as unknown variables. χ^* is computed at each step using expressions and conditions in (5). Concerning resolution method, a finite difference method is used to approximate first and second derivatives: for first derivatives, a backward scheme is used and for second derivatives, a centered second order scheme. Then, at spatial coordinate x_k and $x_{k+1} = x_k + \Delta x$, where Δx is the spatial step one, can write:

$$[M(x) \quad v(x)] = \begin{cases} [M_k \quad v_k] & \text{if } x = x_k \\ [M_{k+1} \quad v_{k+1}] & \text{if } x = x_k + \Delta x \end{cases} \qquad (8)$$

Equations (6) and (7) can be rewritten as:

$$M_{k-1} - 2M_k + M_{k+1} + wdx^2\chi^* v_k = 0 \qquad (9)$$

$$-\frac{dx^2}{EI} M_k - \left(2 + \frac{wdx^2\chi^*}{\kappa GS}\right) v_k + v_{k-1} + v_{k+1} = 0 \qquad (10)$$

On can remark that (9) and (10) can be resume in a matrix system as follow:

$$\begin{bmatrix} 1 & 0 & -2 & wdx^2\chi^* & 1 & 0 \\ 0 & 1 & -\frac{dx^2}{EI} & -\left(2 + \frac{wdx^2\chi^*}{\kappa GS}\right) & 0 & 1 \end{bmatrix} \begin{bmatrix} M_{k-1} \\ v_{k-1} \\ M_k \\ v_k \\ M_{k+1} \\ v_{k+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Concerning boundary conditions at beam right side, an imposed displacement is applied and for beam left side, beam on elastic foundation conditions type is adopted. Defining a compliance matrix as C, relation between normal displacements and rotation firstly, and shear force and bending moment in second, can be expressed according to degree of freedom allow in our configuration. Then, we obtain:

$$\begin{bmatrix} v \\ \varphi \end{bmatrix} = \begin{bmatrix} C_{vT} & C_{vT} \\ C_{\varphi T} & C_{\varphi M} \end{bmatrix} \begin{bmatrix} T \\ M \end{bmatrix}$$

Below is illustrated the generic algorithmic scheme that was developed for double cantilever beam test using cohesive zone model as interface law. Three computations level appears depending on which part of the code is activated. Level 0 is related to raw data, level 1 to loop initiation, level 2 when the code operates inside the loop and finally level 3 when the stopping rule allows getting out the loop.

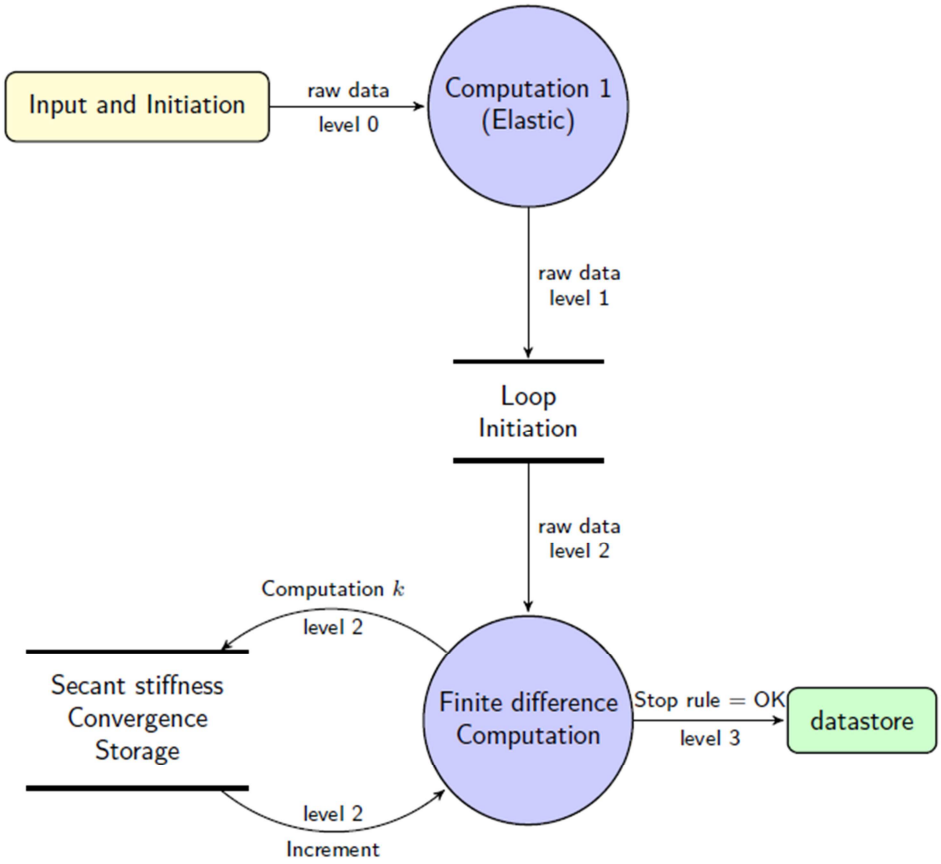


Figure 3 Algorithmic scheme for DCB experiment modeling

4 NUMERICAL RESULTS

In this section, are shown numerical results obtained from a double cantilever test simulation using an elastic perfectly plastic cohesive zone model as interfacial behavior for adhesive layer. We consider Aluminum substrate as Timoshenko beam with 70e9 Young modulus, 0.3 Poisson ratio and 200 mm length. In addition, a 30 mm initial crack length is defined and the specimen is 10 mm thick and 15 mm of width. For the cohesive zone model, three parameters are taken into account: Young modulus E_a, maximum local σ_{max} stress and strain at break ε_{max}. Initial values are respectively set to 2GPa, 50MPa and 6.85%. Opening force P and opening displacement Δ at the specimen right end are computed, as energy release rate J using J-integral formula and apparent crack length. Finally, a constant rate displacement load is applied at specimen right end in order to charge it in Mode-I for initiation and crack propagation. Results are compared to analytical formula provided by single beam theory approach.

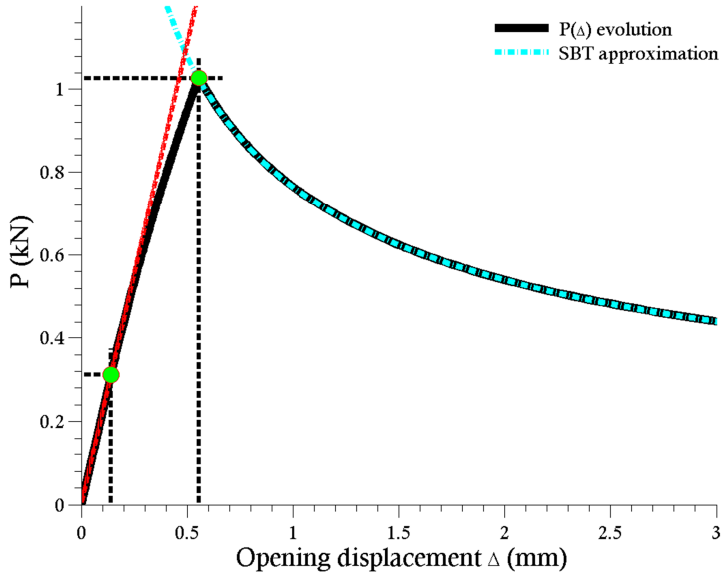


Figure 4 Opening force versus Opening displacement

In Figure 4 is shown the opening force according to the opening displacement load. The curve presents three different parts. The first one is linear and is referring to fast processes like those of elastic behavior and can be related to cohesive zone model elastic part. The second one is characterized by the loss of linearity and corresponds to the plastic part of the CZM. And the last part is related to the crack propagation. For this last, an analytical formula derived from single beam theory allow to relate P and Δ using simple beam theory (Eq. 11), the curve provided by this formula is illustrated by the cyan dotted lines and perfectly match our P (Δ) crack propagation evolution.

$$P = \left(\frac{4EI}{9}\right)^{1/4} (wG_c)^{3/4} \frac{1}{\sqrt{\Delta}} \qquad (11)$$

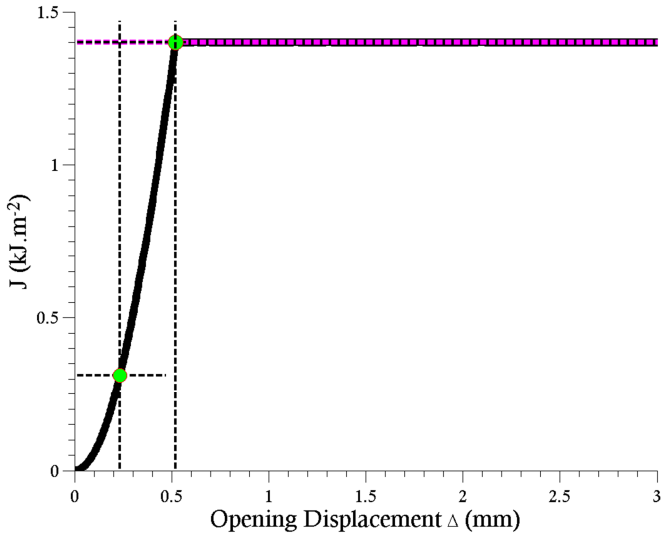


Figure 5 Energy release rate J according to Opening displacement

Figure 5 shows three parts also, and they are related to same processes as for P (Δ) evolution. However, a point has to be made concerning the propagation part. In fact, we assume in this paper a stationary propagation which induced a constant energy release rate during this phase and equal to the area under the cohesive zone model which should be an assumption made when loading speed is low. On can also remark that J becomes constant at $\Delta = 0.5\text{mm}$ and at this value in Figure 4, the force reached his maximum value $P_{max} = 1.1$ kN. Such facts that can be seen as crack propagation criteria. In Figure 6 is represented the same quantity J plotted in terms of crack opening displacement δ, only crack initiation phase is represented because of during propagation phase, we assume that J is constant. According to J-integral approach theory, J is written as path independent contour integral as:

$$J = \oint_{\Gamma} \sigma . d\delta$$

Then, according to cohesive zone model shape, we can mention two essential parts of J (δ) evolution. In fact, the first part of CZM is the elastic one represented by an affine curve, then its integral should be a parabolic type curve, as we see in Figure 6 first phase. Using the same idea, CZM plastic is constant then its integral is a line as illustrated below. In conclusion, the modeling strategy used here by considering DCB specimen as Timoshenko beam lying on elastic perfectly plastic foundation through a Winkler type approach shows good agreement between analytical result provided by J-Integral approach and his computed evolution. The same accuracy can be found on crack propagation phase on Figure 4, where numerical results perfectly matched the theoretical evolution.

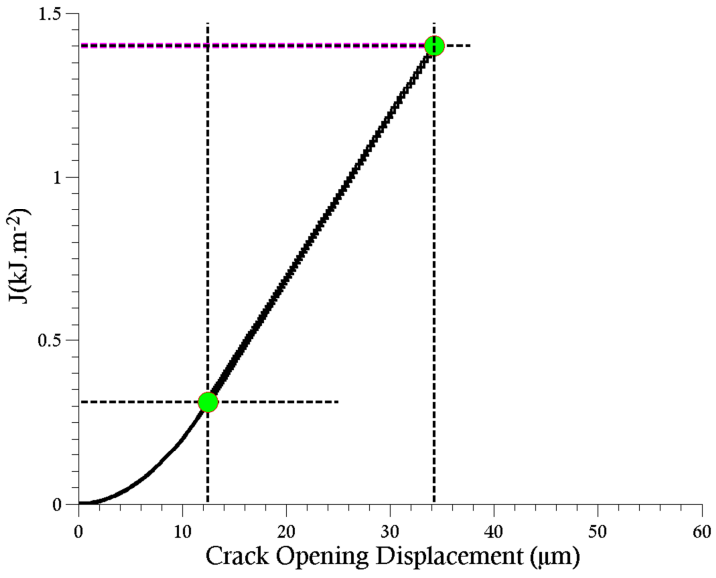


Figure 6 Energy release rate J according to Crack Opening Displacement

5 COHESIVE ZONE MODEL IDENTIFICATION

5.1 Sensitivity analysis

Sensitivity analysis in order to quantify how CZM parameters affect J (δ) evolution is conducted. For this, as influence indicator, we use gradient of J according to parameters and a weight factor is affected to each gradient in order to avoid units and scale gap as shown in Eq. 12. However, regarding to curve partitioning aspects discussed above, it's convenient to compute these quantities, first for the elastic part and second for the plastic phase. Such approach is chosen to facilitate cohesive zone model parameters identification, when coupling between parameters can be negligible or inexistent.

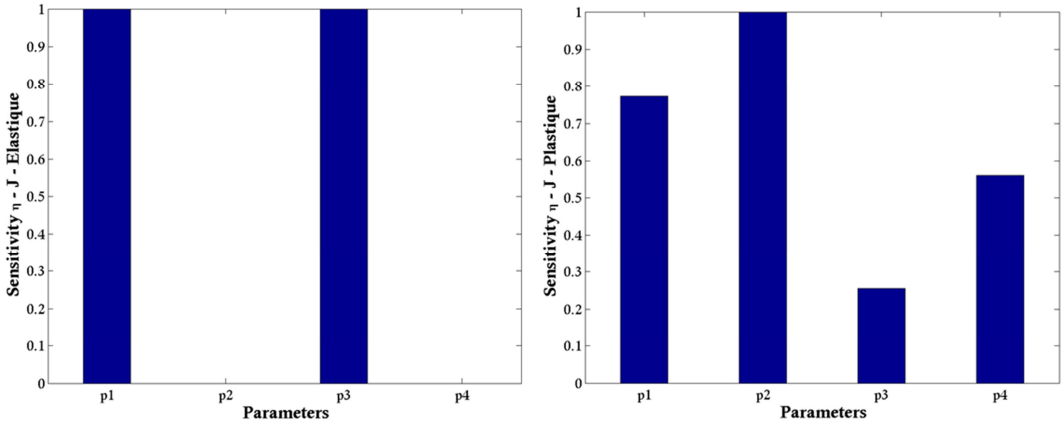


Figure 7 CZM parameters sensitivity index with J (δ) evolution

$$\eta_i = \frac{p_0}{J_0}\frac{\partial J_{p_i}}{\partial p_{\hat{i}}} \qquad 12$$

Where, p_0 and J_0 are respectively parameters initial values and corresponding J values. Parameters p1, p2, p3 and p4 corresponds respectively to t_a, σ_{max}, E_a, and ε_{max}. We can see as shown in Figure 7 that adhesive thickness and Young modulus are the most important parameters in the elastic part. Nonetheless, we can also assume that adhesive thickness can be accurately measured and then his influence could be neglected. Young modulus is then the only driving parameter during the elastic phase as it could be expected. But sensitivity in plastic zone shows that p2, p3 and p4 have an influence on J response and a coupling between them should be taken into account. In this article, only p2 and p4 e.g. σ_{max} and ε_{max} taken separately are considered.

5.2 CZM parameters confidence interval

Cohesive zone model parameters identification is performed using minimization of a cost function with Levenberg-Marquardt algorithm in elastic and plastic part. Young modulus is identified in the first one and maximum stress and critical opening strain in the later. For this, synthetized experimental data are built by adding to simulated ones Gaussian noise with standard deviation of 10. In addition, random CZM parameters values are injected as seed for the model and asymptotic standard deviation is computed when convergence rule is satisfied. Parameters convergence and residuals evolution are provided in Figure 8, 9 and 10. Ratio between standard deviation and fitted parameters are reported in Table 2.

Parameters	σ_p/p_{fit} (%)
E_a	27.51
σ_{max}	8.86
ε_{max}	8.93

Table 2 Parameters ratio: standard deviation obtained over fitted values

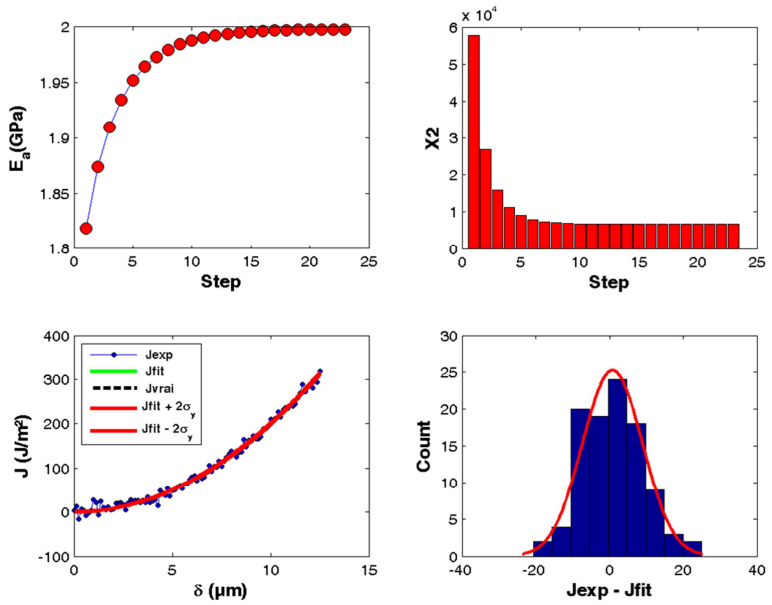


Figure 8 Young modulus identification

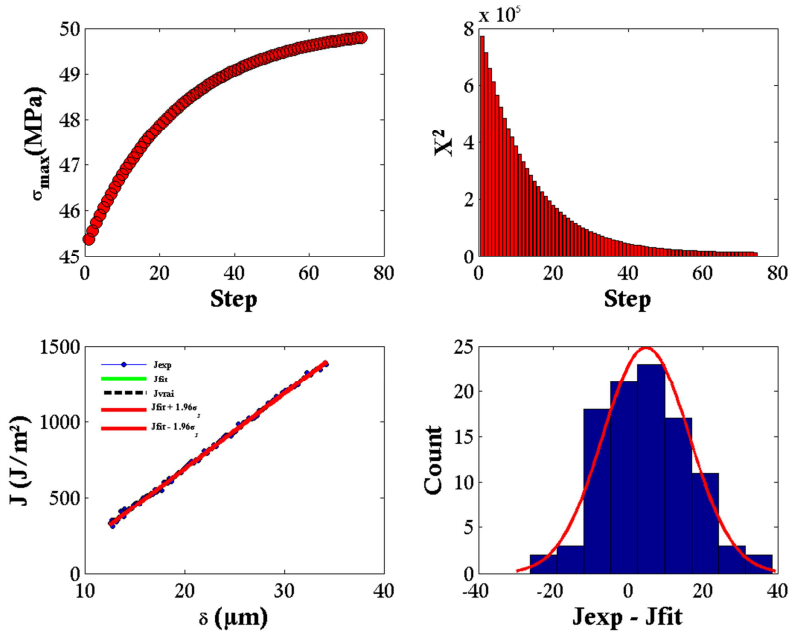


Figure 9 Maximum stress identification

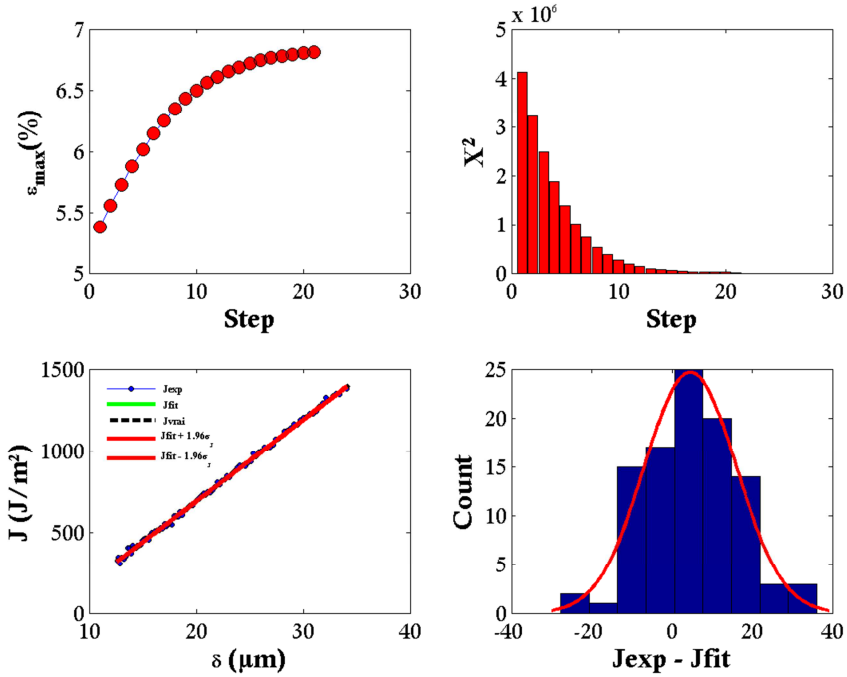


Figure 10 Strain at break stress identification

6 CONCLUSIONS

Crack initiation and propagation conditions during a double cantilever beam test under constant rate loading were investigated. Elastic-plastic cohesive zone model was considered to represent the interface separation law. A finite difference model was derived to evaluate the applied force versus specimen opening evolution and the parameters effect on energy release rate based on J-integral approach. Asymptotic standard deviation obtained from an optimization process, provide tight confidence interval. Since J could be measure with three parameters which is force, quantifiable with the traction machine and cross section rotation measurable with inclinometers, a cohesive zone model could be obtain using J first derivative according to crack opening displacement. The latter is also accessible from digital image correlation.

Future works must be devoted to more sophisticated sensitivity index estimation like those stochastic for example, and a comparison with experimental data in order to validate such model.

REFERENCES

[1] M. Alfano and al, Mode I fracture of adhesive joints using tailored cohesive zone model. *International Journal of Fracture*, **157**,193-204, 2009.

[2] A.M. Waas P.A. Gustafson. The influence of adhesive constitutive parameters in cohesive zone finite element models of adhesively bonded joints. *International Journal of Solids and Structures*, **46**, 2201-2215, 2009.

[3] T.K. Jacobsen B.F. Sorensen. Determination of cohesive laws by the j integral approach. *Engineering Fracture Mechanics*, **70**, 1841-1858, 2003.

[4] R.J. Ward V.C. Li. A novel testing technique for post-peak tensile behavior of cementitious materials. *Fracture Toughness and Fracture Energy*, 183-195, 2000.

[5] N. Valoroso and al. Identification of mode-i cohesive parameters for bonded interfaces based on DCB test. *Engineering Fracture Mechanics*, **104**, 56-79, 2013.

[6] G.I. Barenblatt. The mathematical theory of equilibrium cracks formed in brittle fracture. *Zhurnal prikladnoy mekhaniki i tekhnicheskoy fisiki*, **4**, 3-56, 2000.

[7] D.S. Dudgale. Yielding of steel sheets containing slits. *Journal of the mechanics and physics solids*, **8(2)**, 100-104, 1960.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOPOLOGY OPTIMIZATION FOR ADDITIVE MANUFACTURING WITH CONTROLLABLE SUPPORT STRUCTURE COSTS

Matthijs Langelaar

Delft University of Technology
Mekelweg 2, 2628 CD Delft, the Netherlands
e-mail: m.langelaar@tudelft.nl

Keywords: Topology optimization, additive manufacturing, overhang angle, support structures, self-supporting designs, manufacturing restrictions

Abstract. *Advances in additive manufacturing (AM) allow economical production of components with unprecedented geometric complexity. This offers exciting opportunities for innovative designs, and particularly topology optimization has been identified as a key technique to fully exploit the capabilities of AM. However, also AM involves manufacturing restrictions, such as limitations on the inclination of overhanging parts. To deal with this problem, either sacrificial supporting structures can be added during the process, or only self-supporting designs can be considered. Both approaches have disadvantages, as support structures add material and post-processing costs, while demanding exclusively self-supporting designs may impose strong restrictions on achievable performance. With current methods, designers are limited to a choice between these two extremes. To open up a wider range of designs, this paper presents and demonstrates a topology optimization formulation that allows the designer to find trade-off solutions between design performance and support structure costs, considering both printing and removal costs.*

1 INTRODUCTION

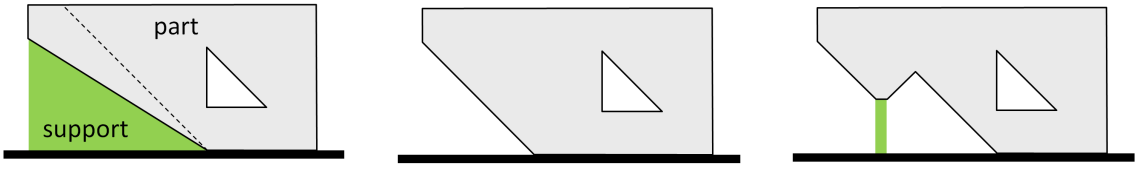
Additive manufacturing (AM) processes such as Selective Laser Melting (SLM) or Electron Beam Melting (EBM) allow economic fabrication of metal components with unprecedented geometric complexity [1]. This offers exciting opportunities for innovative designs, and particularly topology optimization (TO) has been identified as a key technique to fully exploit the capabilities of AM [2, 3, 4]. However, there are still geometric restrictions that must be taken into account. An important restriction that applies to popular powder-bed processes such as SLM and EBM, is that the inclination of printable overhanging parts with respect to the build plate is limited. This limitation is associated with a critical overhang angle, that typically amounts to 45° [5, 6].

Current practice is to first generate an optimized design using TO, and to subsequently find a part orientation that meets the overhang restriction. However, complex parts often lack a fully printable orientation. In that case, either the part must be modified, or sacrificial support structures must be printed together with the part itself (see Fig. 1(a)), and removed afterwards. The first option reduces part optimality, while the latter adds material and machining costs. A more preferable approach would be to include the AM restrictions in the TO process, making the subsequent steps unnecessary.

As a remedy, authors have explored the idea to e.g. optimize the addition of supporting structures, that become permanent parts of the final component [7]. Unfortunately this affects its optimality. Others have taken steps to identify overhang angles already during optimization, in order to generate fully self-supporting optimized designs [8, 9]. Still it was found that the previously published methods leave room for the optimization process to produce inadmissable structures.

To address this problem, this contribution first summarizes our approach to include the critical overhang angle restriction in the TO process. This approach utilizes a simplified AM fabrication model, that is applied in every TO iteration to transform a blueprint design layout into an as-printed part [10, 11]. The AM model is defined in a continuous, differentiable way, and sensitivities of the modeled AM operation can be computed efficiently. Applying TO including this AM fabrication model results in optimized designs that are fully printable, and rigorously comply with the overhang angle limitation. The necessary design modifications however may result in performance reduction compared to freely optimized parts, as illustrated by Fig. 1(b).

In the outlined approach, designers do not have an option to exercise control over the trade-off between performance reduction and support structure costs, which is restrictive. Building on the mentioned AM fabrication model, we therefore propose a TO problem formulation that allows for the simultaneous design of the actual part and the support structures required for its



a) Unrestrained design b) Self-supporting design c) Trade-off design

Figure 1: Schematic 'Design for AM' options: a) parts optimized without considering AM restrictions typically require additional support material, b) fully self-supporting designs may show reduced performance / higher mass, c) the method proposed in this paper finds a compromise solution that balances part performance and support costs.

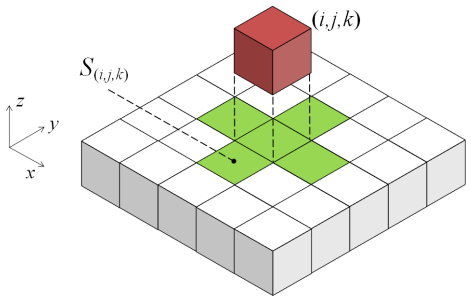


Figure 2: Definition of 3D AM filter, assuming the z-direction as the build direction. The green region $S_{(i,j,k)}$ denotes the supporting region of a red element at position (i, j, k) in a mesh. When insufficient printed material is present in this region, element (i, j, k) cannot be printed.

fabrication. The designer can specify costs associated to the printing of added support material, as well as costs for the machining effort required to remove a given amount of supports. Depending on these costs, the optimization finds the best trade-off between reduction in performance due to part adaptation, and added costs related to support structures (Fig. 1(c)). This allows for a controllable, rational approach to TO for AM, that enables the designer to find optimal designs in terms of cost and performance.

In the following, Section 2 introduces the AM fabrication model and the proposed trade-off formulation, followed by several 3D numerical examples in Section 3 and concluding remarks in Section 4.

2 FORMULATION

2.1 Printable volume detection

In order to detect which parts of a design require additional support, an operation is needed to identify the printable part of a given design layout. This is achieved by applying a simplified AM process simulation, that determines which parts of a layer are printable given the printable parts of the underlying layer. We distinguish two interrelated density values per element: a *blueprint density* b, expressing the material layout specified by the optimizer, and a *printed density* p, indicating the actually printable structure. The following subsections summarize the definition of this AM filter and its sensitivity analysis, for a more in-depth discussion the reader is referred to [10, 11].

2.1.1 AM filter definition

The simplified AM process model is defined on a structured grid, as depicted in Fig. 2. An element at position (i, j, k) is associated with a support region $S_{(i,j,k)}$, consisting of the element directly underneath and its four neighbours. The maximum printed density in $S_{(i,j,k)}$ determines the maximum printable density it can support. Mathematically this is expressed by:

$$p_{(i,j,k)} = \min\left(b_{(i,j,k)}, \max(p \in S_{(i,j,k)})\right). \tag{1}$$

By definition, the base layer ($k = 1$) is fully supported by the baseplate and thus fully printable: $p_{(*,*,1)} \equiv b_{(*,*,1)}$. In all subsequent layers, Eq. 1 is applied by sweeping through the domain in the build direction. In order to make this operation continuously differentiable, the min and max operators are replaced by smooth approximations $\check{s}$ and $\hat{s}$. Further details are given in [10].

2.1.2 Sensitivity analysis under AM filter

A response f is evaluated on the printed part $\mathbf{p}$, which is indirectly defined by $\mathbf{b}$, i.e. $f = f(\mathbf{p}(\mathbf{b}))$. Sensitivities $\partial f/\partial \mathbf{p}$ are readily obtained, e.g. using adjoint sensitivity analysis. For gradient-based optimization, sensitivities in terms of $\mathbf{b}$ are required, which can be computed by:

$$\frac{\partial f}{\partial \mathbf{b}} = \frac{\partial f}{\partial \mathbf{p}}\frac{\partial \mathbf{p}}{\partial \mathbf{b}} = \frac{\partial f}{\partial \mathbf{p}}\mathbf{J}_{pb} \tag{2}$$

The Jacobian $\mathbf{J}_{pb}$ is not sparse and unattractive to compute in full, from a numerical point of view. Instead, we opt for an adjoint formulation. Using the AM filter definition $\mathbf{p}_k = \breve{s}(\mathbf{b}_k, \hat{s}(\mathbf{p}_{k-1}))$, where k is a layer index, it is found that, for layers $k > 1$:

$$\frac{\partial f}{\partial \mathbf{b}_k} = \boldsymbol{\lambda}_k^T \frac{\partial \breve{s}}{\partial \mathbf{b}_k}, \tag{3}$$

$$\boldsymbol{\lambda}_k^T = \frac{\partial f}{\partial \mathbf{p}_k} + \boldsymbol{\lambda}_{k+1}^T \frac{\partial \breve{s}}{\partial \mathbf{p}_k} \qquad \text{for } 1 < k < n, \tag{4}$$

$$\boldsymbol{\lambda}_n^T = \frac{\partial f}{\partial \mathbf{p}_n}, \tag{5}$$

with n indicating the number of layers. For the base layer, simply $\partial f/\partial \mathbf{b}_1 = \partial f/\partial \mathbf{p}_1$. Eqs. 3-5 are inexpensive to evaluate, by sweeping layerwise through the domain from the top layer back to the base.

2.2 Cost/performance trade-off

2.2.1 Problem formulation

The described AM filter rigorously eliminates designs that require support structures for overhanging sections from the design space. However, imposing these restrictions generally results in lower performance. In many cases designers may instead seek a trade-off solution, where the added costs of support material and its removal are balanced against a smaller reduction in performance. This can be achieved using the AM filter introduced in Section 2.1.1, by considering the blueprint design as a combination of the actual part and additional support structures. To do so, we introduce two new design fields: $\mathbf{d}$ representing the density distribution of the actual design, and $\mathbf{s}$ indicating the distribution of support material in the domain. The blueprint design then consists of the union of $\mathbf{d}$ and $\mathbf{s}$, which can be approximated as:

$$b_{(i,j,k)} = \hat{s}\left(d_{(i,j,k)}, s_{(i,j,k)}\right). \tag{6}$$

Next, this blueprint field is processed by the AM filter, producing an as-printed geometry $\mathbf{p}(\mathbf{b})$. This however consists of regions that are only included as support material, and regions that belong to the actual component. The component field $\mathbf{c}$ can be extracted by taking the intersection with the actual design $\mathbf{d}$, i.e.:

$$c_{(i,j,k)} = \breve{s}\left(p_{(i,j,k)}(\mathbf{b}), d_{(i,j,k)}\right). \tag{7}$$

Performance is evaluated on the component geometry $\mathbf{c}$. Moreover, the amount of support material follows from $\mathbf{s}$, and we assume the support removal effort to be proportional to a measure of the part-support interface. The latter is given by the intersection between $\mathbf{d}$ and $\mathbf{s}$, i.e.:

$$i_{(i,j,k)} = \breve{s}\left(d_{(i,j,k)}, s_{(i,j,k)}\right). \tag{8}$$

A density filter is used, hence some overlap exists between $\mathbf{d}$ and $\mathbf{s}$, which allows this definition to be used. To summarize, the formulation now involves two fields controlled by the optimizer: the actual design $\mathbf{d}$ and the supports $\mathbf{s}$. In addition, we have defined a blueprint design for the printer $\mathbf{b}$, the printed geometry $\mathbf{p}$, the component $\mathbf{c}$ and the support/part interface $\mathbf{i}$. Note that other choices are possible, e.g. the actually printed support geometry could be defined similar to $\mathbf{c}$. The present formulation was chosen for its relative simplicity.

This leads to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{d},\mathbf{s}} \quad & f(\mathbf{c}) + c_s V_s(\mathbf{s}) + c_r V_i(\mathbf{d},\mathbf{s}) \\ s.t. \quad & g_i(\mathbf{c}) \leq 0 \\ & 0 \leq \{\mathbf{d},\mathbf{s}\} \leq 1. \end{aligned} \tag{9}$$

Here f denotes the component performance indicator to be optimized, c_s and c_r are cost factors related to the amount of support material and removal costs, respectively, and g_i are certain constraints imposed on the component. V_s and V_i denote the volume of support material and support-component interface regions as defined by Eq. 8, relative to the design domain. By choosing the cost factors adequately in relation to the performance, different trade-off designs can be generated. For very high cost factors, the design tends towards a fully printable geometry without any support material. However, that result is easier to achieve by applying the AM filter to $\mathbf{d}$ only, and omitting the support variables. The benefit of this refined formulation is that trade-off solutions can be found, where some support material is used in those places, where further design adaptation would result in undesirable performance reductions. The influence of different choices of c_s and c_r is demonstrated using numerical examples in Section 3.

2.2.2 Sensitivity analysis

The responses involved in Eq. 9 in part depend on the component performance. The component geometry is obtained from the actual design combined with the printed part (Eq. 7), which itself results from the blueprint design, given by the actual design field and the support material: $g_i(\mathbf{c}(\mathbf{p}(\mathbf{b}(\mathbf{d},\mathbf{s}))),\mathbf{d})$. Given these dependencies, the design sensitivities of the responses involved in Eq. 9 follow from application of the chain rule as:

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{d}} &= \frac{\partial f}{\partial \mathbf{c}}\left(\frac{\partial \mathbf{c}}{\partial \mathbf{p}}\frac{\partial \mathbf{p}}{\partial \mathbf{b}}\frac{\partial \mathbf{b}}{\partial \mathbf{d}} + \frac{\partial \mathbf{c}}{\partial \mathbf{d}}\right) + c_r \frac{\partial V_i}{\partial \mathbf{i}}\frac{\partial \mathbf{i}}{\partial \mathbf{d}}, \\ \frac{\partial f}{\partial \mathbf{s}} &= \frac{\partial f}{\partial \mathbf{c}}\frac{\partial \mathbf{c}}{\partial \mathbf{p}}\frac{\partial \mathbf{p}}{\partial \mathbf{b}}\frac{\partial \mathbf{b}}{\partial \mathbf{s}} + c_s \frac{\partial V_s}{\partial \mathbf{s}} + c_r \frac{\partial V_i}{\partial \mathbf{i}}\frac{\partial \mathbf{i}}{\partial \mathbf{s}}, \\ \frac{\partial g_i}{\partial \mathbf{d}} &= \frac{\partial g_i}{\partial \mathbf{c}}\left(\frac{\partial \mathbf{c}}{\partial \mathbf{p}}\frac{\partial \mathbf{p}}{\partial \mathbf{b}}\frac{\partial \mathbf{b}}{\partial \mathbf{d}} + \frac{\partial \mathbf{c}}{\partial \mathbf{d}}\right), \\ \frac{\partial g_i}{\partial \mathbf{s}} &= \frac{\partial g_i}{\partial \mathbf{c}}\frac{\partial \mathbf{c}}{\partial \mathbf{p}}\frac{\partial \mathbf{p}}{\partial \mathbf{b}}\frac{\partial \mathbf{b}}{\partial \mathbf{s}}. \end{aligned}$$

The individual derivatives follow from the expressions given in Section 2.2.1 and derivatives of the chosen smooth approximations $\breve{s}$ and $\hat{s}$. The term expressing the dependence of the printed part with respect to the blueprint design, $\partial\mathbf{p}/\partial\mathbf{b}$, is not computed in full: instead the more efficient adjoint sensitivity transformation given by Eqs. 3-5 is applied.

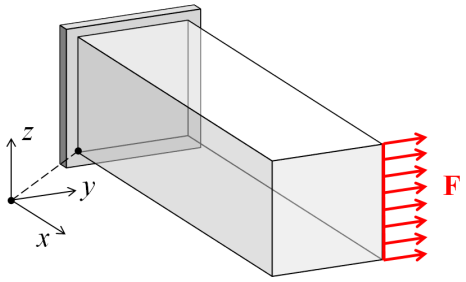


Figure 3: Cantilever beam test case used in numerical compliance minimization tests. The design domain is meshed with $150\times50\times50$ elements, and a load of 1000 N is distributed over the right front edge. The z-direction is chosen as build direction.

3 NUMERICAL EXAMPLES

3.1 Test case

In this paper, to demonstrate the additional freedom that the proposed trade-off formulation offers to the designer, we consider a cantilever beam optimized for minimum compliance, as depicted in Fig. 3. We assume that the orientation of the part has been specified a priori, and deliberately choose an orientation that is not favorable for the design, in order to clearly demonstrate the trade-off solutions the optimization process must find. Note that in real applications, part orientations are also often predetermined, due to demands on surface quality, post-processing operations, chamber size or costs (build height).

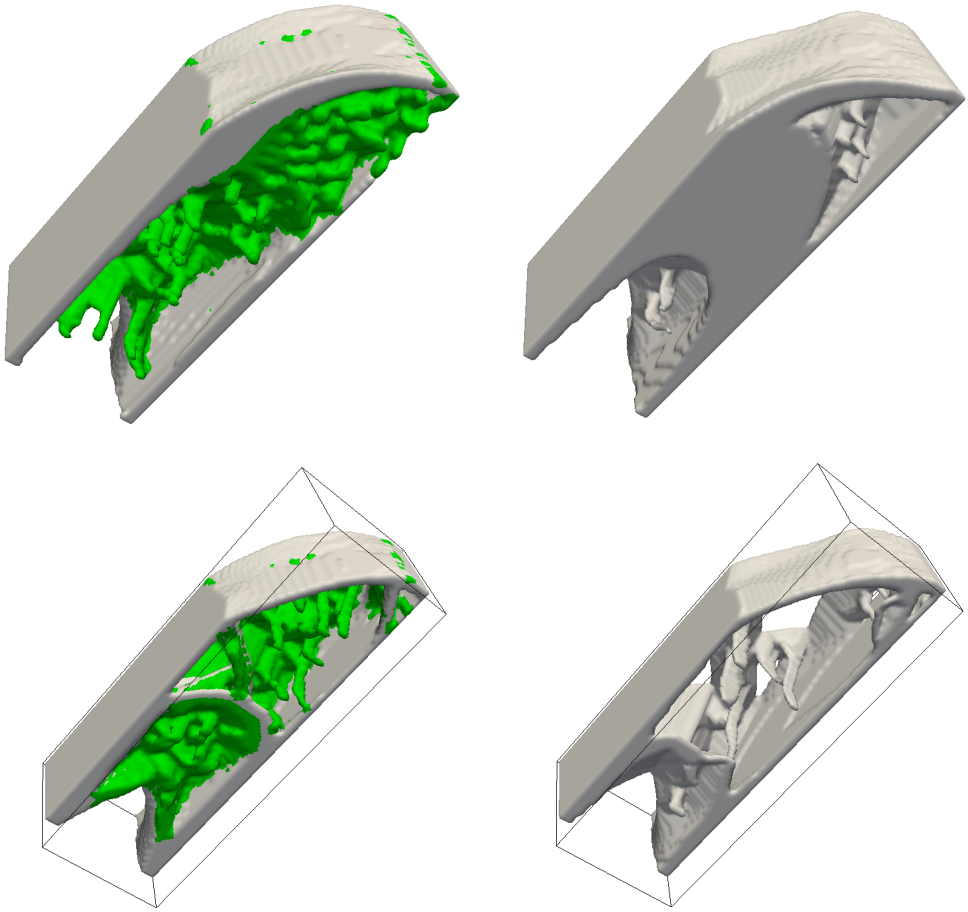
We solve the classical compliance minimization problem, under a 50% volume constraint, using the formulation given by Eq. 9. The problem is solved for different combinations of support- and removal costs c_s and c_r. Cube-shaped trilinear isoparametric finite elemens have been used, and the optimization has been performed using the Method of Moving Asymptotes [12]. Density filtering was used to define the $\mathbf{d}$ and $\mathbf{s}$ fields from the fundamental optimization variables, with a filter radius of 1.5 times the element edge length. 3D results are depicted by their isosurface, extracted at a density level 0.3.

3.2 Results

3.2.1 Reference designs

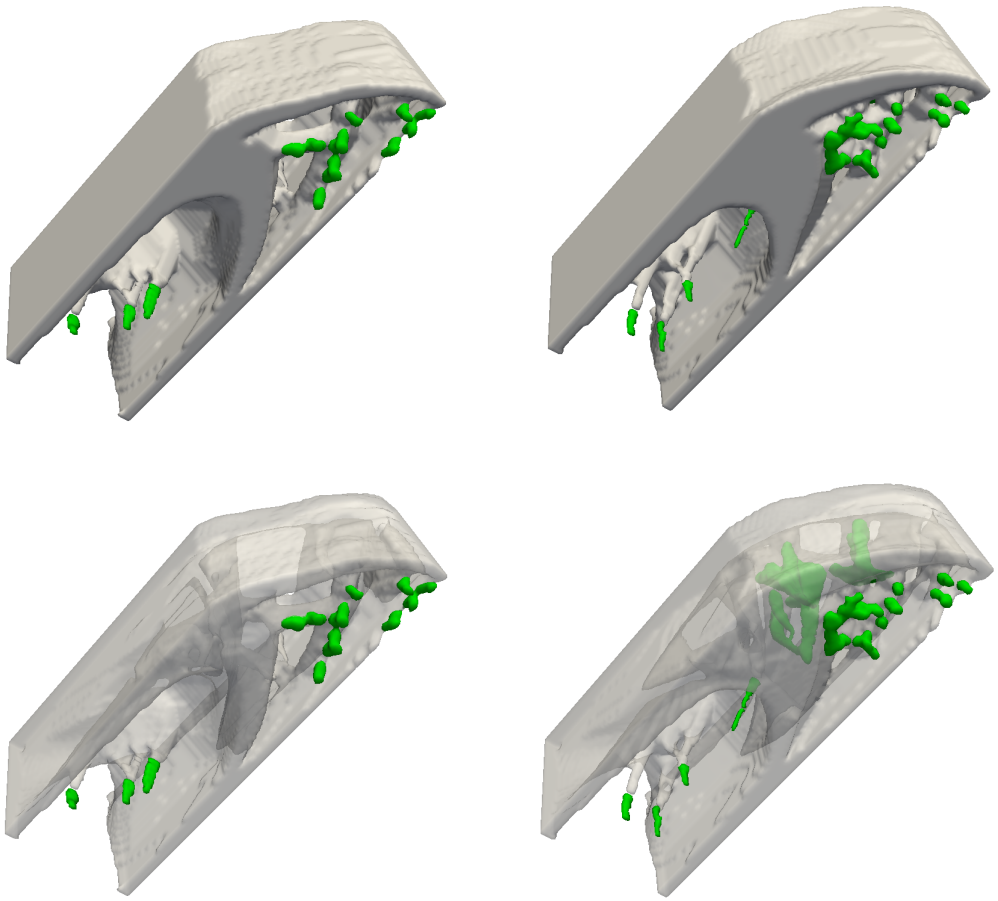
First we consider two extreme cases as reference designs: the unrestrained case, where no AM restrictions are applied (cost factors set to zero), and the fully self-supporting case (cost factors set to 10^3). This corresponds to the cases illustrated in Fig. 1(a) and (b). To interpret the cost factors, note that the numerical magnitude of the optimal compliance is around 25.

The obtained designs are depicted in Fig. 4. Note the presence of ample amounts of support material in Fig. 4(a), even coinciding with the actual component material. There is no driving force for the optimizer to be economical with support structures, hence they appear all throughout the domain. In contrast, the case with costly supports results in a fully self-supporting design, with no support structures at all (Fig. 4(b)). The compliance of the unrestrained case is $C_u = 25.55$, and that of the fully self-supporting case $C_{ss} = 27.63$. Requiring the entire part to meet the AM restrictions has worsened its compliance performance for this problem by 8%.



a) Unrestrained design b) Fully self-supporting design

Figure 4: Cantilever beam designs obtained in compliance minimization using either a) no AM restrictions or b) requiring the final design to meet the overhang restrictions everywhere, in full and cutaway view. Sacrificial support material is depicted in green, the actual component in white. The part is viewed from below the baseplate, in the printing direction.



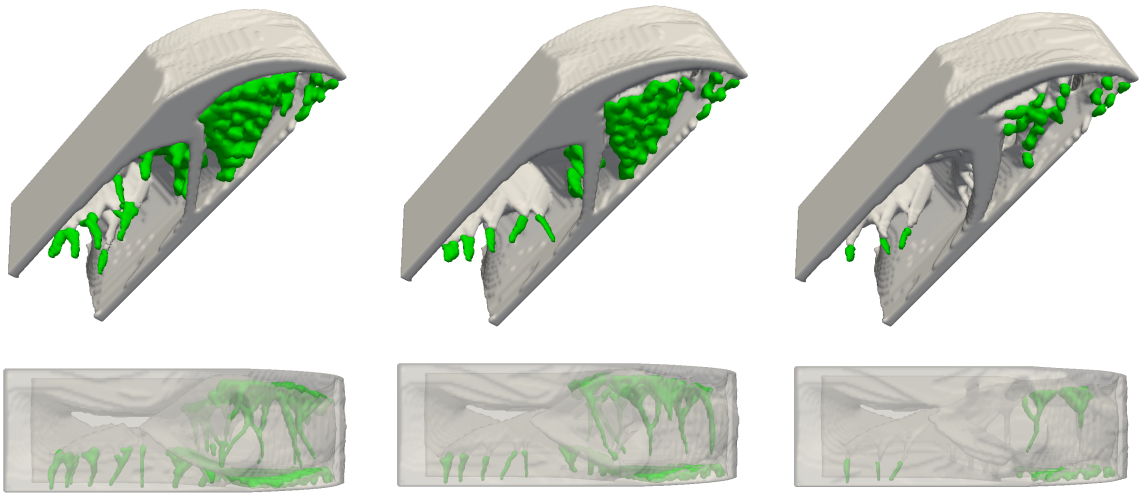
a) $c_s = 10, c_r = 0$ (no removal costs) b) $c_s = 0, c_r = 10$ (no support material costs)

Figure 5: Beam designs obtained considering either a) no removal costs, or b) no support material costs. Sacrificial support material is depicted in green, and the actual component in white, opaque and semi-transparent. The part is viewed from below the baseplate, in the printing direction.

3.2.2 Trade-off designs

Next we present various results that illustrate trends of the effect of the cost factors on the obtained designs. At the moment we have no access to realistic cost calculations, hence the focus of this study is purely on determining if the formulation has the intended characteristics. First, we consider the support and removal costs separately. Although they have different units, both are set numerically to 10.0 with their counterpart equal to zero. Resulting designs are depicted in Fig. 5.

The amount of support material employed by the optimizer is less than that when no costs were associated to its usage (Fig. 4(a)); instead, it is applied more sparingly to specific locations where the costs of either support material or interfaces balances with the gain or loss in performance. Especially in the plots with transparent component geometry in Fig. 5 it is seen that also a difference exists between restricting material usage or interface area: Fig. 5(b) clearly contains more support material. In terms of compliance, the $c_r = 0$ case is close to the fully self-supporting case C_{ss}, while the $c_s = 0$ case does slightly better: 7% higher than C_u, which corresponds to the higher support material usage seen for this design. Note that the chosen cases are not directly comparable: the cost factor 10.0 has a different meaning when



a) $c_s = 0.1, c_r = 1$ b) $c_s = 0.5, c_r = 5$ c) $c_s = 2.5, c_r = 25$

Figure 6: Beam designs obtained considering various levels of support and removal costs. Sacrificial support material is depicted in green, and the actual component in white, opaque and semi-transparent. The part is viewed from below the baseplate, in the printing direction (top row) and from above towards the baseplate (bottom row).

applied to volume or area. Nevertheless, the observed trends confirm that this formulation is able to distinguish different cost aspects.

Next, the ratio between support material and removal costs is fixed to 1:10, and different cost levels are considered. Results are depicted in Fig. 6. It is seen that changing the cost factors leads to adjustments in the optimized design, with the employed amount of support material decreasing with increasing costs.

Relative compliances of these three cases are 1.6%, 3.5% and 4.6% worse than the unrestrained design (C_u, Fig. 4(a)), but all clearly better than forcing the design to be fully self-supporting, which yielded an 8% performance drop (C_{ss}, Fig. 4(b)). This is as expected: adding more severe restrictions leads to further reductions in achievable performance. Note also that the support material layout as seen in the transparent plots shown in Fig. 6 is optimized in the process as well: to limit the amount of sacrificial material, branches are formed that merge together to form a connection to the baseplate. In the present model, no stiffness requirements are taken into account for the support structures, but the problem formulation can in principle be extended with constraints related to the printed geometry field $\mathbf{p}$ as well.

4 DISCUSSION AND CONCLUSIONS

A new formulation that enables designers to find trade-off solutions when using topology optimization for additive manufacturing has been presented. It builds on a filter-based procedure to create exclusively self-supporting designs. While effective, for practical applications, a less restrictive approach that can find a rational compromise between part performance and production costs is often desired.

The presented formulation, employing a simple cost model considering support structure volume and removal costs, generates fully printable combinations of component and support geometry. As expected, optimized part performance decreased as support and removal costs increased. When applying realistic cost factors, this allows designers to properly balance performance and costs. The cost model itself is simple and should be seen as a placeholder for more

refined cost models for specific processes. Also, an important limitation is that the present procedure does not account for the accessibility of the interfaces that require machining for support removal. Even enclosed voids may occur in the designs. Therefore, engineering judgement is still necessary to evaluate the proposed solutions.

The examples confirm that enforcing full compliance with the AM process' overhang limitations comes at a cost in terms of part performance. In this case, the maximum difference amounted only to 8%, and in fact other part orientations do not result in performance reductions [11]. The examples are however only intended to illustrate the principle, that in general restrictions often reduce performance, and that the presented cost-based formulation offers a way to find a compromise.

The current method assumes a fixed part orientation, while in reality freedom may exist to reorient the part. This can be included in the present formulation by considering multiple optimizations with different orientations. However, this directly increases the computational effort, hence also here engineering insight in the selection of suitable orientations is required. Substantial gains or losses can result from part orientation choices, and including this aspect in the optimization is considered an important topic for future work.

REFERENCES

[1] E. Atzeni, A. Salmi, Economics of additive manufacturing for end-usable metal parts, The International Journal of Advanced Manufacturing Technology 62 (9-12) (2012) 1147–1155.

[2] C. Emmelmann, J. Kranz, D. Herzog, E. Wycisk, Laser additive manufacturing of metals, in: Laser Technology in Biomimetics, Springer, 2013, pp. 143–162.

[3] D. Rosen, Design for additive manufacturing: Past, present, and future directions, Journal of Mechanical Design 136 (9) (2014) 090301.

[4] T. Zegard, G. H. Paulino, Bridging topology optimization and additive manufacturing, Structural and Multidisciplinary Optimization 53 (2016) 175–192.

[5] R. Mertens, S. Clijsters, K. Kempen, J.-P. Kruth, Optimization of scan strategies in selective laser melting of aluminum parts with downfacing areas, Journal of Manufacturing Science and Engineering 136 (6) (2014) 061012.

[6] J. Kranz, D. Herzog, C. Emmelmann, Design guidelines for laser additive manufacturing of lightweight structures in TiAl6V4, Journal of Laser Applications 27 (S1) (2015) S14001.

[7] M. Leary, L. Merli, F. Torti, M. Mazur, M. Brandt, Optimal topology for additive manufacture: A method for enabling additive manufacture of support-free optimal structures, Materials & Design 63 (2014) 678–690.

[8] D. Brackett, I. Ashcroft, R. Hague, Topology optimization for additive manufacturing, in: Proceedings of the Solid Freeform Fabrication Symposium, Austin, TX, 2011, pp. 348–362.

[9] A. Gaynor, J. Guest, Topology optimization for additive manufacturing: considering maximum overhang constraint, in: 15th AIAA/ISSMO multidisciplinary analysis and optimization conference, 2014, pp. 16–20.

[10] M. Langelaar, An additive manufacturing filter for topology optimization of print-ready designs, Structural and Multidisciplinary Optimization, in review (2016, forthcoming).

[11] M. Langelaar, Topology optimization of 3D self-supporting structures for additive manufacturing, Additive Manufacturing, in review (2016, forthcoming).

[12] K. Svanberg, The method of moving asymptotes - a new method for structural optimization, International Journal for Numerical Methods in Engineering 24 (2) (1987) 359–373.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

STUDY OF TOPOLOGY OPTIMIZATION PARAMETERS AND SCAFFOLD STRUCTURES IN ADDITIVE MANUFACTURING

A. Garaigordobil[1], **R. Ansola**[2], **and E. Veguería**[2]

[1] University of the Basque Country
Alameda Urquijo s/n . 48013 - Bilbao
e-mail: alain.garaigordobil@ehu.eus

[2] University of the Basque Country
Alameda Urquijo s/n . 48013 - Bilbao
{ruben.ansola, estrella.vegueria }@ehu.es

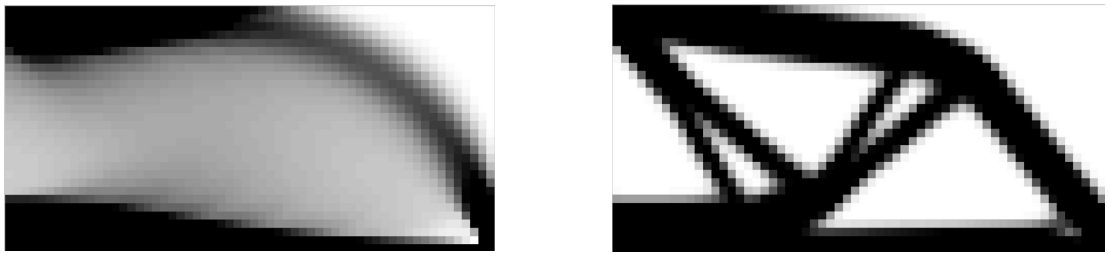
Keywords: Topology Optimization, Additive Manufacturing, Overhang, SIMP, Perimeter Control, Scaffold Structures, Geometric Complexity.

Abstract. *This work falls within the scope of topology optimization procedures and additive manufacturing technologies. In recent years the topology optimization has become a perfect tool to maximize the potential and freedom that these revolutionary manufacturing technologies offer, allowing to conceive designs that utilize available resources optimally. However, there are still many theoretical and practical issues regarding automatic integration of both technologies. This investigation aims to advance in this line of research, where algorithms that provide the ability to control and minimize support structures will be developed. These scaffold structures are usually necessary when additive manufacturing is used to construct the geometries obtained after the optimization process. The algorithms to be implemented in this project would offer the possibility to control the formation of scaffold structure and minimize them when necessary. For this purpose, this introductory paper discusses the effect of different parameters and restrictions that may have an influence on the formation of scaffold structures and are eventually involved in the topology optimization process. Considered parameters include the filter radius applied for mesh independency, penalization factor in the SIMP power law and the perimeter constrain. These strategies are frequently used to assure existence of solutions and mesh independent 0-1 designs, when a high degree of complexity is usually undesired for traditional manufacturing. This paper discusses the effect of these parameters on the length and inclination of members present in the optimum topologies. The requirement for several Additive Manufacturing processes to use support structures for large overhangs provides justification for investigating additional and specific methods for including this measure into the optimization process. In this line of work the authors will also propose a new approach to evaluate the global angle and overhang of a topology with intermediate densities, which can be incorporated as a manufacturing constraint into the topology optimization process in order to explicitly specify the amount of support structures desired in the final solution.*

1. INTRODUCTION

In past years, and due to factors such as limitations on accessible manufacturing technologies, difficulty for interpreting results,… topology optimization has not been used beyond theoretical arguments, and it was when techniques like penalization methods (SIMP, …), mesh independent filter… were developed that topology optimization began to sprout on the industry.

The tendency back then, due to the limitations that manufacturing technologies seemed to have at that time, was to simplified the optimum design using penalizations methods, filters, … Hence, despite of being close, the resulting geometry wasn´t the optimal.



(a) (b)

Figure 1: Differences between a Real Optimum (a) and a Simplified Optimum (b)

The complexity of the structure used to be a problem and many authors worked on analysing it by using filter and different parameters [1]

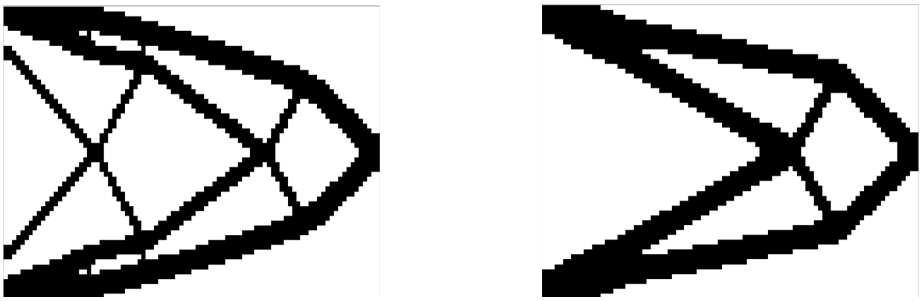
Now the situation is quite the opposite. The introduction of the modern additive manufacturing technologies has enforced the usage of topology optimization in the design process, as combining both one can fully exploit the capabilities of each of them. However, that combination presents a critic point as although the comparatively lack of limitations, additive manufacturing technologies posses one weak point which is commonly known as the overhang problem. This means that during the manufacturing process curling/warping may appear due to high temperature gradients. These situations are conditioned by the angel (relative to the structure's growing direction) of inclination of the member and its length (perpendicular projection to the structure's growing direction).

In order to face these problems, but never to avoid them, what it's been commonly done is to introduce scaffold structures. Nevertheless, the introduction of these structures supposes an increment of needed initial raw material and hence an increment of economic cost as well as an increment on manufacturing time (there are needed some processes to put and remove them), contamination… The current tendency then is to develop methods that are able to avoid these structures. These methods should include the overhang constraint into the topology optimization process.

Anyway, some authors [2] have already mentioned that increasing the complexity of the structure may help obtaining scaffold structure free geometries as the more complex a geometry is, more and shorter members are created and they can have a similar behavior to scaffold structures. Consequently, if the complexity of the structure can be controlled by controlling the typical parameters of the topology optimization problem, it may help reducing the amount of scaffold structures and may make more affordable the additive manufacturing process.

As it is been already mentioned, the overhang problem is conditioned by the length and inclination of the member and by increasing the complexity of the structure we are directly

controlling one of these parameters, the length, and as will be explained later on, minor length equals a minor problem. An example of the mentioned is shown in figure 2.



(a) (b)

Figure 2: Differences between a complex geometry (a) and simpler geometry (b) for the same structural problem

While image 2(a) shows an structure with a good amount of thin and short elements with appropriate inclination (respect to a vertical growing direction) to be manufacture by additive methods with probably no scaffold structures, image 2(b) shows a simpler geometry where members' inclination will be problematic and hence scaffold structures would be needed.

This paper seeks to present the effects of the typical topology optimization problem parameters over the complexity of the geometry and it will be proposed a technique with which the probability/necessity of an structure for needing scaffold structures can be calculated.

2. PROPERTIES OF A MEMBER

Before going in any farther detail, the properties of structural members are presented.

2.1. Inclination, Ω.

The inclination is a critical factor in the fulfillment of the overhang constraint, but it does not influence directly on the complexity of the global geometry. The inclination is measured always respect to the structural growing direction. This is represented by the vector **r**. The inclination is a relative angle and is named Ω.

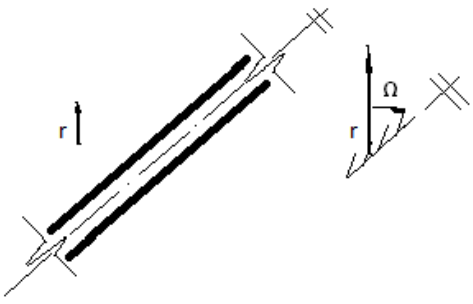


Figure 3: Inclination of a member

Regarding the restriction, problems begin to appear for relative angles greater than 45º, which is considered the theoretical limiting angle, but some authors [2] propose a penalization for the combination of the inclination and the length of the member, considering the

combined effect critical for the fulfillment of the constraint. This gives an idea of the important effect that complexity casts over the problem.

2.2. Length, δ.

The length, δ, is relevant in the fulfillment of the constraint and it influences directly on the complexity of the structure. Low values of δ suppose greater amount of members, therefore a structure with more and thinner members which ends being a more complex structure.

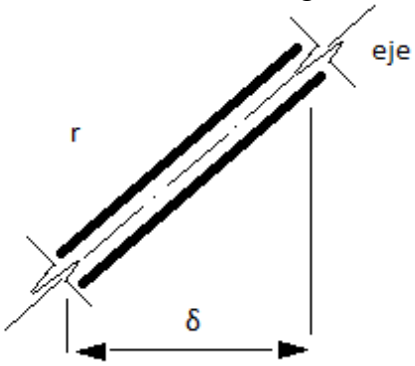


Figure 4: Length of a member

This parameter, δ, also affects the thickness of the members (high value of δ = high value of d).

2.3. Thickness, d.

The thickness, d, does not influence the complexity non the constraint, as the constraint is not mechanic but thermal.

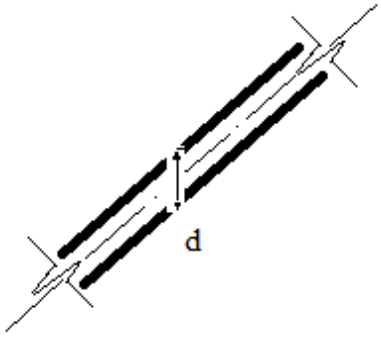


Figure 5: Thickness of a member

3. EFFECTS OF THE OPTIMIZATION PARAMETERS

In this section there are going to be shown the effects of the typical parameters of the topology optimization over the complexity of the final geometry. Considered parameters include the filter radius applied for mesh independency, objective volume and the perimeter constraint.

For filters, since their introduction, they have always been used to avoid the generation of checkerboard-like problems due to the introduction of penalization processes such as the SIMP method. Filters tend to concentrate material where a predisposition for this exists, but always inside a region controlled by the value of the filter radius, rmin. However,

concentrating material means to simplify the geometry, so the bigger the filtering radius is the simpler the geometry will became. Figure 6 shows this effect.

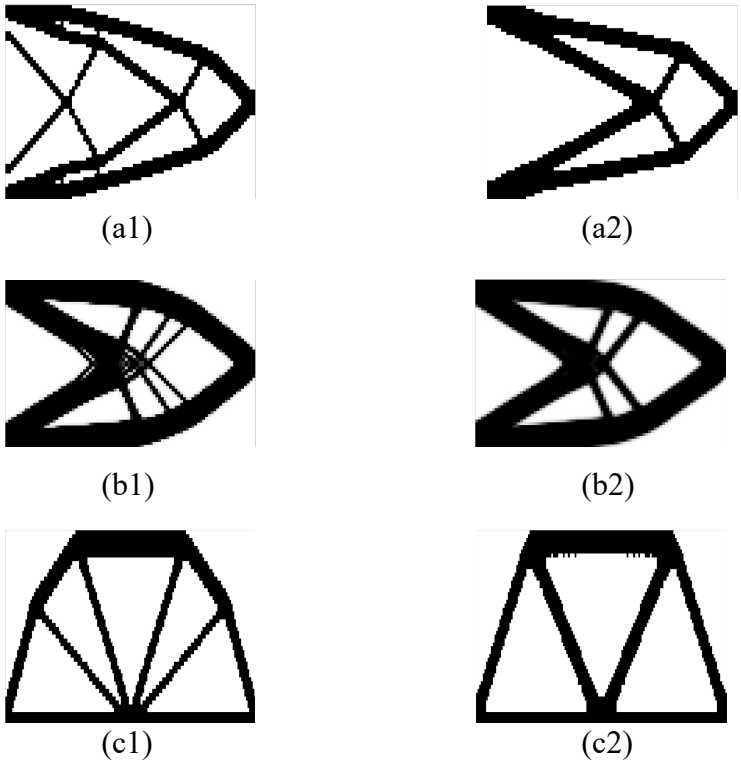


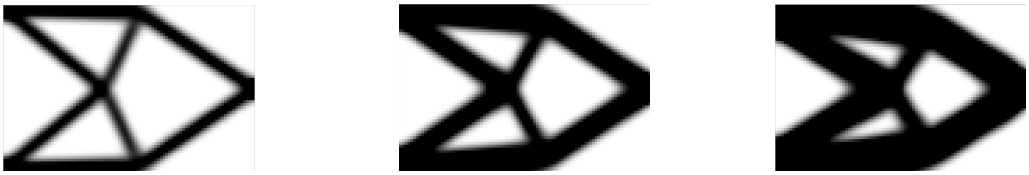
Figure 6: Effect of the filter radius on geometry's complexity. **(a)** Discrete variables. 90x60 elements. V*=0.3, (a1) rmin=1.5, (a2) rmin=2.5. **(b)** Continuous variables. 90x60 elements. V*=0.5, (b1) rmin=1.05 and (b2) rmin=1.25. **(c)** Discrete variables. 150x50 elements. V*=0.3, (c1) rmin=1.5, (c2) rmin=3.5.

Figures (a1), (a2), (b1) and (b2) show the effects of varying the values of the filter radius on a typical cantilever beam optimization problem, for discrete and continuum variables respectively. For both cases the lower value of the filter results in a more complex geometry. From figures (a1) and (a2) in can be appreciated, as it was explained before, that the more complex geometry is more suitable (for a vertical growing direction).

In the other hand, figures (c1) and (c2), although for lower filter radius the geometry is more complex, it is not more suitable (for a vertical growing direction) as some members' relative inclination is greater than the theoretical limit of 45º.

As it can be extracted from this examples there is a clear effect of the filter radius on the final complexity of the structure, but there cannot be established a concrete pattern that could be applied on any geometry and problem. In general terms, low values of the filter radius result in more and thinner members, increasing then the complexity of the geometry but not always resulting in a more suitable geometry for additive manufacturing.

The to be accomplished volume also has a direct effect on the member inclination. The need for occupying the same space with more material supposes, in general, that the inclination of the members will vary due to the increase of member thickness. Figure 7 shows this effect.

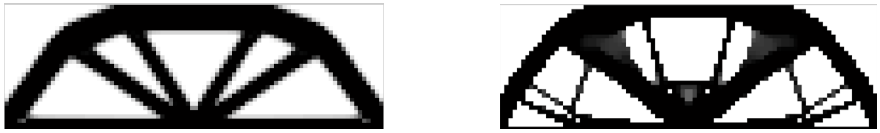


(a) (b) (c)

Figure 7: Effect of the objective volume on the member inclination. Continum variables. 90x60 elements .(a) V*=0.3, (b) V*=0.5, (c) V*=0.65. Rmin=3.5.

From (a) to (c), it can be appreciated how member inclination changes as objective volume increases. Inclination of the boundary elements is what really matters, and it is seen that the greater the final perimeter is, que greater the inclination it is.

In the case of the perimeter, its control is one of the proposed methods for controlling de complexity of the structures. This method was proposed by [3] and its initial objective was to substitute the filtering processes and to reduce/control the complexity of the final geometry. In order to do so, an upper bound was written for the perimeter, reducing this way the number of members and holes. In the the contrary is made and a lower bound is written so that a great amount of members are obtained increasing like that the final complexity of the geometry.



(a) (b)

Figure 8: Effect of the objective perimeter. Continum variables. 90x30 elements .V*=0.5, rmin=1.2. (a) P=626, (b) P=800.

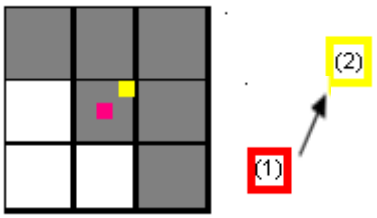
In figure 8 it can be appreciated the optimum structure (a) and the optimum structure with a greater perimeter constraint (b). It is easily seen the increase of the complexity of the geometry when increasing the perimeter, and this happens because the perimeter is the longitudinal measure of every structural boundary, therefore increasing the perimeter increases the amount of holes and members of the final geometry. That way, members that will work as natural scaffold structures can be created avoiding thermal problems and also external scaffold structures.

It can be appreciated how every of the parameters has its own effects on the complexity of a geometry, and hence, on the suitability of the structure to be manufactured by additive manufacturing technologies. Besides, the parameters are connected among them as one affects the other (i.e: the greater the objective volume is, the greater could the perimeter become and more members could be created).

4. THE PROPABILITY FOR THAT A STRUCTURE NEEDS SCAFFOLD STRUCTURES, Ω.

Now that the effect of the different parameters over the complexity of a geometry is been explained, a method for obtaining the probability for that a structure needs scaffold structures is presented, regardless the value of the different parameters. The method presented in this paper is based on discrete variables, but the aim is to develop an universal method.

The objective is to find a number that can describe in which amount the overhang constraint is violated, or what is the same, the probability for that scaffold structures are needed. The process starts calculating variable gradients for 3x3 mask placed with center on the center of every element. These gradients will be calculated in the vector **vcg**, where **vcg** is the direction vector from the geometric center of the mask to its center of gravity. This method is known as SUSAN (Smallest Univalue Segment Assimilating Nucleus) and was developed as a method for locating boundaries [4, 5].



vcg=(xcg,ycg)

$$xcg = \frac{\sum_{i=1}^{n} xi \cdot \rho i}{\sum_{i=1}^{n} \rho i} \qquad (1)$$

$$ycg = \frac{\sum_{i=1}^{n} yi \cdot \rho i}{\sum_{i=1}^{n} \rho i}$$

Figure 9: Physical meaning of vector **vcg**

4.1. Properties of vector vcg.

- The vector **vcg** always points to the material. This property is usefull for defining and distinguishing the conflictive boundaries.

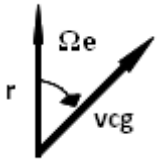
- In the case of continuum variables, the module of **vcg** will indicate if there is a boundary or not. The value of the module indicates how fast the variable varies in the direction of the vector (the maximum variation direction), hence, a high value of the module indicates that the existence of a boundary can be consider in that point.

High |vcg| = Boundary (drastic variation of the design variable)

Once **vcg** is obtained, we must know if there will be any manufacturing problem, and for that the growing direction of the structure must be defined. This direction is represented by the vector **r** which is defined by the designer.

r (designer defined) = (xr, yr) (2)

Now and for every 3x3 mask the angular deviation of **vcg** towards **r** is calculated. Summing these angels of deviation the number Ω is obtained, and every component of this number (the one of every mask) is called Ωe.



$$\Omega e = \text{deviation}(\mathbf{r}, \mathbf{vcg}) \qquad (3)$$

$$\Omega = \sum_{e=1}^{m} \Omega e \qquad (4)$$

Figure 10: Physical meaning of Ω and Ωe

where m is the number of masks. Ω represents the probability for that a structure will need scaffold structures.

A typical situation in any topology optimization problem can be described as follows where problematic boundaries are painted in yellow.

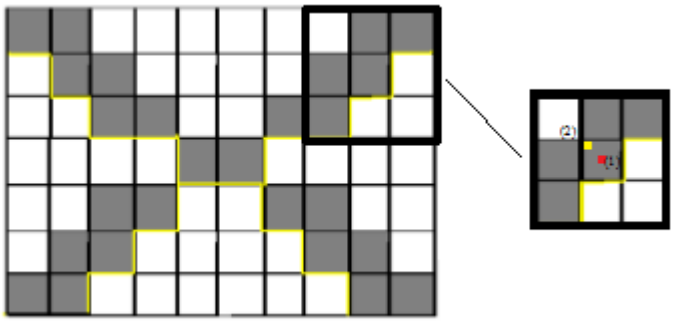


Figure 11: A typical situation in any topology optimization problem

In order to know if a boundary is critic we will use one of the components of its vector **vcg** and compare it with the correspondent of **r**. In this case we use the component ycg. As the vector always points the center of gravity, if the component ycg has the same sign as the component yr of vector **r**, that boundary will be conflictive.

$$ycg/yr < 0 \text{ -------- No problem}$$

$$ycg/yr \geq 0 \text{ -------- Conflictive boundary} \quad (5)$$

Then, in order to safe efforts, only in the case where $ycg/yr \geq 0$ we will proceed to calculate the deviation between the vectors **vcg** and **r**.

However, despite of being in the previous situation, the constraint is only (theoretically) violated if the relative angle y inferior to 45º, so it is helpful to relativize Ωe just to sum only in the cases where the constraint is really violated. Then,

$$\Omega o = \text{deviation}(\mathbf{r},\mathbf{vcg}) - \beta \quad (6)$$

where β is a parameter that represents the limit angle (45º in theory). As it's been mentioned, the overhang constraint is also affected by the member length, so we will let the designer defining the limit value for the angle from what the constraint will be considered violated. This way,

$$\text{If } \Omega o < 0 \text{ -------- The constraint is violated. } \underline{+1}$$

$$\text{If } \Omega o \geq 0 \text{ -------- The constraint is not violated. } \underline{+0}$$

Then, and remembering (4),

$$\Omega e = \begin{cases} 0, & \Omega o \geq 0 \\ 1, & \Omega o < 0 \end{cases} \quad (7)$$

$$\Omega = \sum_{e=1}^{m} \Omega e$$

This way, Ω represents not only the probability of needing scaffold structures or the amount that the constraint is violated, but also the number of conflictive elements. A high value of Ω means that there are so much elements that cannot meet the constraint and therefore that there is a high probability of needing some scaffold structures. Nevertheless, knowing Ω does not mean that we are capable of eluding scaffold structures, but it allows the designer to know more information about the amount of necessity for them.

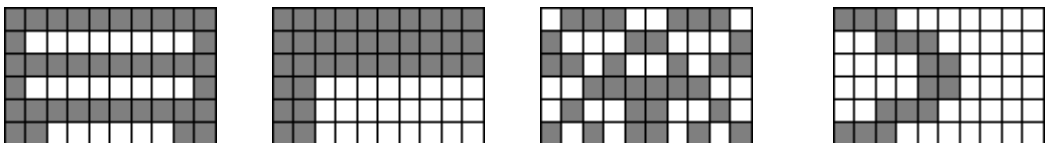
Another question is posed now and it is how to obtain a good interpretation of the number Ω. It's been explained that a high value of Ω means great violation of the constraint, but no information can be obtained beyond that. El objective of the authors is to detect and identify the boundary of every member y to be able to control their orientation, but this is presented as future work. [5] proposed a problem for minimizing the deviation Ωe, but the results didn't show any structural reliability, therefore the authors propose the development of new methods and techniques.

For having a correct interpretation of Ω a normalization parameter is needed, and that will be the perimeter, P. This way we can distinguish among cases that a priori possess similar value of Ω. For example, a structure with high value of Ω and a low value of the perimeter is a great problem due to the low amount of members and great amount of conflictive elements, but a structure with the same number Ω but high perimeter value, is a less critical situation. It is more critical a situation with a low number of holes (members) where vectors **r** and **vcg** will have the same direction in many continuum occasions creating thick members perpendicular to **r**, that one with a lot of little holes in it. For this reason, we define Ω as follows,

$$\Omega = \frac{\sum_{i=1}^{m} \Omega e}{P} \tag{8}$$

Having this information, the interpretation of Ω becomes easier and more effective.

$$\Omega \begin{cases} high \begin{cases} If\ P\ high = A\ lot\ of\ members\ but\ excesive\ constraint\ violation & (a) \\ If\ P\ low = Low\ amount\ of\ members\ and\ high\ violation & (b) \end{cases} \\ low \begin{cases} If\ P\ high = A\ lot\ of\ members\ and\ constraint\ violation, but\ can\ be\ self supporting & (c) \\ If\ P\ low = Low\ amount\ of\ members\ a\ little\ constraint\ violation\ . & (d) \end{cases} \end{cases}$$



(a) (b) (c) (d)

Figure 12: Representation of some examples

5. CONCLUSIONS

- Every typical parameter of the topology optimization problem has some effect on the complexity of the final geometry.
- Low values of the filter radius increase the complexity of the final topology, but the more complex design is not always the most suitable for additive manufacturing construction.
- Same geometry with different objective volume fraction can became non-suitable for additive manufacturing construction. As objective volume grows, member inclination tends to become more critical, specially in the boundaries.
- Objective perimeter shows the greatest effect on the complexity. Higher perimeters help creating more, shorter and thinner members. This effect can help avoiding the need for scaffold structures as the structure can become self-supporting.
- The number Ω represents the probability for that a structure will need of scaffold structures to avoid overhang constraint and be self-supporting. This information may help the designer interpreting how much the constraint is violated and evaluating the quality of the structure.
- The perimeter helps with the normalization of Ω. This parameter helps differentiating among cases with a priori the same value of Ω, giving more detailed and precise information.

REFERENCES

[1] E. L. Cardoso, J. S. O. Fonseca, *Complexity control in the topology optimization of continuum structures, J. Braz. Soc. Mech. Sci. & Eng.vol.25 no.3. Rio de Janeiro. July/Sept, 2003.*

[2] D. Brackett, I. Ashcroft, R. Hague. *Topology Optimization for Additive Manufacturing*, *August 17 2011.*

[3] C.S. Jog, M.P. Bendsoe, *A new approach to variable-topology shape design using a constraint on perimeter. Structural Optimisation 11, 1-12, Springer-Verlag 1996.*

[4] Smith S.M., Brady J.M. *SUSAN. A new Approach to Low Level Image Processing. Technical Report TR95SMS1c from Oxford Centre of Functional MAgnetic Resonance Imaging of the Brain.*

[5] Jens Fynbo, John Rasmussen, Niels Olhoff. *A Directional Topology Optimization Method. Institute of Mechanical Engineering, Aalborg University*.

[6] Andrew T. Gaynor, James K. Guest. *Topology Optimization for Additive Manufacturing: Considering Maximum Overhang Constraint. Johns Hopkins University, Baltimore, MD, 21218, USA.*

[7] M. Reuben Serphos. *Incorporating AM-specific Manufacturing Constraint into Topology Optimization. Faculty of Mechanical, Maritime and Materials Engineering. Delft University of Technology, May 30, 2014.*

[8] X.Y. Yang, Y.M. Xie, J.S. Liu, G.T. Parks, P.J. Clarkson. *Perimeter control in the bidirectional evolutionary optimization method. Structural and Multidisciplinary Optimization 24, 430-440, Springer-Verlang, 2003.*

[9] O. Sigmund. *A 99 line topology optimization code written in Matlab. Structural and Multidisciplinary Optimization 21, 120-127, Springer-Verlang, 2001.*

[10] Cristina Alonso, Osvaldo M. Querin, Rubén Ansola. *A Sequential Element Rejection and Admission (SERA) method for compliant mechanism design. Structural and Multidisciplinary Optimization 47, 795-807, Springer-Verlang, 2013*.

[11] Tomás Zegard, Glaucio H. Paulino. Bridging topology optimization and additive manufacturing. *Structural and Multidisciplinary Optimization, Springer-Verlang, 2015.*

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

FINITE ELEMENT APPROXIMATION OF A TIME-DEPENDENT TOPOLOGY OPTIMIZATION PROBLEM

M. Bruggi[1], N. Parolini[2], F. Regazzoni[2] and M. Verani[2]

[1]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano
P.zza Leonardo da Vinci 32, 20133 Milano, Italia.
e-mail: matteo.bruggi@polimi.it

[2] MOX - Dipartimento di Matematica, Politecnico di Milano
P.zza Leonardo da Vinci 32, 20133 Milano, Italia.
e-mail: {nicola.parolini,marco.verani}@polimi.it, francesco1.regazzoni@mail.polimi.it

Keywords: Topology optimization, finite element method, additive manufacturing

Abstract. *In most additive manufacturing technologies, support structures are required to sustain overhanging surfaces. These additional structures have negative effects on both processing time and material consumption. Moreover, additional post-processing effort is required for their removal. Therefore, reducing the use of the supports would have a beneficial impact on the overall manufacturing process.*

An optimization procedure aiming at the optimal placement (and design) of the supports in additive manufacturing should include the intrinsic time-dependent nature of the process. More precisely, it should take into account not only the final configuration, but all the intermediate shapes that are obtained during the additive process. As a consequence, we believe that it is necessary to go beyond standard topology optimization methods, where typically only the final shape is optimized.

The model proposed in this work relies on the solution of a time-dependent minimal compliance problem based on the classical Solid Isotropic Material with Penalization (SIMP) method. In particular, we first introduce a continuous optimization problem with the state equation defined as the time-integral of a linear elasticity problem on a space-time domain. The objective functional is given by the mean compliance over a time interval. The optimality conditions for this optimization problem are then derived and a fixed-point algorithm is introduced for the iterative computation of the optimal solution.

Numerical examples showing the differences between a standard SIMP method, which only optimizes the shape at the final time, and the proposed time-dependent approach are presented and discussed.

1 INTRODUCTION

Additive manufacturing, also known as 3D printing, is widely used to create prototypes from digital models. Successive layers of material are laid down by a three–dimensional printer under computer control. In most additive manufacturing technologies, support structures are required to sustain overhanging surfaces. Up to now, not any shape or geometry can be printed in real time, because of the need for providing a suitable set of supports when synthesizing the three–dimensional object. Support structures remarkably affect not only the processing times but also the material consumption. Therefore, their rational use would greatly improve the overall process of 3D printing.

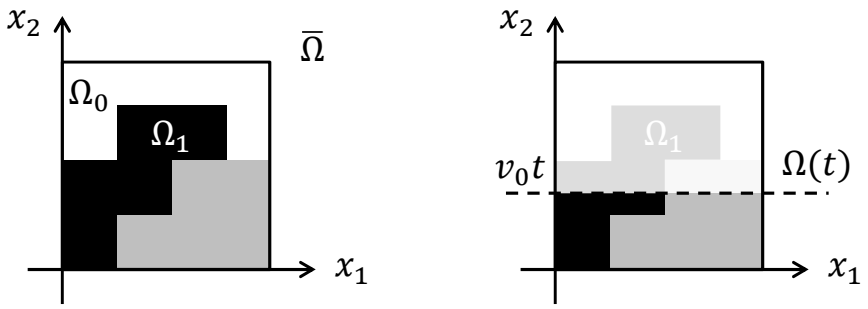
Powerful numerical tools that go under the name of methods of topology optimization are rapidly developing in many branches of engineering to find optimal layouts that maximize any kind of performance, see [1]. The classical formulation searches for the distribution of the available amount of isotropic material such that the so–called compliance (twice the elastic strain energy computed at equilibrium) is minimized. A suitable interpolation depending on the local values of the unknown density field penalizes the mechanical properties of the elastic body to achieve 0–1 solutions, see e.g. the well–known SIMP (Solid Isotropic Material with Penalization) [2]. Methods of mathematical programming are adopted to solve the arising minimization problem, see [3], generally resorting to the adoption of the finite element method to solve the equilibrium equation and compute the objective function and its sensitivity with respect to the design variables.

Additive manufacturing is a fertile area of research for topology optimization. According to the recent and comprehensive contribution in [4], additive manufacturing fills the gap between topology optimization and application, since any computed optimal design can be printed with minimal limitations on its complexity. Among the others, [5] investigates the issues and opportunities for the application of topology optimization for 3D printing, addressing the production of meso–scale structures to cope with intermediated density regions, whereas [6] copes with the multiple–material topology optimization of compliant mechanisms created via 3D printing.

Topology optimization can be used not only to generate optimal objects to be printed via additive manufacturing, but even to optimize support structures. An optimization procedure aiming at the optimal placement (and design) of the supports in additive manufacturing should include the inherent time–dependent nature of the process. More precisely, it should account not only for the final configuration, but also for *all* the intermediate shapes that are handled during the printing.

Goal of this work is to propose a new approach to this problem, resorting to the solution of a time–dependent minimal compliance formulation based on the classical SIMP. More precisely, a continuous optimization problem adopting a state equation defined as the time–integral of a linear elasticity problem on a space–time domain is formulated. The objective function is given by the time-averaged compliance, whereas the optimality conditions for this optimization problem are derived and a fixed-point algorithm is introduced for the iterative computation of the optimal solution. The discretization of the optimization problem is finally obtained by considering n intermediate time instants (t_i) (and the corresponding spatial domains $\Omega(t_i)$) and solving a sequence of linear elasticity problems on $\Omega(t_i)$ with the finite element method.

The outline of the paper is as follows. Sections 2 and 3 define the topology optimization problem to design the supports of a 3D printed object having mean minimum compliance over a time interval, in the continuous and discrete form respectively. Section 4 provides numerical examples showing the differences between a standard SIMP method, which only optimizes the



a) b)

Figure 1: a) Reference space domain. b) Space domain relative to time t. $\Omega(t)$ is located below the dotted line, corresponding to the height $v_0 t$).

shape at the final time, and the proposed time–dependent approach. Section 5 provides final remarks on the presented preliminary simulations and outlines the ongoing research.

2 THE OPTIMIZATION PROBLEM

The objective of this work is to define a topology optimization problem to design the supports $\mathcal{S}$ of a 3D printed object $\mathcal{O}$ so that the manufactured item (i.e. $\mathcal{O} \cup \mathcal{S}$) exhibits the minimum mean compliance over a given time interval $I = [0, T]$, being T the duration of the production process. To achieve this goal, we will first introduce a suitable time-dependent domain together with a linear elasticity state problem governing the displacement $\mathbf{u}(t)$ of the manufactured item at each time-instant t.

Let us consider an hold-all cylindrical space domain $\Omega = E \times (0, h) \subset \mathbb{R}^{d-1} \times \mathbb{R}$, with $d = 2, 3$ and E a subset of $\mathbb{R}^{d-1}$. Each point in Ω reads as $\mathbf{x} = (\mathbf{x}^*, y)$, where $\mathbf{x}^*$ denotes the planar component while y is the vertical one. Once the printing process is complete, i.e. for $t = T$, the target object $\mathcal{O}$ will occupy a certain subset $\Omega_1 \subset \Omega$, while for $t < T$ it will occupy intermediate configurations $\Omega_1(t)$ such that $\Omega_1(t) \subset \Omega$. In view of the above discussion, the value h represents the height of the object at the final time T. For future use, we also introduce the subdomain $\Omega_0 \subset \Omega$ identifying the region where *a priori* the user does not want to introduce any support. Next, we introduce a time-dependent domain $\Omega(t)$ that changes during the additive manufacturing process and represents the region where the 3D printer can add material (either belonging to the object or to the supports). We assume that $\Omega(t)$ grows in the direction given by the coordinate y with constant velocity v_0, i.e. $\Omega(t) = \{(\mathbf{x}^*, y) \in E \times (0, y) : 0 < y < v_0 t\}$. Accordingly, we have $\Omega_1(t) = \Omega(t) \cap \Omega_1$ (see Figure 1). Clearly, at the final time $T = h/v_0$, we have $\Omega(T) = \Omega$ and $\Omega_1(T) = \Omega_1$.

Now, we describe the topology optimization problem which will be instrumental to optimally place the supports of the target object to be printed. The first step is to introduce the design variable which describes the material distribution. As it is common in topology optimization, it is introduced as a non-dimensional density distribution $\tilde{\rho}$. More specifically, in our framework

we define the set of admissible densities as follows

$$\mathcal{U}_{ad} = \Big\{ \tilde{\rho}(\mathbf{x}) \in L^2(\Omega) \, : \tilde{\rho} \geq \tilde{\rho}_{min} \qquad \text{a.e. in } \Omega, \tag{1a}$$

$$\tilde{\rho} \leq 1 \qquad \text{a.e. in } \Omega, \tag{1b}$$

$$\tilde{\rho} = \tilde{\rho}_{min} \qquad \text{a.e. in } \Omega_0, \tag{1c}$$

$$\tilde{\rho} = 1 \qquad \text{a.e. in } \Omega_1, \tag{1d}$$

$$\int_\Omega \tilde{\rho} \, \mathrm{d}\mathbf{x} \leq C \Big\}. \tag{1e}$$

The density $\tilde{\rho}$ is bounded between $\tilde{\rho}_{min}$ and 1. The region where $\tilde{\rho} = 1$ identifies the subdomain occupied by the material (either the object to be printed or the supports), while $\tilde{\rho} = \tilde{\rho}_{min}$ identifies the empty region. Clearly, in the region Ω_1 occupied by the target object $\mathcal{O}$ the density is *a priori* fixed at one. Moreover $\tilde{\rho}$ is subject to the volume constraint $\int_\Omega \tilde{\rho} \, \mathrm{d}\mathbf{x} \leq C$, where C represents the total amount of material available for the printing process.

Let $\Gamma_D \subset \partial\Omega$ be the portion of the boundary where the object is anchored (usually the bottom of space the domain). We require that, at each time instant t, the state variable $\mathbf{u}(t)$ is the solution of the following linear elasticity problem on $\Omega(t)$ written in weak for

$$a_{\tilde{\rho}}(\mathbf{u}, \mathbf{v}; t) = l(\mathbf{v}; t) \qquad \forall \, \mathbf{v} \in H^1_{0,\Gamma_D}(\Omega) \quad \text{a.e. } t \in [0, T], \tag{2}$$

where we have introduced the bilinear form

$$a_{\tilde{\rho}}(\mathbf{u}, \mathbf{v}; t) = \int_{\Omega(t)} \mathbf{E}(\tilde{\rho}(\mathbf{x})) \nabla_s \mathbf{u} : \nabla_s \mathbf{v} \, \mathrm{d}\mathbf{x} \tag{3}$$

and the linear functional

$$l(\mathbf{v}; t) = \int_{\Omega_1(t)} \rho_0 \mathbf{g} \cdot \mathbf{v} \, \mathrm{d}\mathbf{x}, \tag{4}$$

being $\mathbf{E}$ the elasticity tensor, $\nabla_s = \frac{\nabla + \nabla^T}{2}$ the symmetric gradient, ρ_0 the mass density of the material used by the printer and $\mathbf{g}$ the gravity acceleration.

The particular choice of the linear functional l amounts to assuming that the printed object is subject to an external load represented by the weight of the target object only (i.e. we are assuming negligible the contributions of the supports).

In the original SIMP method [2], the intermediate density values are penalized through the interpolation scheme $\mathbf{E}(\tilde{\rho}(\mathbf{x})) = \tilde{\rho}(\mathbf{x})^p \mathbf{E}_0$ (where $\mathbf{E}_0$ is the elasticity tensor of the considered material). The lower bound $\tilde{\rho}_{min}$ is required to avoid singularities in the elasticity problem. A modified version of the power law approach was proposed in [8]:

$$\mathbf{E}(\tilde{\rho}(\mathbf{x})) = \tilde{E}_{\tilde{\rho}}(\mathbf{x}) \mathbf{E}_0 \quad , \tag{5}$$

$$\tilde{E}_{\tilde{\rho}}(\mathbf{x}) = \tilde{E}_{min} + \tilde{\rho}(\mathbf{x})^p (1 - \tilde{E}_{min}) \tag{6}$$

and $0 \leq \tilde{\rho} \leq 1$. In this work, we will consider a hybrid formulation, by using (6) with $0 < \tilde{\rho}_{min} \leq \tilde{\rho} \leq 1$, so that we can recover the original formulation by setting $\tilde{E}_{min} = 0$, and the modified version by setting $\tilde{\rho}_{min} = 0$.

2.1 The continuous optimization problem

Now, we are ready to introduce our optimization problem. In particular, we ask that the unknown optimal density distribution of the supports minimizes the mean compliance of the printed object over the time interval I. Thus, we consider the following objective functional:

$$\mathcal{J}(\tilde{\rho}) = \frac{1}{T}\int_0^T l(\mathbf{u(t)};t)\,\mathrm{d}t, \tag{7}$$

where, for given density distribution $\tilde{\rho}$, the function $\mathbf{u(t)}$ solves the state equation (2) at each time t. Thus, the minimum mean compliance problem reads:

$$\begin{aligned} &\min_{\tilde{\rho}\in\mathcal{U}_{ad}} \quad \mathcal{J}(\tilde{\rho}) \\ &\text{s.t.} \quad \frac{1}{T}\int_0^T \int_{\Omega(t)} \tilde{E}_{\tilde{\rho}}(\mathbf{x})\mathbf{E}_0\nabla_s\mathbf{u}(t) : \nabla_s\mathbf{v}(t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \frac{1}{T}\int_0^T l(\mathbf{v};t)\,\mathrm{d}t \quad \forall\,\mathbf{v}\in\mathcal{V} \end{aligned} \tag{8}$$

It is possible to analyze the continuous time-dependent optimization problem (8), recasting the *pointwise-in-time* formulation (2) in an equivalent *integral-in-time* formulation. In this way the evolutionary problem can be addressed by employing well known solution methods for the classical SIMP problem, such as OC, MMA and CONLIN, see [1]. The analysis will be presented in [9].

3 DISCRETIZATION OF THE PROBLEM

In this section, we introduce the discretization of the problem (2) and we derive the discrete counterpart of the optimization problem (8) that will be solved using the OC (Optimality Conditions) method.

3.1 Finite-element space discretization

At the generic time instant t, we consider a computational grid $\mathcal{T}_h(t)$ partitioning the domain $\Omega(t)$ and we denote with $X_h(t)$ the continuous linear finite element space defined on $\mathcal{T}_h(t)$. We introduce the finite element space of functions compatible with the boundary conditions:

$$V_h(t) = \left\{ \mathbf{u}_h \in X_h(t) \quad \text{s.t.} \quad \mathbf{u}_h = 0 \quad \text{on} \quad \Gamma_D \right\}. \tag{9}$$

Let $\{\boldsymbol{\varphi}_j\}_{j=1}^{N_h(t)}$ be a basis of $V_h(t)$, where $N_h(t)$ is the dimension of the space. The space discretization of state equation (2) at time t reads:

$$\begin{aligned} &\text{find} \quad \mathbf{u}_h(t) \in V_h(t) \quad \text{s.t.} \\ &\qquad a_{\tilde{\rho}}(\mathbf{u}_h,\mathbf{v}_h;t) = l(\mathbf{v}_h;t) \qquad \forall\,\mathbf{v}_h \in V_h(t). \end{aligned} \tag{10}$$

By writing $\mathbf{u}_h(t)$ as linear combination of elements of the basis

$$\mathbf{u}_h(x,t) = \sum_{j=1}^{N_h(t)} u_j(t)\boldsymbol{\varphi}_j, \tag{11}$$

we get the Galerkin approximation associated to time t:

$$\mathbf{A}_t\mathbf{U}_t = \mathbf{F}_t \tag{12}$$

where

$$[\mathbf{A}_t]_{ij} = a_{\tilde{\rho}}(\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_i;t), \qquad [\mathbf{U}_t]_j = u_j(t), \qquad [\mathbf{F}_t]_i = l(\boldsymbol{\varphi}_i;t). \tag{13}$$

3.2 Discretization in time

We consider a uniform subdivision of the time domain I and we solve the linear system (12) at the N time instants $t_n = nT/N, n = 1, \dots, N$. The collection of vectors $\{\mathbf{U}_{t_0}, \dots \mathbf{U}_{t_N}\}$ is the full discretization of the state variable $\mathbf{u}$ over the time interval I. For the sake of simplicity, we consider a structured grid with a vertical discretization such that the grid $\mathcal{T}_h$ on the full space domain Ω can be split into the N horizontal layers $E \times [v_0 t_{n-1}, v_0 t_n]$ for $n = 1, \dots N$. Under this hypothesis, given $\mathcal{T}_h$ and a basis $\{\boldsymbol{\varphi}_j\}_{j \in V}$ of the finite element space V_h defined on $\mathcal{T}_h$, we can recover $V_h(t)$ as the space generated by the subset of $\{\boldsymbol{\varphi}_j\}_{j \in V}$ of functions whose support has non-empty intersection with $\Omega(t_n)$.

When multiple layers ($N > 1$) are used in the computation of the optimal time-averaged compliance, the resulting approach will be referred to as *multi-layer approach*. Otherwise, for $N = 1$ we recover the standard minimal compliance optimization that we will refer to as *single-layer* approach.

3.3 Full discretization

Let K_e be a generic element of the mesh $\mathcal{T}_h$, with index $e \in B$. We denote with B^0 and B^1 the sets of indexes of elements belonging to Ω_0 and Ω_1, respectively. The subset of mesh elements contained in the domain $\Omega(t_n)$ is denoted as $B_n \subset B$. Moreover, let n_e indicate the value of the smallest time step such that the element K_e belongs to $\Omega(t_n)$ and let V_e be the set of indexes of degrees of freedom associated with the element e. Thus the following relations hold:

$$\begin{aligned} &\bigcup_{e \in B} K_e = \Omega, \\ &\bigcup_{e \in B_n} K_e = \Omega(t_n), \\ &e \in B_n \iff n \geq n_e. \end{aligned} \tag{14}$$

We remark that the finite element approximation ρ_h of the density is piecewise constant over the triangulation $\mathcal{T}_h$, while the state variable $\mathbf{u}$ is discretized by continuous in space (with basis $\{\boldsymbol{\varphi}_j\}_{j \in V}$) piecewise constant in time finite elements, i.e.

$$\begin{aligned} \rho_h(\mathbf{x}) &= \sum_{e \in B} \tilde{\rho}_e \mathbb{1}_{K_e}(\mathbf{x}), \\ \mathbf{u}_h(\mathbf{x}, t) &= \sum_{n=1}^{N} \sum_{j \in V} u_j^n \boldsymbol{\varphi}_j(\mathbf{x}) \mathbb{1}_{(t_{n-1}, t_n]}(t). \end{aligned} \tag{15}$$

We define

$$\begin{aligned} K_{ij}^e &= \int_{K_e} \mathbf{E}_0 \nabla_s \boldsymbol{\varphi}_i : \nabla_s \boldsymbol{\varphi}_j \, \mathrm{d}\mathbf{x}, \\ f_i^e &= \int_{K_e} \rho_0 \mathbf{g} \cdot \boldsymbol{\varphi}_i \, \mathrm{d}\mathbf{x}, \end{aligned} \tag{16}$$

and the following SIMP interpolation holds:

$$\tilde{E}(\tilde{\rho}_e) = \tilde{E}_{min} + \tilde{\rho}_e^p (1 - \tilde{E}_{min}). \tag{17}$$

Finally, the constraint (1e) reads as follows:

$$\sum_{e\in B} |K_e|\tilde{\rho}_e \leq C. \tag{18}$$

In view of the above discussion, the discrete counterpart of the minimization problem (8) reads as follows:

$$\begin{aligned}
\min_{\tilde{\rho}_e} \quad & \frac{1}{N}\sum_{n=1}^{N}\sum_{e\in B_n} \mathbb{1}_{B^1}(e) \sum_{i\in V_e} f_i^e u_i^n \\
\text{s.t.} \quad & \frac{1}{N}\sum_{n=1}^{N}\sum_{e\in B_n} \tilde{E}(\tilde{\rho}_e) \sum_{i,j\in V_e} K_{ij}^e u_i^n v_j^n \\
& \qquad = \frac{1}{N}\sum_{n=1}^{N}\sum_{e\in B_n} \mathbb{1}_{B^1}(e) \sum_{i\in V_e} f_i^e v_i^n \quad \forall \{v_i^n\}_{n=1,\dots,N}^{i\in V} \subset \mathbb{R} \\
& \sum_{e\in B} |K_e|\tilde{\rho}_e \leq C \\
& \tilde{\rho}_{min} \leq \tilde{\rho}_e \leq 1 \quad \forall e \in B \\
& \tilde{\rho}_e = \tilde{\rho}_{min} \quad \forall e \in B^0 \\
& \tilde{\rho}_e = 1 \quad \forall e \in B^1.
\end{aligned} \tag{19}$$

The Lagrangian function for the discretized problem is defined as follows:

$$\begin{aligned}
\mathcal{L} = & \frac{1}{N}\sum_{n=1}^{N}\sum_{e\in B_n} \mathbb{1}_{B^1}(e) \sum_{i\in V_e} f_i^e u_i^n - \frac{1}{N}\sum_{n=1}^{N}\sum_{e\in B_n} \left(\tilde{E}(\tilde{\rho}_e) \sum_{i,j\in V_e} K_{ij}^e u_i^n u_j^n - \mathbb{1}_{B^1}(e) \sum_{i\in V_e} f_i^e u_i^n \right) \\
& + \Lambda\left(\sum_{e\in B} |K_e|\tilde{\rho}_e - C \right) + \sum_{e\in B} \lambda_e^+ \left(\tilde{\rho}_e - 1\right) + \sum_{e\in B} \lambda_e^- \left(\tilde{\rho}_{min} - \tilde{\rho}_e\right).
\end{aligned} \tag{20}$$

where Λ is the Lagrangian multiplier for the constraint (1e), λ_e^+ for the constraint (1b) and λ_e^- for (1a). The constraints (1c) are (1d) are not plugged in the Lagrangian, but will be considered later by projecting the solution on the space $\mathcal{U}_{ad}$ in (24).

If we derive the Lagrangian with respect to the state variable we get the discrete version of adjoint equation, while deriving with respect to the design variables $\tilde{\rho}_e$ we get:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\rho}_e}(\xi) = \sum_{e\in B} \left[-\frac{1}{N}\sum_{n=n_e}^{N} \left(p\tilde{\rho}_e^{p-1}(1-\tilde{E}_{min}) \sum_{i,j\in V_e} K_{ij}^e u_i^n u_j^n \right) + \Lambda|K_e| + \lambda_e^+ - \lambda_e^- \right] \xi. \tag{21}$$

By defining

$$\Psi_e = \frac{1}{N|K_e|}\sum_{n=n_e}^{N} \left(p\tilde{\rho}_e^{p-1}(1-\tilde{E}_{min}) \sum_{i,j\in V_e} K_{ij}^e u_i^n u_j^n \right) \tag{22}$$

we get the optimality conditions for the discretized problem which reads as follows

$$\begin{cases}
\Psi_e = \Lambda & \text{if } \tilde{\rho}_{min} < \tilde{\rho}_e < 1 \\
\Psi_e \leq \Lambda & \text{if } \tilde{\rho}_e = \tilde{\rho}_{min} \\
\Psi_e \geq \Lambda & \text{if } \tilde{\rho}_e = 1.
\end{cases} \tag{23}$$

3.4 A fixed point algorithm

Based on (23), it is possible to derive an Optimality Condition (OC) method that computes a sequence $\{\tilde{\rho}_e^K\}$ approximating $\tilde{\rho}_e$ by resorting to the following fixed-point algorithm: for $K \geq 1$

$$\tilde{\rho}_e^{K+1} = \begin{cases} \tilde{\rho}_{min} & \text{if } e \in B^0 \\ 1 & \text{if } e \in B^1 \\ \max\{(1-\zeta)\tilde{\rho}_e^K, \tilde{\rho}_{min}\} & \text{if } \tilde{\rho}_e^K (Q_e^K)^\eta \leq \max\{(1-\zeta)\tilde{\rho}_e^K, \tilde{\rho}_{min}\}, e \notin (B^1 \cup B^0) \\ \min\{(1+\zeta)\tilde{\rho}_e^K, 1\} & \text{if } \tilde{\rho}_e^K (Q_e^K)^\eta \geq \min\{(1+\zeta)\tilde{\rho}_e^K, 1\}, e \notin (B^1 \cup B^0) \\ \tilde{\rho}_e^K (Q_e^K) i^\eta & \text{elsewhere} \end{cases} \tag{24}$$

where $Q_e^K = \Psi_e^K / \Lambda^K$.

Since the space integral of updated density is a continuous non decreasing function of the multiplier Λ, its updated value Λ^K can be computed by employing a bisection algorithm. The variable ζ is a move limit, and η is a tuning parameter. Both values can be adjusted to improve efficiency of the algorithm. Typical values are respectively 0.1 and 0.5. In the actual implementation some filtering procedure must be taken into account to get a well-posed problem, see e.g. [7].

3.5 Description of the algorithm

In this section, we briefly summarize the different steps required to apply the proposed methodology, from the pre-processing needed to setup the simulation to the actual optimization loop.

1. *Pre-processing*
   - Choose a cylindric reference space domain Ω and identify sub-domains Ω_0 and Ω_1.
   - Subdivide time interval $[0, T]$ in N time steps.
   - Build a spatial mesh $\mathcal{T}_h$ on Ω, fine enough to describe the geometrical details of the sub-domains Ω_0 and Ω_1 and conforming with the horizontal layers at $y = v_0 t_1, \ldots, v_0 t_N$.
   - Build the finite element space V_h with its basis functions.
   - Choose an initial design for the variables $\tilde{\rho}_e$ (for instance a uniform distribution).
2. *Optimization*
   - For each time step t_n, compute the displacement field with the current value of the design variable (see (12)).
   - Compute for each mesh element the value of Ψ_e according to (22).
   - Compute the current value of Λ_K by bisection.
   - Update the design variables $\tilde{\rho}_e$ as in (24).
   - Repeal until a stopping criterion is satisfied.

4 NUMERICAL RESULTS

In this section we present some numerical test cases to assess the properties of the proposed topology optimization scheme for the design of support structures. In particular, we will highlight the differences between the solution obtained with the proposed multi-layer approach minimizing the mean compliance during the printing process and the one obtained from a standard single-layer minimal compliance problem for the whole object.

4.1 Test case 1

In the first test case we consider a self-supporting structure defined by an inclined tapered beam with the thinner extreme joint to the ground and the thicker extreme supported by a vertical pillar (in black in Fig 2). The domain is $\Omega = (0,1) \times (0,0.5)$ and the prescribed active region $\Omega \setminus \Omega_0 \cup \Omega_1$ (where the supports can be placed) is the area below the structure (in grey in Fig. 2). Both the object and the supports are made of thermoplastic polyurethane with $\rho_0 = 1.1 \cdot 10^3$ Kg/m^3 and $E = 3 \cdot 10^7$ Pa. The optimization was performed on three different mesh resolutions ($50 \times 25, 100 \times 50, 200 \times 100$) using a sensitivity filter with a fixed filter size of 0.2 (independent of the mesh size). The results obtained using a standard single-layer approach for the final object shape are reported in Fig. 3 (left), while the results of the proposed multi-layer using 25 time intervals are reported in Fig. 3 (right).

First, we note that in both approaches the main features of the solution are captured even with a coarse grid and the solution is robust as the space resolution is increased. However, position and size of the supports generated by the two approaches are different. In particular, for the multi-layer approach the supports are shifted on the right and thickened in order to reduce the compliance in the first phases of the printing process when the tapered beam is not yet self-supporting.

In order to better understand the reasons behind the difference between the optimal solutions obtained with the two approaches, we have evaluated, for the two configurations, the history of the compliance during the printing process by solving a sequence of 50 linear elasticity problems, each one at a different time instant belonging to the time interval I. The results presented in Fig. 4 clearly show that the shape optimized using the multi-layer approach guarantees a significant reduction of the compliance over a large portion of the time interval considered. Obviously, the value of the compliance at the final time is lower for the single-layer approach since in this case the compliance at $t = T$ is exactly the functional that is minimized.

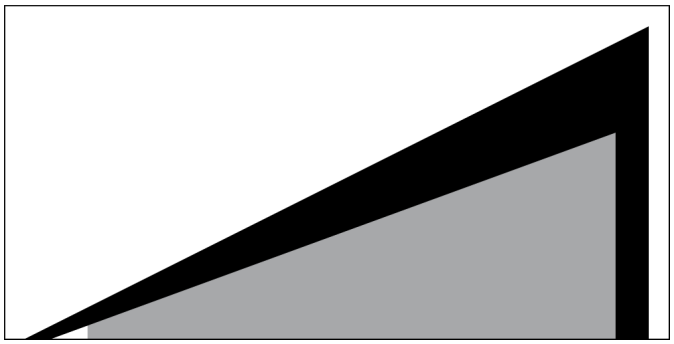


Figure 2: Computational domain for test case 1

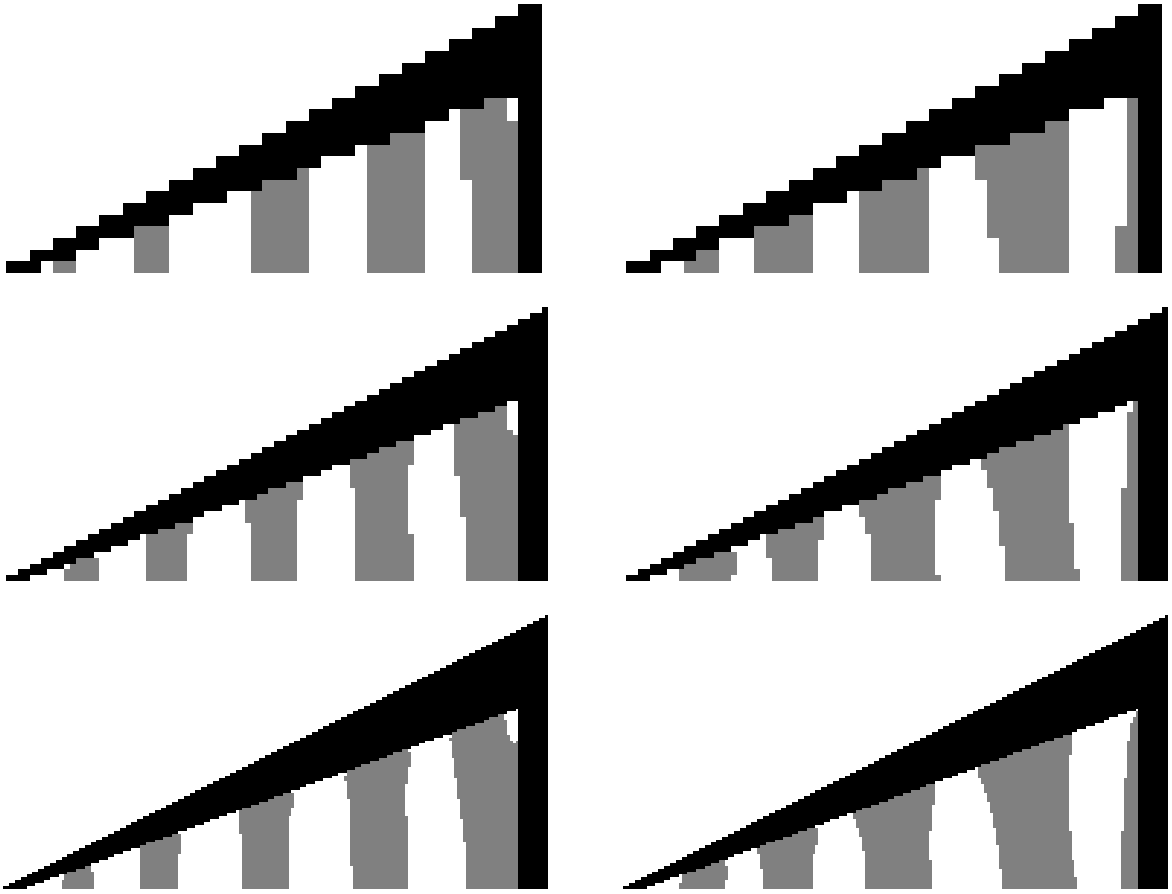


Figure 3: Results of test case 1: single-layer minimization (left), multi-layer minimization (right). Increasing mesh resolution from top to bottom.

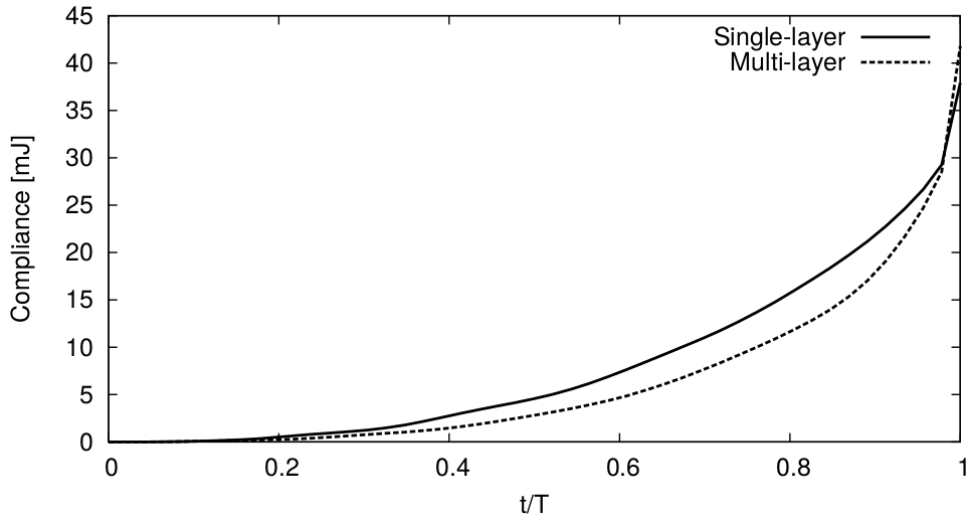


Figure 4: Time-evolution of the compliance for the two configurations obtained by the single-layer (solid line) and multi-layer (dashed-line) approaches in test case 1.

4.2 Test case 2

In the second test case, with the aim of precisely identifying the conditions under which our novel approach is more appropriate than the standard one, we continue the analysis of the differences between the single-layer compliance minimization and the multi-layer mean compliance minimization (on 25 time instants). We consider the two structures (squared and rounded) displayed in black in Fig. 5 which are fixed on the ground. The domain is $\Omega = (0,1) \times (0,0.25)$ and the prescribed active region where the support can be added is identified by the grey region below the structure. The material properties and filtering procedure are the same as for test case 1.

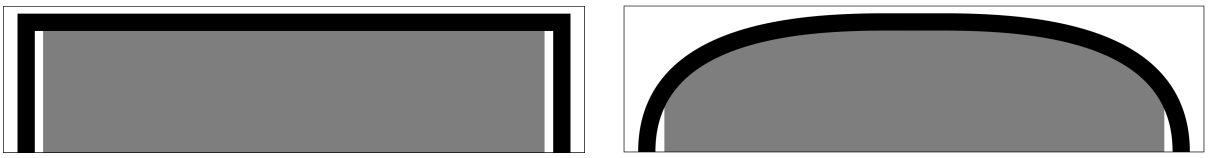


Figure 5: Computational domains for test case 2: squared structure (left) and rounded structure (right)

Let us first consider the squared structure and compare the results of the support optimization obtained by using the single-layer or the multi-layer approach (see Fig. 6). Differently from the previous test case, here the optimal supports obtained by the two approaches are almost identical. This is not surprising, due to the particular geometry of the squared structure. Indeed, only the last few layers give a significant contribution to the mean compliance, so that the functionals to be minimized by the two approaches are quite similar.

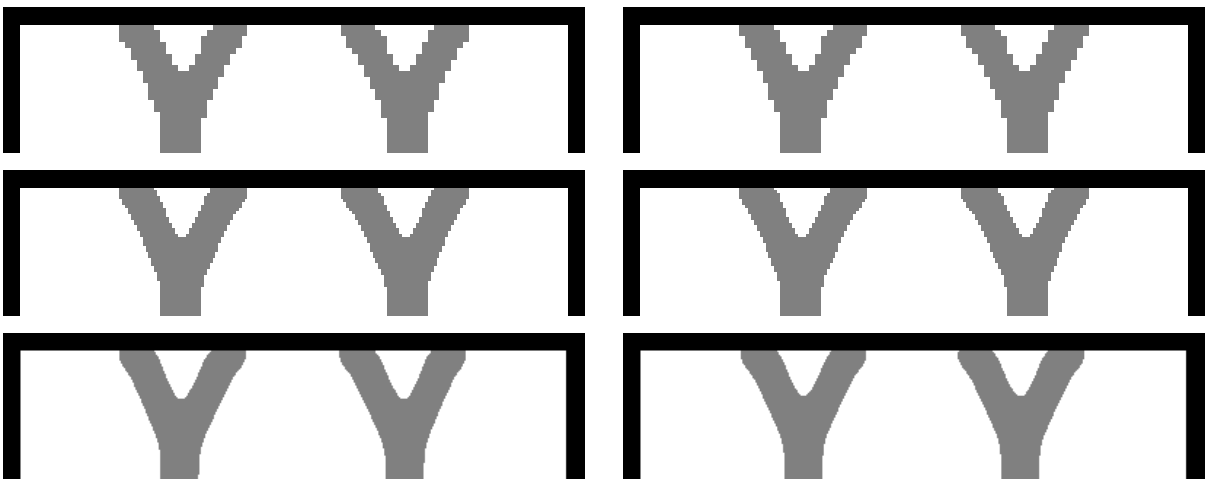


Figure 6: Results of test case 2 (squared structure): single-layer minimization (left), multi-layer minimization (right). Increasing mesh resolution from top to bottom.

Remarkably, when the rounded structure is considered, the two approaches lead to optimal solutions which are very different as displayed in Fig. 7. Here a double tree-like structure for the single-layer approach and a pillar structure for the multi-layer approach are obtained. In the latter case, at the early stages of the printing process, two overhanging structures need to be supported and this fact drives the optimization towards a structure composed by six supporting structures distributed non-uniformly along the domain width.

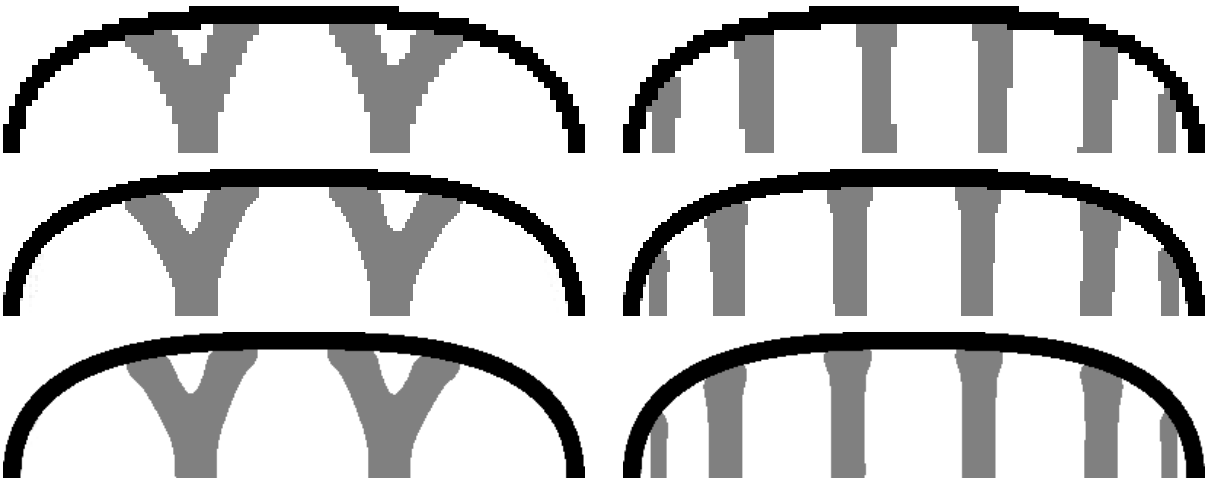


Figure 7: Results of test case 2 (rounded structure): single-layer minimization (left), multi-layer minimization (right). Increasing mesh resolution from top to bottom.

5 CONCLUSIONS

We have presented a new topology optimization algorithm for time-dependent linear elasticity problems. The algorithm has been devised for facing the time-dependent nature of additive manufacturing processes, with the aim of computing the optimal material distribution which minimizes the mean compliance over a time interval.

The algorithm has been derived applying the Optimality Condition framework to the time-dependent elasticity problem leading to the introduction of a fixed-point iterative scheme.

Numerical tests have been performed with the aim of investigating the properties of the proposed approach. Moreover, a detailed comparison with the standard minimal compliance optimization has been carried out. The results show that our method is reliable in handling situations typical in additive manufacturing where overhangs appear in the early stages of the printing process aiming at the production of self-sustained structures.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Moxoff s.r.l. for pointing out this interesting topic and for the partial support to this research project.

REFERENCES

[1] M. P. Bendsøe, O. Sigmund, *Topology optimization theory, methods and applications*, New York, Springer, 2003.

[2] M. P. Bendsøe, *Optimal shape design as a material distribution problem*, Structural Optimization, **1** (1989), 193–202.

[3] K. Svanberg, *Method of moving asymptotes - a new method for structural optimization*, International Journal for Numerical Methods in Engineering, **24** (1897), 193–202.

[4] T. Zegard, G. H. Paulino, *Bridging topology optimization and additive manufacturing*, Structural and Multidisciplinary Optimization, **53**(1) (2016), 175–192.

[5] R. Rezaie, M. Badrossamay, A. Ghaei, H. Moosavi, *Topology optimization for fused deposition modeling process*, Paper presented at the Procedia CIRP, **6** (2013), 521-526.

[6] A. T. Gaynor, N. A. Meisel, C. B. Williams, J. K. Guest, *Multiple–material topology optimization of compliant mechanisms created via PolyJet three-dimensional printing*, Journal of Manufacturing Science and Engineering, Transactions of the ASME, **136**(6) (2014), 1–10.

[7] O. Sigmund O, J. Petersson, *Numerical instabilities in topology optimization: a survey on procedures dealing with checkerboards, mesh-dependencies and local minima*, Structural Optimization, **16** (1998), 68–75.

[8] O. Sigmund, *Morphology-based black and white filters for topology optimization*, Structural and Multidisciplinary Optimization **33** (4-5) (2007), 401–424.

[9] M. Bruggi, N. Parolini, F. Regazzoni, M. Verani, *Time-dependent topology optimization: an application to additive manufacturing processes*. In preparation.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DESIGNING MANUFACTURABLE VISCOELASTIC DEVICES USING A TOPOLOGY OPTIMIZATION APPROACH WITHIN A TRULY-MIXED FEM FRAMEWORK

Paolo Venini[1], **Marco Pingaro**[2], **and Carlo Cinquini**[1]

[1]University of Pavia, Department of Civil Engineering and Architecture
via Ferrata, 3 - 27100 Pavia, ITALY
e-mail: {paolo.venini,carlo.cinquini}@unipv.it

[2]IUSS - Istituto Universitario di Studi Superiori di Pavia
Palazzo del Broletto - Piazza della Vittoria n.15, 27100 Pavia, ITALY
e-mail: marco.pingaro@iusspavia.it

Keywords: H_∞–norm topology optimization, Viscoelasticity, Truly–mixed FEM, Additive manufacturing.

Abstract. *A new approach to topology optimization is presented that is based on the minimization of the input/output transfer function H_∞ norm. Additionally, by properly selecting input and output vector, the approach is recognized to minimize an entirely new definition of frequency–based dynamic compliance. The method is applied to viscoelastic systems in plane strain conditions that are investigated by using the Arnold–Winther finite–element resorting to a generalized solid phenomenological model. Preliminary indications on how to address the actual manufacturability of the optimal specimen are eventually outlined.*

1 INTRODUCTION

Research on topology optimization of dynamic response dates back to the nineties and was at that time mostly intended as eigenvalue optimization [1] within a general max–min strategy that aims at maximizing the lowest eigenvalues and the like. The maximization of single and multiple eigenfrequencies and relevant band–gaps for vibrating structures is addressed in [2] whereas a similar problem for Euler–Bernoulli beams has been proposed in [3]. More recently methods for the optimal design of forced beams and continua in dynamic regime have been presented that may be roughly grouped into two main categories: approaches that make use of the dynamic compliance concept, see [4],[5] and [6] among others, and frequency domain methods that are investigated in [7], [8] and [9] just to mention a few.
Focus of this paper is on the proposal of a new frequency–based topology optimization strategy for dynamic structures that aims at the minimization of the input/output transfer function H_∞–norm. The distinctive feature of the proposed approach is that it seems to represent the natural extension to dynamics of by now classical methods for static topology optimization, [10]. Furthermore, by a proper selection of system input and output vectors, a new concept of frequency–domain compliance for dynamic systems is gathered and ready for topology optimization. Crucial for the success of the proposed methodology is the development of a semi–analytic formula for computing the gradient of the H_∞ norm with respect to the design parameter vector (element densities) that represents an extension of the formula presented in [11]. The proposed topology optimization strategy is quite general and may be actually applied to any kind of dynamical system in state–space (or equivalently transfer–function) form such as standard elastic systems whose state–space vector encompasses displacements and velocities. However, for this paper sake a new class of linear 2D viscoelastic systems is considered along the path suggested in [12] that uses a weak–symmetry stress finite–element as opposed to what is done herein where the truly–element proposed in [13] is adopted. By using velocities (instead of displacements) as kinematic variables, the state vector of the resulting dynamical system includes velocities themselves and stresses and this makes the numerical analysis peculiar with respect to more classical elastic idealizations. The same class of viscoelastic systems was investigated in [14] wherein the the focus of the design was on stress–constraints and incompressibility.
Eventually, a few considerations on the extensions needed by the formulation to comply with additive manufacturing requirements are highlighted, mainly following the general framework proposed by Sigmund and coworkers in [15] and [16]. As to optimal design of viscoelastic materials exhibiting extreme properties, reference is made to [17]. Numerical investigations on minimum compliance problems are presented and discussed to validate the general framework.

2 A NEW H_∞-NORM FRAMEWORK FOR TOPOLOGY OPTIMIZATION

2.1 (Descriptor) state space (D)SS and transfer function TF repsentations

The class of dynamical systems to be investigated herein may be given the time domain descriptor state-space format

$$\left\{ \begin{array}{rcl} \boldsymbol{E}\dot{\boldsymbol{x}} & = & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{w} \\ \boldsymbol{z} & = & \boldsymbol{C}\boldsymbol{x} \end{array} \right. , \tag{1}$$

or Laplace domain, transfer–function representation

$$\boldsymbol{Z} = \boldsymbol{G}(s)\boldsymbol{W}, \text{ where: } \boldsymbol{G}(s) = \boldsymbol{C}(s\boldsymbol{E} - \boldsymbol{A})^{-1}\boldsymbol{B}, \tag{2}$$

where $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{w}$ are the state, output and load vectors respectively, $\boldsymbol{E}$ is (an extended version of) the mass matrix, $\boldsymbol{A}$ is the structural state matrix, $\boldsymbol{B}$ a topological matrix distributing the loads to the degrees–of–freedom, $\boldsymbol{C}$ the output matrix that basically selects the components of the state vector whose response need be optimized, and $\boldsymbol{Z}$ and $\boldsymbol{W}$ the Laplace transforms of $\boldsymbol{z}$ and $\boldsymbol{w}$, respectively. It is assumed that matrices $\boldsymbol{E}$ and $\boldsymbol{A}$ depend on a design-variable vector $\boldsymbol{p}$, i.e. in explicit form one writes

$$\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{p}), \quad \boldsymbol{A} = \boldsymbol{A}(\boldsymbol{p}). \tag{3}$$

2.2 H_∞ norm definition and physical meaning

A systematic exposition of the H_∞ norm concept is beyond the scopes of this paper and reference is made to [18] for a in-depth treatment including applications to robust control of uncertain dynamical systems. The transfer–function representation of Equation 2 is referred to but all the reasoning applies to the state–space form of Equation 1 as well. To start with we consider the simpler case of a single–input/single–output system (that may be however be characterized by a multidimensional state vector $\boldsymbol{x}$). The H_∞ norm of a single–input/single–output (SISO) linear system with given transfer function $\boldsymbol{G}(s)$ is defined as the peak gain of the frequency response, i.e.

$$||\boldsymbol{G}||_\infty = \sup_\omega |\boldsymbol{G}(i\omega)|. \tag{4}$$

Recalling that $|\boldsymbol{G}(i\omega)|$ is the factor by which the amplitude of a sinusoidal input with angular frequency ω is magnified by the system, it is seen that the H_∞ norm is a measure of the largest factor by which any sinusoid is magnified by the system. In this respect the H_∞ norm of a dynamical system represents the natural extension to dynamics of static response measures in that it gives the analyst the maximum system response all over the frequency range in the spirit of Fourier analysis. Switching now to multi–input/multi–output (MIMO) systems, the H_∞ norm is the peak gain across all input/output channels. For an $n \times m$ transfer function matrix $\boldsymbol{G}(s)$, a natural way to achieve this is to introduce the maximum gain of $\boldsymbol{G}(i\omega)$ at the frequency ω. Let $\boldsymbol{b} = [v_1, \ldots, v_m]^T \in C^m$ be a complex–valued vector with Euclidean norm

$$||\boldsymbol{b}||_2 = \left(|v_1|^2 + \cdots + |v_m|^2\right)^{1/2}. \tag{5}$$

The maximum gain of $\boldsymbol{G}$ at frequency ω may be written as

$$\begin{aligned} ||\boldsymbol{G}(i\omega)|| &= \max_{\boldsymbol{b}} \left\{ \frac{||\boldsymbol{G}(i\omega)\boldsymbol{b}||_2}{||\boldsymbol{b}||_2} : \boldsymbol{b} \neq \boldsymbol{0}, \boldsymbol{b} \in C^m \right\} \\ &= \max_{\boldsymbol{b}} \{ ||\boldsymbol{G}(i\omega)\boldsymbol{b}||_2 : ||\boldsymbol{b}||_2 = 1, \boldsymbol{b} \in C^m \}. \end{aligned} \tag{6}$$

Finally, the H_∞ norm of the transfer function matrix $\boldsymbol{G}(s)$ is defined as

$$||\boldsymbol{G}||_\infty = \sup_\omega ||\boldsymbol{G}(i\omega)||. \tag{7}$$

One may further show that the matrix norm $||\boldsymbol{G}(i\omega)||$ is equal to the *maximum singular value* $\overline{\sigma}(\boldsymbol{G}(i\omega))$ of the matrix $\boldsymbol{G}(i\omega)$. Therefore, the H_∞ norm may also be written as

$$||\boldsymbol{G}||_\infty = \sup_\omega \overline{\sigma}(\boldsymbol{G}(i\omega)). \tag{8}$$

A few approaches for computing the H_∞–norm of a dynamical system are available in the literature, see e.g. [18], whose analysis is beyond the scopes of this paper. It suffices here to say that a few methods have been tested within the MATLAB [19] based code developed for the numerical simulations to be presented next that displayed comparable performance in terms of accuracy and speed. The MATLAB functions `getPeakGain` and `hinfnorm` have been actually used that, for a given dynamical system saved in descriptor state-space (`dss`) or transfer–function (`tf`) formats, return its H_∞ norm and the frequency ω at which the peak gain takes place that are needed for the computation of the relevant gradient vector $\frac{\partial ||\boldsymbol{G}||_\infty}{\partial \boldsymbol{p}}$.

2.3 The abstract topology optimization problem

The abstract topology optimization problem dealt with hereinafter may therefore be written as

$$\begin{cases} \min\limits_{\boldsymbol{p}} F(\boldsymbol{p}) & = & ||\boldsymbol{G}(i\omega,\boldsymbol{p})||_\infty \\ \text{s.t.} & & \boldsymbol{G}(i\omega,\boldsymbol{p}) = \boldsymbol{C}(i\omega\boldsymbol{E}(\boldsymbol{p}) - \boldsymbol{A}(\boldsymbol{p}))^{-1}\boldsymbol{B} \\ & & V(\boldsymbol{p}) \leq V_{\max} \\ & & 0 \leq \boldsymbol{p} \leq 1 \end{cases} \tag{9}$$

where 9_3 is a global volume constraint and 9_4 is the usual limitation on the element densities $\boldsymbol{p}$. As to practical topology optimization problems that fit the above framework the following are worth mentioning:

1. minimization of the displacement response at selected points of the structure. With reference to Equation 1_2, this may be achieved by selecting a Boolean $\boldsymbol{C}$ matrix whose unitary entries are such that the product $\boldsymbol{Cx}$ extracts from the state vector $\boldsymbol{x}$ those only components of the response the designer wants to be minimized;

2. minimization of the stress response at selected points of the structure. When using a mixed finite element approximation method, the goal is achieved with the very same technique as for displacement response since stresses $\boldsymbol{\sigma}$ belong to the state vector $\boldsymbol{x}$ as well. Furthermore, when a mixed finite element approach is adopted the output vector $\boldsymbol{z}$ may include both displacements and stresses so as to end up with an optimal structure with respect to kinematic and static quantities at the same time;

3. minimization of (a new concept of) dynamic compliance. Dynamic compliance has recently received much attention by the engineering community [5]. The classical static concept of compliance is typically extended to dynamics by choosing an objective functions that is the integral average over a finite time interval of the product of the loads times dual (in virtual work sense) displacements. In the spirit of the proposed approach, a new concept of dynamic compliance is introduced herein that is suitable for numerical investigations. Reference is again made to Equation 1_2 that provides the system output vector in the form $\boldsymbol{z} = \boldsymbol{Cx}$. By considering a single–output system (in which case $\boldsymbol{C}$ is a single–line vector) whose entries are the load intensities F_i positioned at the same entries as the dual displacements u_i, one gets a single–output of type $\boldsymbol{z} = \boldsymbol{Cx} = \sum_i F_i u_i$, i.e. the system dynamic compliance $\mathscr{C}$. By this selection of matrix $\boldsymbol{C}$ one may then get a topologically optimal structure that minimizes the system compliance over the entire frequency range of the acting loads.

2.4 Computing the gradient $\dfrac{\partial ||\boldsymbol{G}(i\omega,\boldsymbol{p})||_\infty}{\partial \boldsymbol{p}}$

Computing the H_∞–norm sensitivity of a transfer function matrix with respect to a plant parameter may be shown to be quite an hard task that, at least in principle, may be accomplished via a finite–difference approximation as long as the system does not depend on too many parameters, as is unfortunately the case for a structure undergoing topology optimization. To make the optimization procedure feasible, a fast and reliable semi–analytical approach is therefore necessary for the numerical procedure to convergence (herein we use the Method of Moving Asymptotes [20]). Not many algorithms are available in the literature and [11] seems to represent the only contribution on this delicate subject. First of all, the transfer function matrix in Equation (2) is re-written by posing $s = j\omega$, i.e.

$$\boldsymbol{G}(i\omega,\boldsymbol{p}) = \boldsymbol{C}\left(i\omega\boldsymbol{E}(\boldsymbol{p}) - \boldsymbol{A}(\boldsymbol{p})\right)^{-1}\boldsymbol{B}. \tag{10}$$

If now $\boldsymbol{M}(\boldsymbol{p})$ is any differentiable and invertible matrix, a basic result in matrix differential calculus allows one to write

$$\frac{\partial \boldsymbol{M}^{-1}}{\partial p_j} = -\boldsymbol{M}^{-1}\frac{\partial \boldsymbol{M}}{\partial p_j}\boldsymbol{M}^{-1}, \tag{11}$$

that, once applied to $\left(i\omega\boldsymbol{E}(\boldsymbol{p}) - \boldsymbol{A}(\boldsymbol{p})\right)^{-1}$ in Equation 10, gives

$$\frac{\partial \boldsymbol{G}(i\omega,\boldsymbol{p})}{\partial p_j} \;=\; -\boldsymbol{C}\left(i\omega\boldsymbol{E}-\boldsymbol{A}\right)^{-1}\left[i\omega\frac{\partial \boldsymbol{E}(\boldsymbol{p})}{\partial p_j} - \frac{\partial \boldsymbol{A}(\boldsymbol{p})}{\partial p_j}\right]\left(i\omega\boldsymbol{E}-\boldsymbol{A}\right)^{-1}\boldsymbol{B}\ . \tag{12}$$

A singular value decomposition of $\boldsymbol{G}$ need then be computed that reads

$$\boldsymbol{G}_{k\times m} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H \tag{13}$$

where H denotes the conjugate transpose, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices, i.e. $\boldsymbol{U}^H\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}^H\boldsymbol{V} = \boldsymbol{I}$. The diagonal elements of $\boldsymbol{\Sigma}$ are called singular values of $\boldsymbol{G}$ whereas column vectors $\boldsymbol{U}_i \in \boldsymbol{U}$ and $\boldsymbol{V}_i \in \boldsymbol{V}$ are respectively called left and right singular vectors of $\boldsymbol{G}$ associated to the singular value $\sigma_i \in \boldsymbol{\Sigma}$:

$$\begin{array}{rcl} \boldsymbol{\Sigma} & = & \text{diag}(\sigma_1,\sigma_2,\ldots,\sigma_\ell),\ \text{singular values}\ ,\ell = \min(k,m) \\ \boldsymbol{U} & = & \text{unitary matrix of left singular vectors} \\ \boldsymbol{V} & = & \text{unitary matrix of right singular vectors} \end{array} \tag{14}$$

The main result arrived at in [11] (actually in the case of standard state–space system for which $\boldsymbol{E} \equiv \boldsymbol{I}$) is a rigorous proof of the sensitivity formula that may be shown to be written as

$$\left.\frac{\partial ||\boldsymbol{G}(i\omega;\boldsymbol{p})||_\infty}{\partial p_i}\right|_{\boldsymbol{p}=\boldsymbol{p}_0} = \text{Real}\left[\boldsymbol{U}_1^H\frac{\partial \boldsymbol{G}(i\omega,\boldsymbol{p}_0)}{\partial p_i}\boldsymbol{V}_1\right], \tag{15}$$

where ω is the frequency at which the peak value of the transfer function matrix $\boldsymbol{G}$ is experienced, i.e. the one where the H_∞ norm is actually computed.

2.5 Computational issues on the H_∞ norm and its gradient

The computation of the H_∞ norm of a complex (MIMO) system with a massive state–vector $\boldsymbol{x}$ is quite a complicated task that is object of active research by the control community. Depending on the complexity of the system, the computation of the H_∞ norm may require from several minutes to half an hour as order of magnitude. However, within a topology optimization session having the H_∞ norm as objective function, one needs to compute the H_∞ norm (and the peak frequency) of a sequence of systems, say $\boldsymbol{G}_n(i\omega, \boldsymbol{p}_n)$ and typically the difference of the (2)norm of two consecutive design parameters $\boldsymbol{p}_n$ and $\boldsymbol{p}_{n+1}$ is quite small and (under the obvious hypothesis of continuity with respect to the parameters) such is therefore the difference between the H_∞ norm of the consecutive systems $\boldsymbol{G}_n(i\omega, \boldsymbol{p}_n)$ and $\boldsymbol{G}_{n+1}(i\omega, \boldsymbol{p}_{n+1})$ as well as that of the peak frequencies ω_n and ω_{n+1}. This suggests the following continuation method that allows a dramatic reduction of the CPU time needed to compute the H_∞ norm:

1. For the first two iterations, i.e. $n = 1, 2$, compute the exact H_∞ norms $||\boldsymbol{G}_1(i\omega, \boldsymbol{p}_1)||_\infty$, $||\boldsymbol{G}_2(i\omega, \boldsymbol{p}_2)||_\infty$ and the relevant peak frequencies ω_1 and ω_2, e.g. using the resident Matlab functions `getPeakGain` or `hinfnorm` [19];

2. for each subsequent iteration $n > 2$, set $\Delta\omega_{n+1} = \omega_n - \omega_{n-1}$ and consider the frequency interval centered on ω_n given as $I_{n+1} = (\omega_n - \kappa\Delta\omega_{n+1}, \omega_n + \kappa\Delta\omega_{n+1})$, where κ governs the amplitude of I_{n+1} ($\kappa = 5$ has been used in the simulations after a quick trial–and–error procedure);

3. Sample I_{n+1} with a regular grid of points (50 points are used in the computations to be presented next) and compute the frequency response of the system for all such points in the frequency domain. This amounts to a forward dynamic computation that may be accomplished using the Matlab function `freqresp` that is far quicker than the actual computation of $||\boldsymbol{G}_{n+1}(i\omega, \boldsymbol{p}_{n+1})||_\infty$;

4. set $||\boldsymbol{G}_{n+1}(i\omega, \boldsymbol{p}_{n+1})||_\infty$ equal to the maximum over all the sampled points of the absolute value of the frequency responses returned by `freqresp`.

It may be useful to check the procedure approximately once every 50 iterations by computing exactly the H_∞ norm and evaluate whether the approximate strategy is stable or not (in the numerical computations to be presented next, the continuation strategy proposed has proven to be always stable).

As for the computation of the H_∞–norm gradient, reference is made to [21] where a comparison between the CPU time needed to get a finite–difference approximation versus using Equation 15 is performed showing that Equation 15 allows for a CPU time reduction of approximately three orders of magnitude.

3 THE VISCOELASTIC BIDIMENSIONAL SYSTEM UNDER OPTIMIZATION

3.1 Strong form

The viscoelastic model originally introduced in [12] is considered that is based on a stress–additive decomposition as in Figure 1.

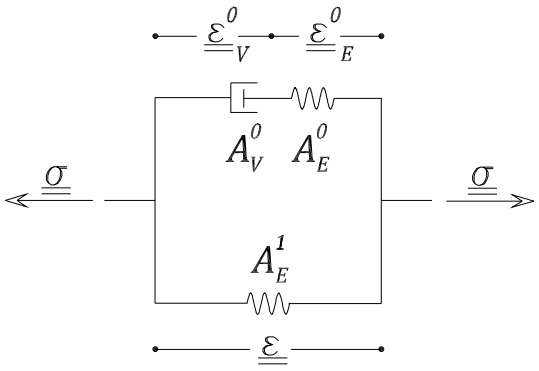


Figure 1: Standard solid phenomenological model

As usual when adopting Hellinger–Reissner variational principles, compliance tensors relating strains to stresses are introduced that allow one to write

$$\begin{cases} \underline{\underline{\underline{\underline{A}}}}^0_E \dot{\underline{\underline{\sigma}}}_0 + \underline{\underline{\underline{\underline{A}}}}^0_V \underline{\underline{\sigma}}_0 = \underline{\underline{\varepsilon}}(\underline{v}) \\ \underline{\underline{\underline{\underline{A}}}}^1_E \dot{\underline{\underline{\sigma}}}_1 = \underline{\underline{\varepsilon}}(\underline{v}) \end{cases} \tag{16}$$

where $\underline{\underline{\underline{\underline{A}}}}^0_E$ and $\underline{\underline{\underline{\underline{A}}}}^0_V$ are the elastic and viscous compliance tensors of the viscoelastic component, $\underline{\underline{\underline{\underline{A}}}}^1_E$ is the elastic compliance tensor that is in parallel with the viscoelastic one and $\underline{v}$ is the velocity field. Hereinafter plane strain conditions shall be considered allowing one to write each compliance tensor in matrix form as

$$A^{0,1}_{E,V} = \frac{1+\nu}{E^{0,1}_{E,V}(\boldsymbol{p})} \begin{bmatrix} 1-\nu & -\nu & 0 \\ -\nu & 1-\nu & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{17}$$

where, as usual in topology optimization of isotropic structures, it is assumed that the Young modulus only depends on the material density $\boldsymbol{p}$. One should notice that a stress–velocity formulation is being used that presents several advantages over more classical stress–displacement approaches, including the ease with which dynamic effects may be considered in the analysis. Therefore, compatibility relations are written in terms of strain velocities as

$$\underline{\underline{\varepsilon}}(\underline{v}) = \underline{\underline{\nabla}}^s \underline{v} = \frac{1}{2}\left(\underline{\underline{\nabla}} \underline{v} + \underline{\underline{\nabla}} \underline{v}^T\right), \tag{18}$$

whereas the dynamic equilibrium reads

$$-\rho \dot{\underline{v}} + \underline{\mathrm{div}}\, \underline{\underline{\sigma}} = -\rho \underline{g}. \tag{19}$$

3.2 Truly–mixed formulation discretized with the Arnold–Winther finite element

By observing that the total stress $\underline{\underline{\sigma}}$ may be additively decomposed as $\underline{\underline{\sigma}} = \underline{\underline{\sigma}}_0 + \underline{\underline{\sigma}}_1$, the continuous variational formulation of the problem at hand may be obtained by eliminating the strain tensor $\underline{\underline{\varepsilon}}$ in Equations 16 and 18, testing the resulting equation by two virtual stress fields

$\underline{\underline{\tau}}_0, \underline{\underline{\tau}}_1$, and the equilibrium equation 19 by a virtual velocity field $\underline{w}$, so as to write:

Find $(\underline{\underline{\sigma}}_0, \underline{\underline{\sigma}}_1, \underline{v}) \in H(\text{div}, \Omega) \times H(\text{div}, \Omega) \times L^2(\Omega)$ such that

$$\begin{cases} \langle \underline{\underline{\underline{\underline{A}}}}^0_E \dot{\underline{\underline{\sigma}}}_0, \underline{\underline{\tau}}_0 \rangle + \langle \underline{\underline{\underline{\underline{A}}}}^0_V \underline{\underline{\sigma}}_0, \underline{\underline{\tau}}_0 \rangle + \langle \underline{v}, \underline{\text{div}}\, \underline{\underline{\tau}}_0 \rangle = 0 \\ \langle \underline{\underline{\underline{\underline{A}}}}^1_E \dot{\underline{\underline{\sigma}}}_1, \underline{\underline{\tau}}_1 \rangle + \langle \underline{v}, \underline{\text{div}}\, \underline{\underline{\tau}}_1 \rangle = 0 \\ -\langle \rho \dot{\underline{v}}, \underline{w} \rangle + \langle \underline{\text{div}}\, \underline{\underline{\sigma}}_0, \underline{w} \rangle + \langle \underline{\text{div}}\, \underline{\underline{\sigma}}_1, \underline{w} \rangle = \langle \rho \underline{g}, \underline{w} \rangle \end{cases} \tag{20}$$

$\forall \underline{\underline{\tau}}_0 \in H(\text{div}, \Omega), \forall \underline{\underline{\tau}}_1 \in H(\text{div}, \Omega), \forall \underline{w} \in L^2(\Omega)$.

In more compact form, one may rewrite the governing system in Equation 20 in matrix-vector notation as usual within the framework of mixed methods, i.e.

$$\begin{bmatrix} A^0_E(\boldsymbol{p}) & 0 & 0 \\ 0 & A^1_E(\boldsymbol{p}) & 0 \\ 0 & 0 & -M(\boldsymbol{p}) \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{\sigma}}_0 \\ \boldsymbol{\sigma}_1 \\ \boldsymbol{v} \end{bmatrix} + \begin{bmatrix} A^0_V(\boldsymbol{p}) & 0 & B^T \\ 0 & 0 & B^T \\ B & B & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\sigma}_0 \\ \boldsymbol{\sigma}_1 \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \rho \boldsymbol{g} \end{bmatrix} \tag{21}$$

The triangular Arnold-Winther finite element used in this paper is the lowest–order of the family of finite elements introduced in the pioneering paper [13]. Figure 2 shows the relevant degrees of freedom that may be listed as follows. As to the stresses, one should notice that the symmetry of the stress tensor is imposed strongly so that the components to be approximated are $\sigma_{11}, \sigma_{22}, \sigma_{12}$ and one ends up with 24 degrees of freedom:

- the 3 components of the stress tensor $\sigma_{11}, \sigma_{22}, \sigma_{12}$ at each vertex of the triangle (9 dofs);
- the moments of order zero and one of the traction vector $\underline{\underline{\sigma}} \cdot \underline{n}$ along each edge of the triangle (12 dofs);
- the average of the the components of the stress tensor over the triangle, i.e. $\int_T \sigma_{11}$, $\int_T \sigma_{22}$, $\int_T \sigma_{12}$, (3 dofs).

As to the velocity vector $\underline{v}$, a standard element–wise linear globally discontinuous approximation is adopted.

3.3 Recovering the state–space and transfer function formats

The state–space or transfer–function format of Equations 1 or 2 may be recovered by introducing the state vector

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\sigma}_0 \\ \boldsymbol{\sigma}_1 \\ \boldsymbol{v} \end{bmatrix}, \tag{22}$$

and the structural matrices

$$\boldsymbol{E}(\boldsymbol{p}) = \begin{bmatrix} A^0_E(\boldsymbol{p}) & 0 & 0 \\ 0 & A^1_E(\boldsymbol{p}) & 0 \\ 0 & 0 & -M(\boldsymbol{p}) \end{bmatrix}, \ \boldsymbol{A}(\boldsymbol{p}) = - \begin{bmatrix} A^0_V(\boldsymbol{p}) & 0 & B^T \\ 0 & 0 & B^T \\ B & B & 0 \end{bmatrix}. \tag{23}$$

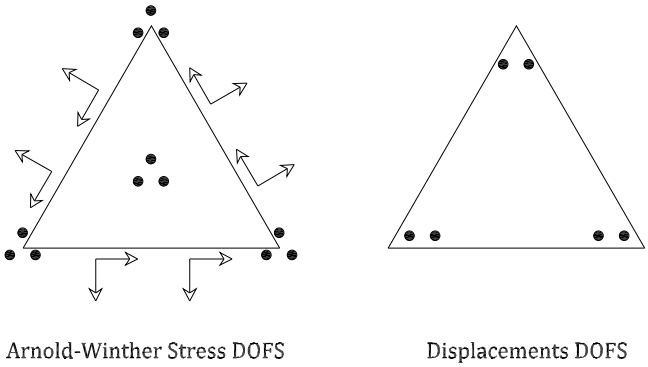


Figure 2: Arnold-Winther DOFS

As to the loading term $\boldsymbol{Bw}$, an explicit formulation should be derived on a case by case basis but a sufficiently general possibility is to pose

$$\boldsymbol{w} = \begin{bmatrix} 0 \\ 0 \\ \rho \boldsymbol{g} \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I} \end{bmatrix}. \tag{24}$$

As to the dependence of stiffness and mass density on the design variables, as usual in topology optimization, each element is given a (mathematical) density $p \in [0, 1]$ that is used within a SIMP approach [10] so as the following interpolations are considered

$$\begin{aligned} E(p) &= E_{min} + p^q (E_{full} - E_{min}) \\ \rho(p) &= \rho_{min} + p(\rho_{full} - \rho_{min}), \end{aligned} \tag{25}$$

where the exponent q is set equal to 3 as usual and Equation 25_1 applies to elastic and viscous phases of the material described by the standard model in Figure 1.

4 ADDRESSING THE MANUFACTURABILITY ISSUE

Manufacturability of the designed specimen is addressed via the three–field density approach for which reference is made to [15] for a comprehensive exposition and [16] for the physical meaning of the three involved densities within standard micro/nano lithography manufacturing processes. Many similarities of these approaches may be found with the findings of Guest and co–workers, see [22] for single–phase projection and [23] for multiple phase projection. The three density fields entering the formulation originate a chain like density transformation that reads [15]

$$p \rightsquigarrow \widetilde{p} \rightsquigarrow \widehat{\widetilde{p}}, \tag{26}$$

where p is the bounded mathematical density, $\widetilde{p}(p)$ is the filtered density typically obtained from p by a convolution–type filter (to avoid checkerboarding among other undesirable effects) and $\widehat{\widetilde{p}}$ is the projected density that allows to remove grey regions that unavoidably affect the filtered density $\widetilde{p}$ and to impose a minimum length scale to the specimen to be designed and possibly manufactured.

4.1 Density and Heaviside projection filters

A density filter is first used to switch from the mathematical density p to the intermediate density $\widetilde{p}_e$ of the generic element e that reads

$$\widetilde{p}_e = \frac{1}{\sum_{i \in N_e} H_{ei}} \sum_{i \in N_e} H_{ei} p_i, \tag{27}$$

where N_e is the set of all elements whose center–to-center distance $\Delta(e,i)$ from element e is less than the prescribed filter radius $r_{\min}$ and the weighting factor H_{ei} is defined as

$$H_{ei} = \max(0, r_{\min} - \Delta(e,i)). \tag{28}$$

For this paper sake only the solid phase is then projected. To this goal, the Heaviside projection filter

$$\widehat{\widetilde{p}}(\widetilde{p}) = \begin{cases} 1 & \text{if} \quad \widetilde{p} > 0 \\ 0 & \text{if} \quad \widetilde{p} = 0 \end{cases} \tag{29}$$

is relaxed to gain differentiability and the following filter is actually used [22, 15]

$$\widehat{\widetilde{p}} = 1 - e^{-\beta \widetilde{p}} + \widetilde{p} e^{-\beta}, \tag{30}$$

where $\beta = 0$ means no filtering and for $\beta \to \infty$ the Heaviside filter is recovered. As suggested in [23, 24], a continuation method is actually implemented that starts with a low value of β that is increased along with the iterations to impose the Heaviside filtering conditions on the final design.

4.2 The modified topology optimization problem

After the mathematical density $\boldsymbol{p}$ is filtered ($\widetilde{\boldsymbol{p}}$) and projected ($\widehat{\widetilde{\boldsymbol{p}}}$), the (formally) new H_∞ topology optimization problem reads

$$\begin{cases} \min\limits_{\boldsymbol{p}} F(\boldsymbol{p}) & = & ||\boldsymbol{G}(i\omega, \boldsymbol{p})||_\infty \\ \text{s.t.} & & \boldsymbol{G}(i\omega, \boldsymbol{p}) = \boldsymbol{C}(i\omega \boldsymbol{E}(\boldsymbol{p}) - \boldsymbol{A}(\boldsymbol{p}))^{-1} \boldsymbol{B} \\ & & V\left(\widehat{\widetilde{\boldsymbol{p}}}\right) \leq V_{\max} \\ & & 0 \leq \boldsymbol{p} \leq 1 \end{cases} \tag{31}$$

The sensitivity of the H_∞ norm with respect to the design densities $\widehat{\widetilde{\boldsymbol{p}}}$ is computed using the formula in Equation 15 whereas the chain rule is used to compute gradients with respect to the initial mathematical densities $\boldsymbol{p}$, i.e.

$$\frac{\partial F}{\partial \boldsymbol{p}} = \frac{\partial F}{\partial \widehat{\widetilde{\boldsymbol{p}}}} \frac{\partial \widehat{\widetilde{\boldsymbol{p}}}}{\partial \widetilde{\boldsymbol{p}}} \frac{\partial \widetilde{\boldsymbol{p}}}{\partial \boldsymbol{p}}, \tag{32}$$

where

$$\frac{\partial \widetilde{p}_e}{\partial p_j} = \frac{1}{\sum_{i \in N_e} H_{ei}} H_{ej}, \quad \frac{\partial \widehat{\widetilde{p}}_e}{\partial \widetilde{p}_e} = \beta e^{-\beta \widetilde{p}_e} + e^{-\beta}. \tag{33}$$

4.3 Assessing the quality of the design

The quality of the design should be assessed with respect to the following two parameters:

1. the first one is of course the value of the objective function to be minimized, i.e. the H_∞ norm for this paper sake;
2. secondly, to check the effectiveness of the proposed strategy to cope with manufacturability issues, the following gray indicator shall be evaluated [15]

$$M = \frac{\sum_{i=1}^{N} 4\widehat{\widetilde{p}}(1-\widehat{\widetilde{p}})}{N}, \tag{34}$$

where N is the dimension of vector $\widehat{\widetilde{\boldsymbol{p}}}$, i.e. the number of elements, and the lower M the better.

5 Numerical study

The bidimensional structure Figure 3 in plane strain conditions is considered with physical properties as in Table 1. A Single-Input Single-Output (SISO) realization is investigated imposing that the two loads F_A and F_B are acting simultaneously with the same (normalized) intensity whereas by properly choosing the topological matrix $\boldsymbol{C}$ (Equation 1), the output vector $\boldsymbol{z}$ (Equation 1) is given as

$$\boldsymbol{z} = \mathscr{C} = F_A v_A + F_B v_B, \tag{35}$$

i.e. the the output represents the (time derivative of the) dynamic compliance exploited all over the frequency range thanks to the H_∞–norm concept.

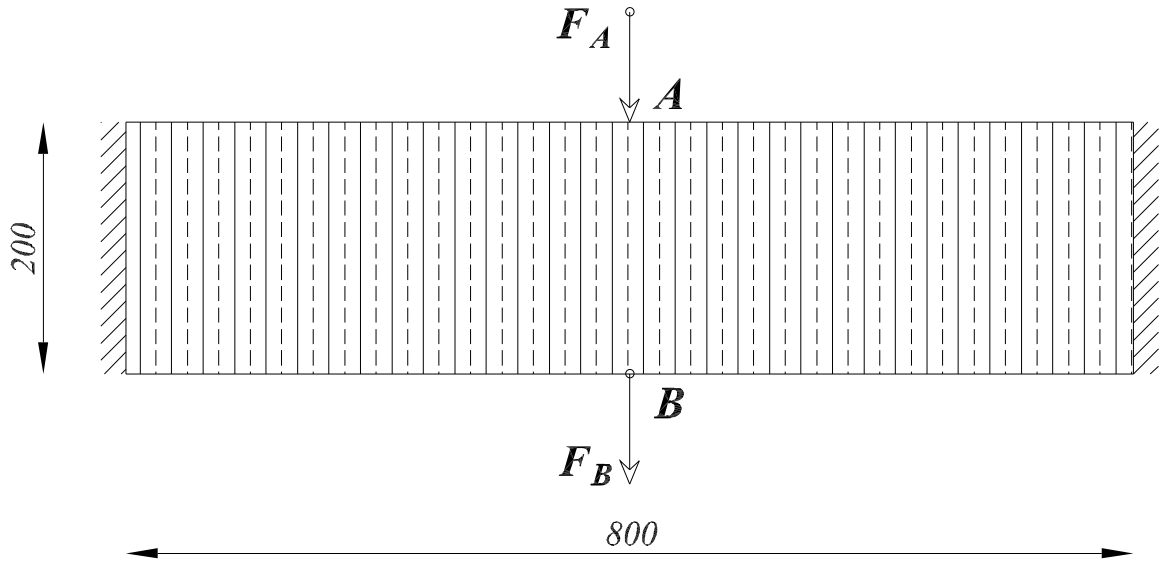


Figure 3: The bidimensional structure under optimization

Due to the symmetry of the system, only the left half is investigated by using 96×48 square cells further subdivided into 4 Arnold–Winther finite elements each. As to the overall volume

E_E^0 [Mpa]	E_V^0 [MPa s]	E_E^1 [MPa]	ρ_0 [kg/m^3]
2	0.2	2	10

Table 1: Material properties of the adopted viscoelastic material

constraint, the maximum volume $V_{\max}$ is set to 40% of the total volume. As to the continuation method to update the value of the parameter β characterizing the projection filter (Equation 30), an initial value $\beta = 2$ is considered and doubled then every 25 iterations.

Figure 4 shows the MMA performance toward the minimization of $||\boldsymbol{G}(i\omega, \boldsymbol{p})||_\infty$. Two issues are worth mentioning at this regard:

- the overall convergence path is not monotonic and remarkable oscillations take place for iterations between the 20th and the 40th;
- when the value of β is updated the objective function decreases for that very same iteration than increases and a decreasing path shows up until β is further updated. Such (smaller) oscillatory behavior should not be attributed to the MMA algorithm but to a volume constraint violation due to the fact that the projection filter is not volume–preserving. This violation could be prevented by a proper scaling of the filter but for this paper sake it was chosen to use the projection filter *as is*, and let the MMA algorithm take care of the volume constraint.

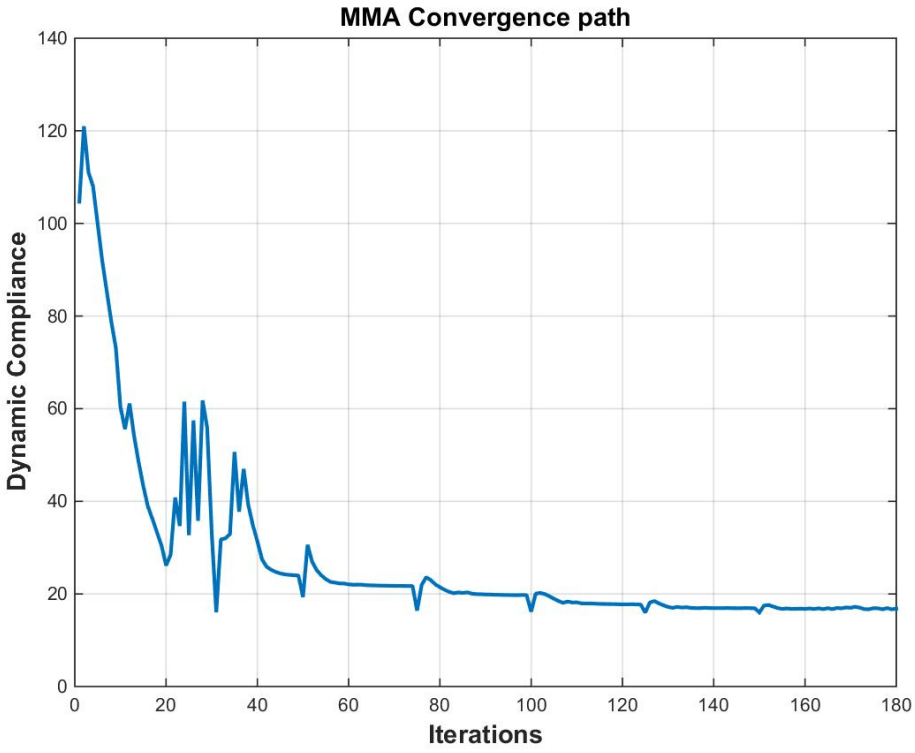


Figure 4: MMA convergence path

Figure 5 shows the optimal topology for the problem at hand (the right part was drawn by symmetry with respect to the mid vertical axis). Basically no grey regions were experienced and the grey indicator in Equation 34 was found to be 0.0075 and no sharp corners show up as to the solid phase. As to the void phase, no remedy was implemented but the adoption of a multiple phase projection strategy in the spirit of [23] is currently under development.

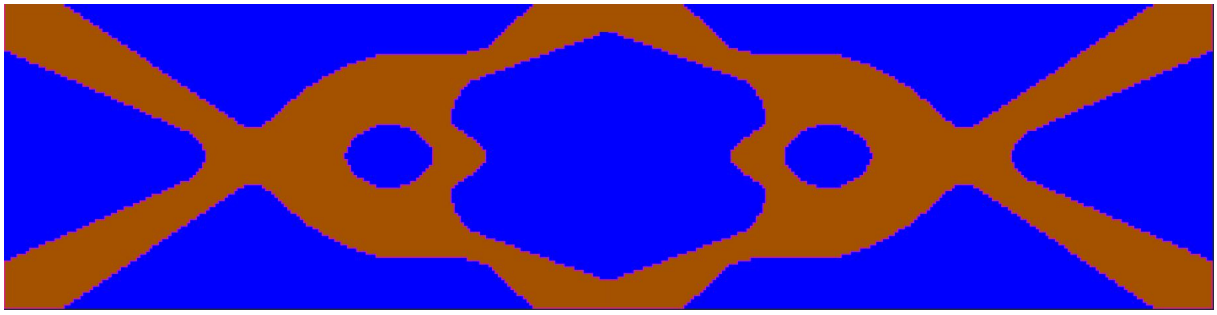


Figure 5: Optimal structure at convergence

Figure 6 shows a comparison between the frequency response functions for the initial uniform structure and the optimal topology at convergence. The peak gain, i.e. $||\mathscr{C}||_\infty$, is reduced to about one fifth of its initial value as one may also check by looking at Figure 4 that should be considered quite a remarkable performance.

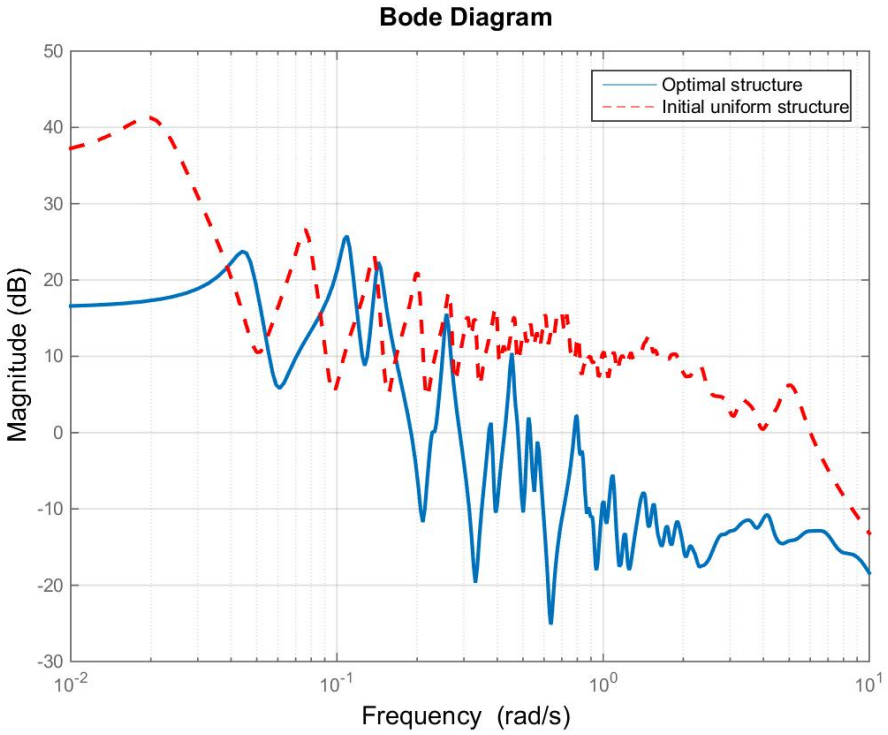


Figure 6: Frequency response functions - Optimal vs initial uniform structure

6 CONCLUSIONS

A new approach to topology optimization of dynamical systems has been introduced that is based on the minimization of the H_∞ norm of the input/output transfer function. The framework is abstract enough to include multiple loads as well as multi–output objectives with basically no modifications. Such generality is also enjoyed by the specific dynamical systems object of optimization as long as its governing equations fit a (descriptor) state–space or transfer–function format. For this paper sake a viscoelastic bi–dimensional system in plane strain conditions has been considered obeying a viscoelastic constitutive law and the truly–mixed Arnold–Winther finite element has been used for the spatial discretization. Eventually, the issue of manufacturability of the design has been addressed using the three–field density representation approach [15, 22]. Extensions to more complex geometries and load conditions are currently under development that include three–dimensional systems and multiple–phase filtering techniques [23].

7 ACKNOWLEDGMENTS

Fondazione Cariplo (grant *Safer Helmets*) is gratefully acknowledged.

REFERENCES

[1] Z.H.Ma, N. Kikuchi, H.C. Cheng, Topological design for vibrating structures. *Computer Methods in Applied Mechanics and Engineering*, **121**, 259–280, 1995.

[2] J. Du, N. Olhoff, Topological design of freely vibrating continuum structures for maximum values of simple and multiple eigenfrequencies and frequency gaps, *Structural and Multidisciplinary Optimization*, **34**, 91–110, 2007.

[3] N.Olhoff, B. Niu, G. Cheng, Optimum design of band-gap beam structures. *International Journal of Solids and Structures*, **49**, 3158–3169, 2012.

[4] J. Dahl, J. Jensen, O. Sigmund, Topology optimization of transient wave propagation problems in one dimension, *Structural and Multidisciplinary Optimization*, **36**, 585–595, 2008.

[5] J. Jensen, P. Nakshatrala, D. Tortorelli, On the consistency of adjoint sensitivity analysis for structural optimization of linear dynamic problems, *Structural and Multidisciplinary Optimization*, **49**, 831–837, 2014.

[6] M. Pingaro, P. Venini, A fast approach to analysis and optimization of viscoelastic beams, *Computers and Structures*, **168**, 46–55, 2015.

[7] Y. Yu, I.G. Jang and B.M. Kwak, Topology optimization for a frequency response and its application to a violin bridge, *Structural and Multidisciplinary Optimization*, **48**(3), 627–636, 2013.

[8] G.H. Yoon, Structural topology optimization for frequency response problem using model reduction schemes, *Computer Methods in Applied Mechanics and Engineering*, **199**(25), 1744–1763, 2010.

[9] J. Jung, J. Hyun, S. Goo, S. Wang, An efficient design sensitivity analysis using element energies for topology optimization of a frequency response problem, *Computer Methods in Applied Mechanics and Engineering*, **296**, 196–210, 2015.

[10] M. Bendsoe, O. Sigmund, *Topology optimization: Theory, methods and applications*. Springer, 2003.

[11] D.P. Giesy, K.B. Lim, H_∞ norm sensitivity formula with control system design applications, *Journal of Guidance, Control and Dynamics*, **16**(6), 1138–1145, 1993.

[12] M. Rognes, R. Winther, Mixed finite element methods for linear viscoelasticity using weak simmetry, *Mathematical Models and Methods in Applied Sciences*, **20**(6), 955–985, 2010.

[13] D. Arnold, R. Winther, Mixed finite elements for elasticity, *Numerische Mathematik*, **92**, 401–419, 2002.

[14] P. Venini, Stress constraints and incompressible materials in topology optimization: state of the art and new perspectives, J. Kruis, Y. Tsompanakis and B.H.V. Topping eds. *Computational Techniques for Civil and Structural Engineering*, Saxe-Coburg Publications, Stirlingshire, UK, Chapter 19, pp 439-465, 2015. doi:10.4203/csets.38.19

[15] B.S. Lazarov, F. Wang, O. Sigmund, Length scale and manufacturability in density-based topology optimization, *Archives of Applied Mechanics*, **86**, 189–218, 2016.

[16] M. Jansen, B.S. Lazarov, M. Schevenels, O. Sigmund, On the similarities between micro/nano lithography and topology optimization projection methods, *Structural and Multidisciplinary Optimization*, **48**, 717–730, 2013.

[17] E. Andreassen, J.S. Jensen, A practical multiscale approach for optimization of structural damping, *Structural and Multidisciplinary Optimization*, **53**, 215–224, 2016.

[18] B.A. Francis, *A Course in H_∞ Control Theory*, Lecture notes in control and information sciences, vol. 88. Springer-Verlag, 1987.

[19] The Mathworks, Inc., Natick, Massachusetts. MATLAB version 8.5.0.197613 (R2015a) 2015.

[20] K. Svanberg, The method of moving asymptotes a new method for structural optimization, *International Journal for Numerical Methods in Engineering*, **24**, 359–373, 1987.

[21] P.Venini, M. Pingaro, A new approach to optimization of viscoelastic beams: minimization of the input/output transfer function H_∞–norm, *Structural and Multidisciplinary Optimization*, under review, 2016.

[22] J.K. Guest, J.H. Prevost, T. Belytschko, Achieving minimum length scale in topology optimization using nodal design variables and projection functions, *International Journal for Numerical Methods in Engineering*, **61**, 238–254, 2004.

[23] J.K. Guest, Topology optimization with multiple phase projection, *Computer Methods in Applied Mechanics and Engineering*, **199**, 123–135, 2009.

[24] E. Andreassen, A. Clausen, M. Schevenels, B.S. Lazarov, O. Sigmund, Efficient topology optimization in MATLAB using 88 lines of code, *Structural and Multidisciplinary Optimization*, **43**, 1–16, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
E. Ramm, O. Sigmund, P. Duysinx, W. K. Liu
Crete Island, Greece, 5–10 June 2016

ISO-XFEM FOR TOPOLOGY OPTIMIZATION OF STRUCTURES UNDER MULTIPLE LOAD CASES AND ACCELERATION LOADING

Meisam Abdi, Ian A. Ashcroft, and Ricky D. Wildman

Faculty of Engineering, University of Nottingham, Nottingham, UK
{meisam.abdi,ian.ashcroft,ricky.wildman}@nottingham.ac.uk

Keywords: Topology Optimization, Acceleration Loading, Isoline, Isosurface, XFEM, Additive Manufacturing.

Abstract. *Iso-XFEM is an evolutionary based topology optimization method for generating high resolution topology optimised solutions using isolines/isosurfaces of a structural performance criterion and eXtended Finite Element Method (XFEM). The conventional approach for topology design of structures includes the use of a density based topology optimization method, such as SIMP, to find the optimal density distribution within the design domain, and thresholding the densities at an arbitrary value to obtain the design boundary. This will then require additional post-processing, such as smoothing and shape optimization, to generate a manufaturable design with smooth and clearly defined boundaries. In the proposed method, the use of XFEM with isoline/isosurface based boundary representation enables generation of topology solutions represented with smooth boundaries which can be straight away translated to a triangular surface file format, for instance STL format, in order to direct to manufacture. From analysis point of view, this allows the use of a coarser mesh during the optimization process to achieve a smooth solution, while this omits/significantly reduces the post processing required before manufacturing. The aim of this paper is to extend the Iso-XFEM method to include multiple loading cases and acceleration loading. Different 2D and 3D test cases are employed to show the effectiveness of the method.*

1 INTRODUCTION

Topology optimization aims to find the best possible layout for a structure within a specified design domain under a set of loads and boundary conditions. There has been a significant interest in topology optimization methods and applications over the last three decades starting from groundbreaking paper of Bendsøe and Kikuchi (1988) introducing homogenization method [1]. Other methods including Solid Isotropic Material with Penalization (SIMP) [2,3], Evolutionary Structural Optimization (ESO) [4,5], level set method [6,7] and genetic algorithms [8] were introduced after that. Although many of the proposed topology optimization algorithms have been developed and been applied on several problems, such as Michelle-type structures and cantilever beams with rectangular domains, there has been less attention on applying these algorithms on 3D real-life structures and real loading scenarios. In some cases the mathematical complexity or the size of the FE design domain doesn't allow the algorithm to be properly implemented. OptiStruct (Altair Engineering Inc.) is an example of software designed to enable the SIMP method of topology optimization to be applied to real components. Other software such as Nastran (MSC Software) and Abaqus FEA (Dassault Systèmes) also have option to apply similar density-based approaches to find the solution to topology optimization problems. Although the topology optimization modules of these software applications are being widely used for research and engineering purposes, a drawback of the density-based approaches (and many other element-based approaches) is that they cannot provide a clear and smooth representation of the design boundaries in converged topologies. This issue brings difficulties in interpreting the solutions, combining them with CAD and manufacturing the topologies. Therefore the solutions usually need post-processing, reanalysing and shape optimization before manufacturing.

Iso-XFEM was developed in a previous study to address the issues related to the boundary representation of the topology [9,10,11]. The idea was to use a simple evolutionary based optimization algorithm (similar to BESO) while improving the boundary representation by implementing isoline/isosurface approach during the optimization. An XFEM integration scheme was also used to increase the accuracy of FE solutions near the design boundary. The method was successfully applied to 2D and 3D structures with complex design domain [11], and the results showed a significant improvement in boundary representation and structural performance of the solutions over the conventional BESO. The aim of this paper is to extend the Iso-XFEM method into problems with multiple load cases and acceleration loading. In the next sections, an overview of the Iso-XFEM method followed by two examples of stiffness design using this method is presented. Then the extension of the method into optimization of structures under multiple load cases and acceleration loading with a few examples of each is presented.

2 AN OVERVIEW OF ISO-XFEM METHOD

The main three elements of the Iso-XFEM method include isoline/isosurface approach to represent the design boundary, XFEM to calculate the elemental sensitivities (a structural performance criterion) near the boundary, and an evolutionary based optimization algorithm. These three elements are explained in this section.

2.1 Isoline/isosurface approach

Isolines/isosurfaces are the lines/surfaces that represent the points of a constant value, named the isovalue, in a 2D/3D space. In structural optimization applications [9,11,12,13], the boundaries are defined by the intersection of the structural performance (SP) distribution with a minimum level of performance (MLP), which is typically increasing during the optimization

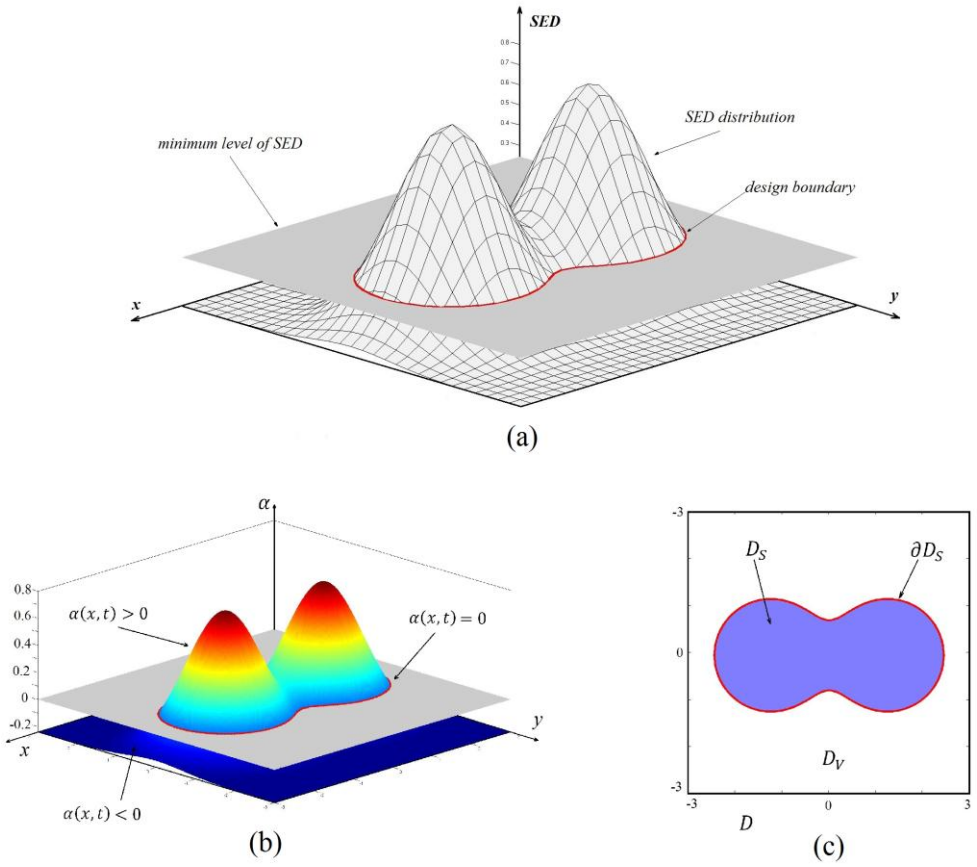
process. Figure 1(a) shows a 2D fixed grid design domain discretized with a 30x30 mesh, where the intersection of strain energy density (SED) distribution as a structural performance criterion with a minimum level of SED gives the design boundary. The relative performance, α, is defined as:

$$\alpha = SP - MLP \tag{1}$$

The design domain can be partitioned into void phase, boundary and solid phase, with respect to the values of relative performance:

$$\alpha(x): \begin{cases} > 0 & solid\ phase\ (D_S) \\ = 0 & boundary\ (\partial D_S) \\ < 0 & void\ phase\ (D_V) \end{cases} \tag{2}$$

Figures 1(b) & 1(c) show how the design space, D, from figure 1(a) is partitioned into D_S, ∂D_S and D_V using the relative performance function $\alpha(x)$, distributed over the design space.



Figur 1: (a) Boundary representation using isolines of a structural performance function (SED in here). The intersection of SP distribution with MLP defines the current state of the boundary. (b) Implicit representation of a 2D design space and the structure's geometry using relative structural performance. (c) Design space decomposed into solid region ($\alpha(x) > 0$), void region ($\alpha(x) < 0$) and boundary ($\alpha(x) = 0$).

2.2 XFEM

By implementing the above isoline/isosurface approach, the design boundary is superimposed on the fixed grid finite elements, making three groups of elements in the FE design space: solid elements, void elements, and boundary elements (the elements which lie on the boundary). The contribution of solid and void elements to the FE framework could simply be considered by assigning solid and void (very weak) material properties to those elements, respectively. In the case of boundary elements, in order to accurately represent the design boundary whilst avoiding expensive remeshing operations, an XFEM approach can be employed. XFEM approximation space for modeling holes and inclusions is given by [14]:

$$u(x) = \sum_i N_i(x)\, H(x) u_i \tag{3}$$

where $N_i(x)$ are the classical shape functions associated to the nodal degrees of freedom, u_i. The value of the Heaviside function $H(x)$ is equal to 1 for the nodes and regions in the solid part of the design and switches to 0 for nodes and regions in the void part of the design domain. This XFEM scheme was realized by dividing the solid domain of the boundary elements into sub-triangles (in 2D problems as shown in figure 2(a)) or sub-tetrahedra (in 3D problems as shown in figure 2(b)), and then performing numerical integration over solid triangles/tetrahedra using Gauss quadrature method.

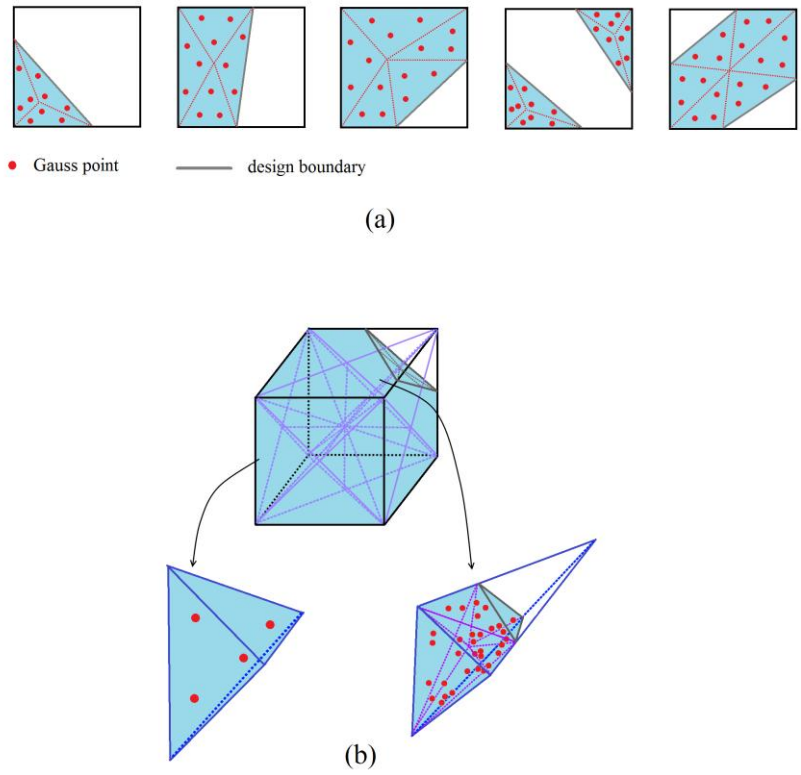


Figure 2: XFEM interation scheme. (a) solid domain of 2D boundary elments are devided into sub-triangles. (b) Solid domain of 3D boundary elements are devided into sub-tetrahedra.

2.3 Evolutionary based optimization method

The optimization algorithm used in the Iso-XFEM method is evolutionary based, i.e. based on the simple assumption that the optimized solution can be achieved by gradually removing the inefficient material from the design domain. However, unlike Evolutionary Structural Optimization (ESO) in which the material removal is carried out at an elemental level, in this approach the optimization operates at a global level of structural performance by the use of isoline/isosurface design approach. An appropriate performance criterion is used to characterize the efficiency of material usage in the design domain. Material is then removed from low relative performance regions ($x;\ \alpha(x)<0$) and redistributed to the high relative performance regions ($x;\ \alpha(x)>0$). The target volume of the design for the current iteration needs to be calculated before any region is added to or removed from the structure. The target volume of the design for the current iteration is given by

$$V_{it} = \max(V_{it-1}(1 - ER), V^c) \tag{4}$$

where ER is the volume evolution rate and V^c is the specified volume constraint. Once the target volume of the current iteration is found, the minimum level of performance which gives this volume needs to be identified. This could be achieved through an iterative process, for instance by defining upper and lower bands for MLP (which are equal to the maximum and minimum SP in first iteration, respectively), finding the volumes corresponding to the upper and lower bands, averaging and updating the upper and lower bands until the difference between the volumes corresponding to the upper and lower bands is smaller than a minimum value.

3 EXAMPLES OF STIFFNESS DESIGN USING ISO-XFEM

3.1 A comparison study

By employing Iso-XFEM method for topology optimization, we would expect to achieve solutions with smoother boundary than those obtained using conventional element based methods, for instance SIMP and BESO [9,11]. If a density based method, like SIMP, is employed for topology optimization, there is a need to threshold at an arbitrary density to achieve a discrete solution. However this could result in reducing the optimality of the solution. The aim of this test case was to investigate the optimality of Iso-XFEM solution for a benchmark problem and compare it with the solution obtained from a commercial software after thresholding densities.

To address this, Iso-XFEM was used to solve topology optimization problem for the cantilever structure shown in figure 3(a) using 40×20×2 hexahedral elements. Material properties were Young modulus $E = 1$, and Poisson's ratio $\nu = 0.3$. To minimize compliance, strain energy density was used as structural performance criterion. A volume evolution rate of $ER = 0.02$ was used. The solution converged after 40 evolutionary iterations (figure 3(b)). The same starting mesh was used to solve the cantilever problem using Optistruct, as shown in figure 4. Figure 4(a) is Optistruct solution (optimal density distribution) without implementing minimum member size constraint. It can be seen that the design boundary is not clearly defined as the solution is represented with intermediate relative densities. Figure 4(b) shows the densities thresholded in relative density of 0.5 which results to a solution with approximately the same volume as the prescribed volume constraint. However, it can be seen that thresholding the densities resulted in an unfeasible design as the structure has lost its members' connectivity. It is possible to preserve members' connectivity by thresholding at a lower density. However,

this would result in a solution with a higher volume than the volume constraint, which requires further optimization and post-processing.

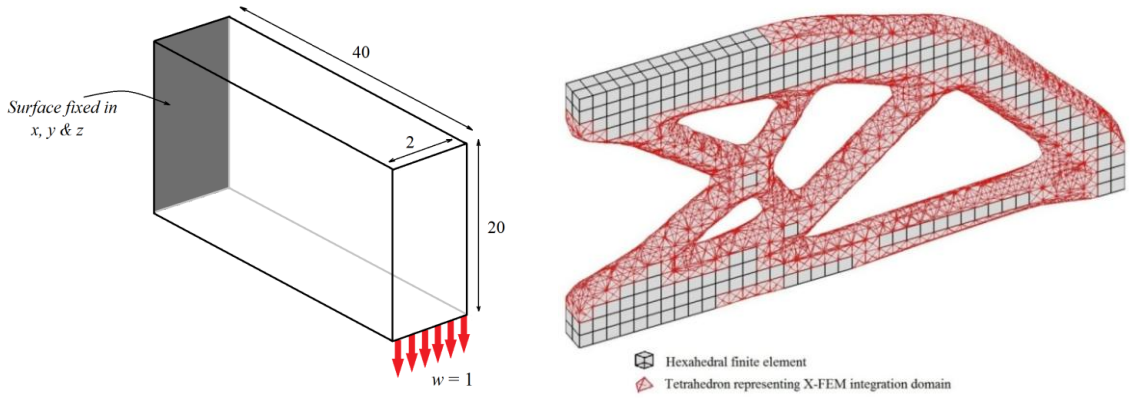


Figure 3: (a) Design domain and boundary conditions (b) Iso-XFEM solution.

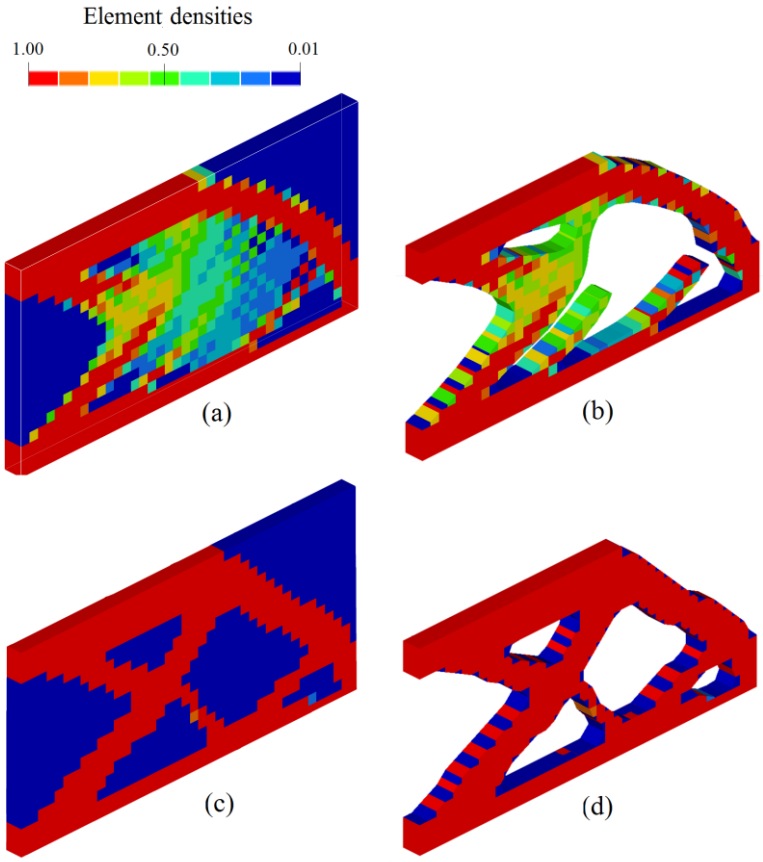


Figure 4: (a) Optistruct solution of cantilever structure without applying minimum member size constraint, and (b) densities of this solution thresholded at isovalue of 0.5. (c) Optistruct solution with minimum member size constraint, and (d) densities of this solution thresholded at isovalue of 0.5.

In order to have a near 0/1 solution and prevent the loss of members' connectivity after thresholding the densities, a minimum member size of 2×element size was applied, resulting in the solution shown in figure 4(c). It can be seen that this is a near 0/1 solution and the members' connectivity is well defined, however, with rough boundaries. Figure 4(d) shows

the solution obtained by thresholding the densities of the solution shown in figure 4(c). The densities were thresholded at isovalue of 0.5, resulting in the same volume as the prescribed volume constraint, thus, enabling the comparison of all solutions with the Iso-XFEM solution, shown in 3(b).

To quantify the optimality of the Optistruct and Iso-XFEM solutions, the strain energy of the solutions has been measured by performing FEA on the solutions, as presented in table 1. It can be seen that the design obtained by thresholding densities of optistruct solution has a higher value of objective than Optistruct solutions with/without minimum member size constraint. This is because the design obtained by thresholding the densities of a topology optimised solution, is not an optimised solution anymore, and will require additional shape optimization and further post-processing until it becomes an optimised manufacturable design. However, it can be seen that the strain energy of the Iso-XFEM solution is lower than the other three designs, indicating that the Iso-XFEM solution is a more optimal solution compared to the Optistruct solutions before/after thresholding the densities. The results of this experiment clearly show that although thresholding the densities can improve the boundary representation, it can reduce the optimality of the solutions.

Different solutions	SE (N.m)
Optistruct solution without minimum member size constraint (figure 4a)	7.34
Optistruct solution with minimum member size constraint (figure 4c)	7.17
Design obtained by thresholding densities of the Optistruct solution, with minimum member size constraint (figure 4d)	7.63
Iso-XFEM solution (figure 3b)	6.56

Table 1: Comparison of strain energies of Optistruct and Iso-XFEM solutions

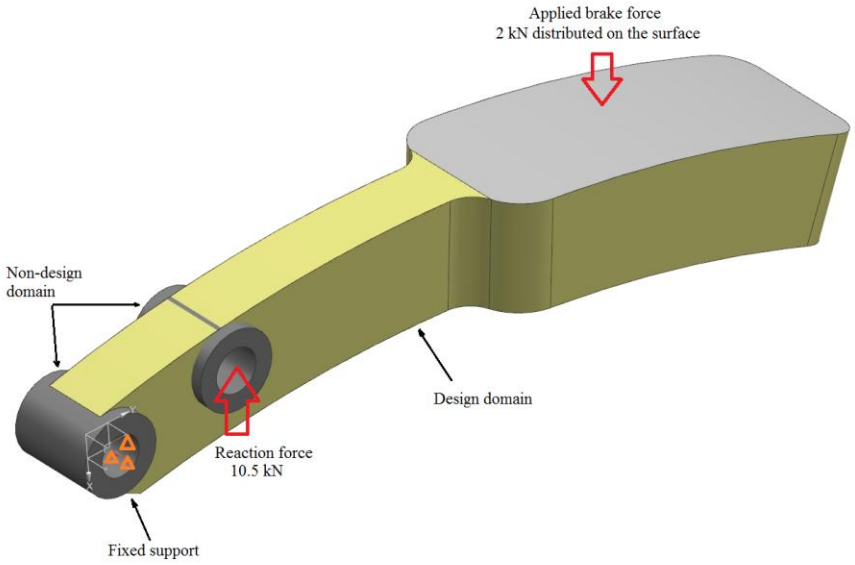
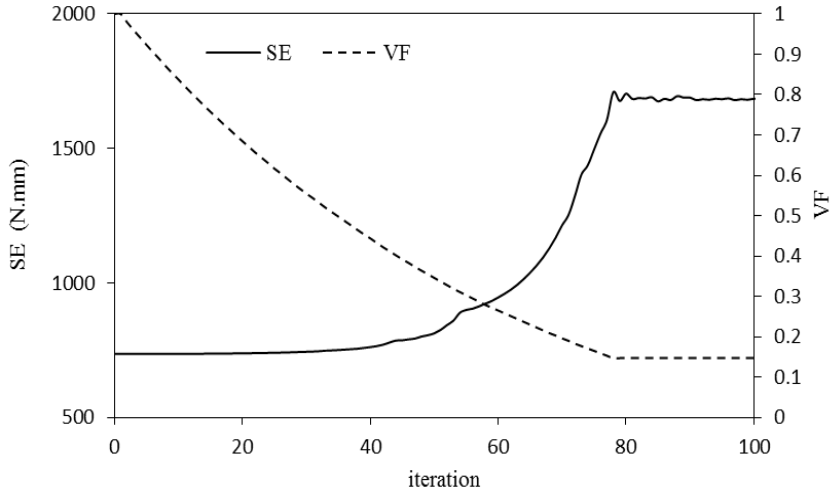


Figure 5: Design domain and boundary conditions of the brake pedal.

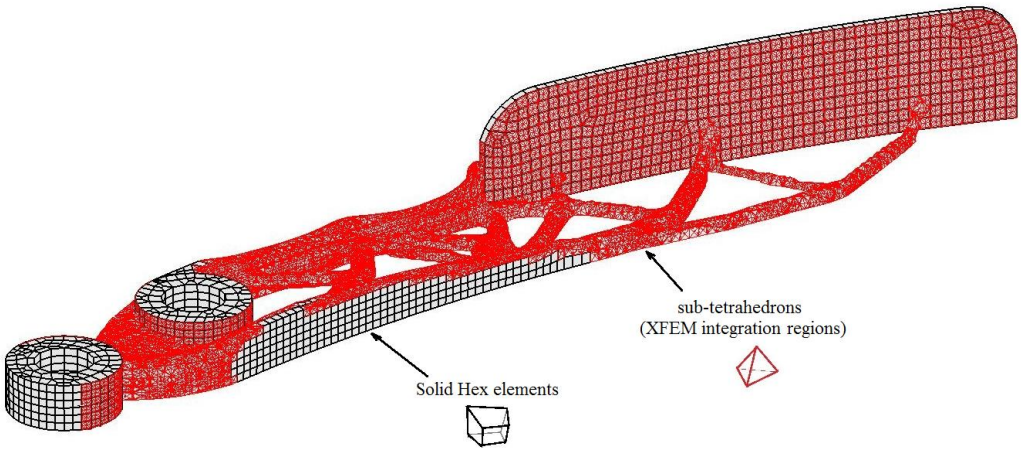
3.2 An industrial test case: a brake pedal

The method was applied to optimize a brake pedal with the loads and boundary conditions shown in figure 5. The objective was to minimize the compliance for the target volume of 15% of the initial design domain. The material used was Ti-6Al-4V. Due to the symmetry of the problem, only half of the structure was considered for the analysis using approximately

16000 hexahedral elements. A volume evolution rate of *ER = 0.02* was used. The solution converged after 90 evolutionary iterations as shown in figure 6(a). Figure 6(b) shows the Iso-XFEM solution for the brake pedal. It can be seen that despite using a coarse starting mesh, a fairly smooth solution has been achieved. Figure 7(a) shows the optimized design after inclusion of all non-design elements fallowed by a few iterations of smoothing. The part was built through Selective Laser Melting (SLM) process using Renishaw AM250 machine, as shown in figure 7(b).

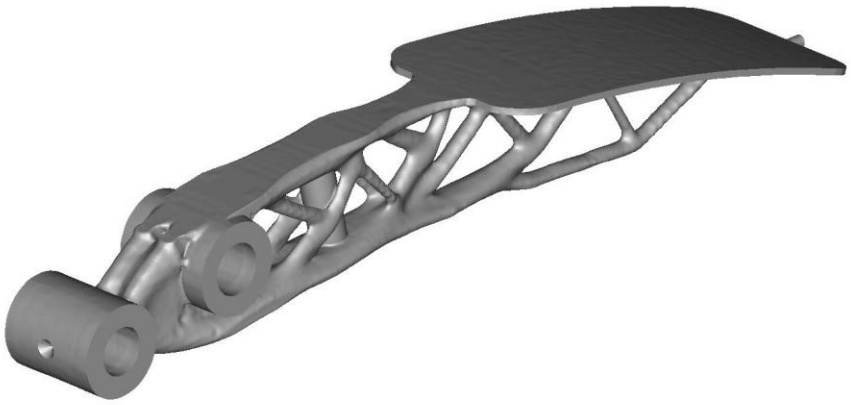


(a)

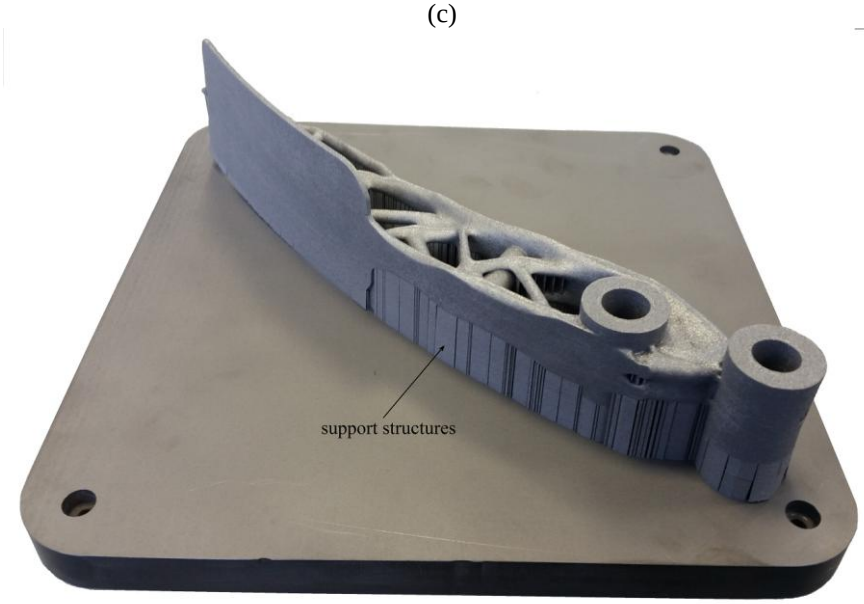


(b)

Figure 6: (a) Evolution histories of objective function (SE) and volume fraction (VF) of the brake pedal (b) Iso-XFEM solution (half of the pedal).



(c)



(d)

Figure 7: (a) Final design of the brake pedal after inclusion of all non-design regions (b) the part built using SLM process. Support structures will be removed after stress relieving process.

4 ISO-XFEM FOR MULTIPLE LOAD CASES

Most real-life structures are subjected to multiple load cases. Moreover, moving loads can be simplified to a finite number of load cases acting sequentially along the load path. In the case of stiffness optimization, in order to account for all load cases, the objective function can be defined as weighted average of mean compliance of all load cases [15]:

$$minimize: f(x) = \sum_{i=1}^{m} w_i C_i$$
$$subject\ to: \sum_{e=1}^{n} v_e^S = V^c \tag{5}$$

where m is the total number of load cases, n is the total number of elements, C is compliance, w_i is the weighting factor of ith load case, and v_e^S is the volume of the solid part of the element, and V^c is the design volume constraint. In this case, the structural performance associated to element e can be defined as

$$SP_e = \frac{1}{2}\sum_{i=1}^{m} w_i (u_e^T k_e u_e)_i / v_e \tag{6}$$

where u_e and k_e are the element's displacement vector and stiffness matrix, respectively, and v_e is the total volume of the element.

4.1 Example

The rectangular plate with the load and boundary conditions shown in figure 8(a) is considered. The objective was to find the stiffest design for the target volume of 30% of the initial design domain. Two unit loads were applied on the top edge of the rectangular plate. Material properties were Young modulus E= 1, and Poisson's ratio v = 0.3. Equal weighting factors were assumed for two load cases ($w_1 = w_2 = 0.5$). A mesh of 60×30 was used for the FE model of the plate. A volume evolution rate of ER = 0.02 was used for the optimization.

Figures 8(b) and 8(c) compare the topology obtained when the two loads are applied as a single load case (8b) with the one obtained from two different load cases (loads in different times), (8c). It can be seen that two different topologies have been obtained. However, the solution obtained for the structure under two load cases (figure 8(c)) can be more stable than the other solution, since it has converged to a triangulated frame structure which is more stable than a trapezoidal frame structure. Figure 9 shows the Evolution histories of objective function and volume fraction for the structure under two load cases. It can be seen that a stable material removal was carried out through the Iso-XFEM optimization process.

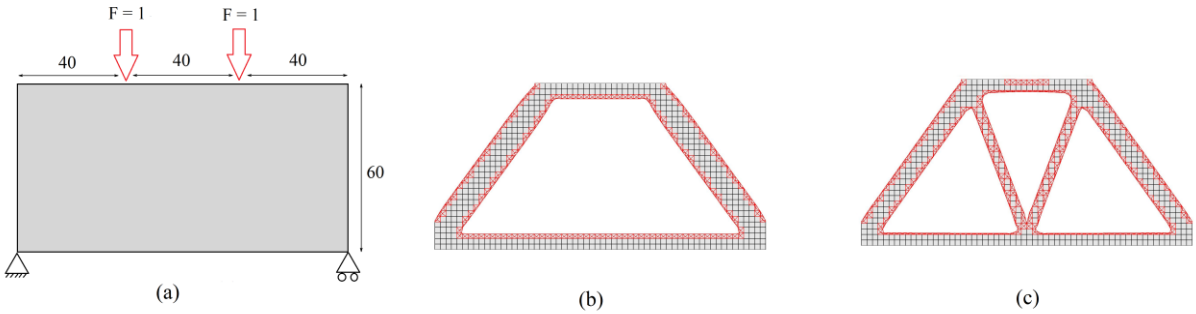


Figure 8: (a) Design domain and boundary conditions of the 2D structure (b) Iso-XFEM solution of the structure under a single load case of two unit force (c) Iso-XFEM solution of the structure under two load cases of a unit force each.

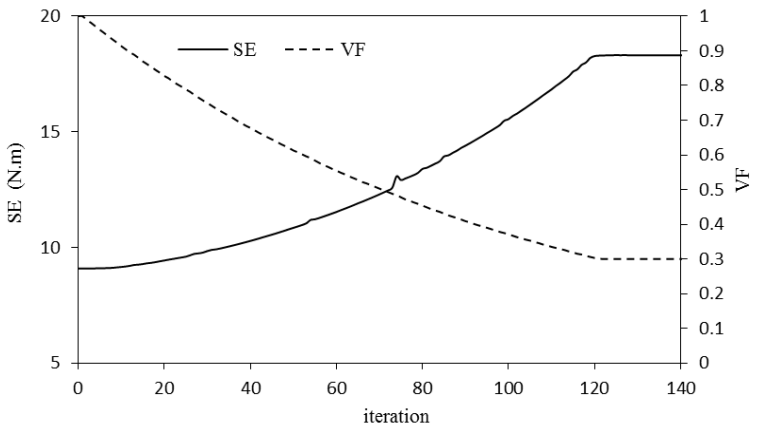


Figure 9: Evolution histories of objective function (SE) and volume fraction (VF) of the rectangular design domain subjected to two load cases.

5 ISO-XFEM FOR GRAVITY AND ACCELERATION LOADING

Gravity load is an important consideration in many engineering design applications such as deigning civil structures. Other forms of acceleration loading may also exist in a mechanical or an aerospace component. Acceleration loading is different from fixed external loading in that the applied force includes design dependent acceleration load, requiring the force vector F in the FE system to be updated in each optimization iteration. Therefore, we use the same objective function as compliance (total strain energy) for stiffness design, and update the nodal forces by

$$F_g^j = \sum_{e=1}^{k} v_e^s \rho g / k \tag{7}$$

where F_g^j is the acceleration force of node j in direction of acceleration g, ρ is the density, and k is the number of elements connected to node j. Note that for simplicity, we assumed that number of elements connected to a node is equal to the number of nodes of an element (for instance for 4-node quadrilateral elements k = 4, and for 8-node hexahedral elements, k = 8).

5.1 A 2D plate under gravity load

A simply supported 1 m × 0.5 m plate shown in figure 10(a) is considered. The objective was to find minimize the compliance of the plate under gravity load for the target volume fraction of 15% of the initial design domain. Material properties were Young modulus of 200 GPa, Poisson's ratio of 0.3, and density of 78 kg/m^3. A mesh of 60×30 was used for the FE model of the plate. A volume evolution rate of ER = 0.02 was used for the optimization. Figure 10(b) shows the converged solution. The results are in agreement with BESO solution in [15].

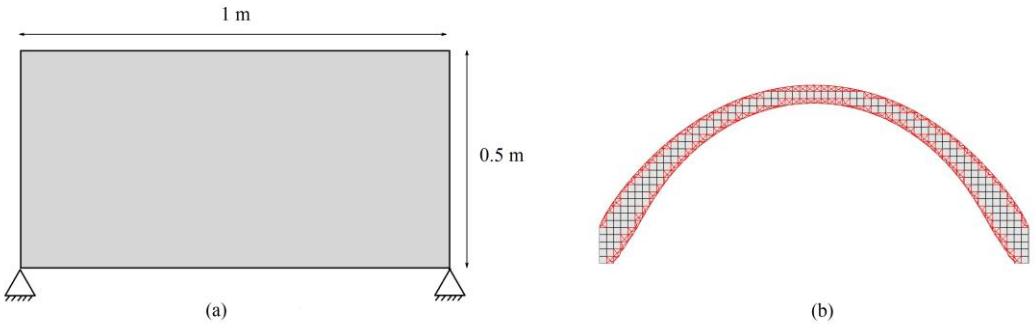


Figure 10: (a) Design domain and boundary conditions of the 2D structure under gravity load (b) Iso-XFEM solution.

5.2 3D structure under gravity load

The cuboid design domain shown in figure 11(a) was used as the 3D test case under gravity load. The structure was simply supported at all four bottom corners. The objective was to minimize the compliance of the structure under self-weight for the target volume fraction of 5% of the initial design domain. Material properties were Young modulus of 200 GPa, Poisson's ratio of 0.3, and density of 78 kg/m^3. A volume evolution rate of ER = 0.04 was used for the optimization. The structure is symmetric in two directions. Therefore only a quarter of the design space was considered for FEA using 32×32×40 elements. The final Iso-XFEM solution is shown in figure 11(b). Figure 11(c) shows the design obtained after mirroring the Iso-

XFEM solution and converting the solution to STL file format. No further post-processing/smoothing was performed to generate the design shown in figure 11(c).

Figure 12 shows the evolution histories of objective function and volume fraction of the 3D structure under self-weight. It can be seen that the Iso-XFEM material removal process for this structure is highly stable. Also, unlike compliance minimization problems for the structures under external loads where by removing material, the compliance increases, here, the material removal resulted in decreasing the compliance. Evidently, this is because the gravity load is a design dependent load and by removing material less load was applied to the structure.

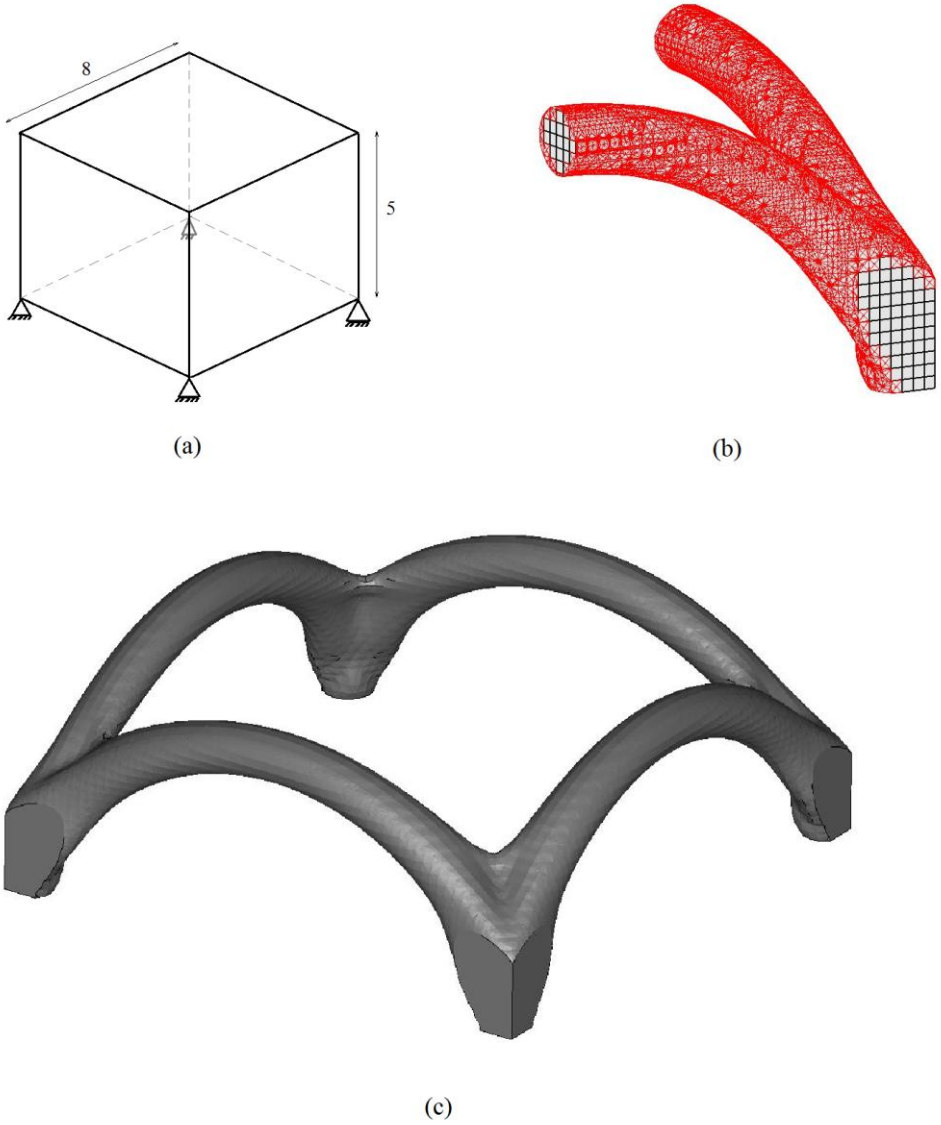


Figure 11: (a) Design domain and boundary conditions of the 3D structure under gravity load (b) Iso-XFEM solution considering only a quarter of the structure (c) Final design.

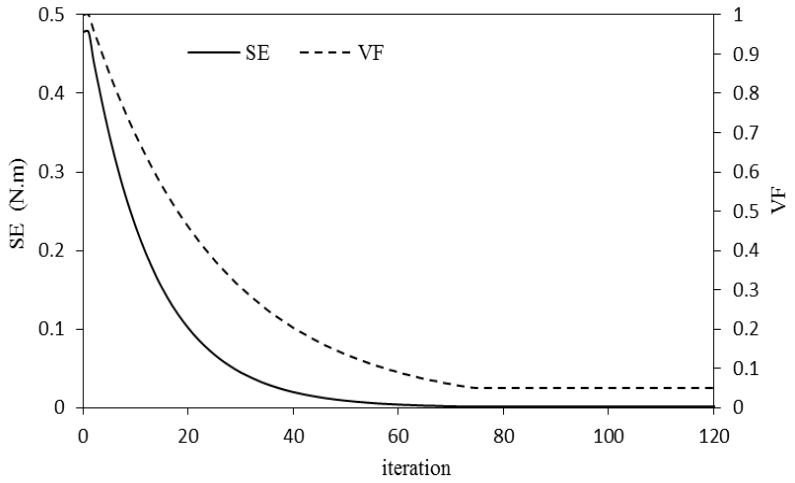


Figure 12: Evolution histories of objective function (SE) and volume fraction (VF) of the cuboid design domain under gravity load.

6 SUMMARY AND CONCLUSION

Iso-XFEM is an evolutionary structural optimization method which can be used for shape and topology optimization of continuum structures. The use of XFEM with isosurface approach has enabled achieving high resolution topology optimized solutions which require no more or only a little post processing before manufacturing. It was shown that the method is capable of optimizing geometrically complex structures, for instance the brake pedal studied in this paper.

In this study, the method was further extended to include topology optimization of structures subjected to multiple load cases and acceleration loading. It was shown that design for multiple load cases can be different from the one for a single load case. The method was also found to be stable for optimization of structures under acceleration loading.

The Iso-XFEM method was found to be a promising method for topology optimization of real-life structures with real loading scenarios. The boundary of the solutions is represented with triangles (rather than element densities in density based topology optimization methods). This makes it straightforward to convert the solutions into STL format for additive manufacturing. The amount of post-processing before manufacturing is highly depended to the mesh size used in the analysis. Considering the fact that even using a coarse mesh, relatively high resolution solutions can be achieved using this method, by taking account the resolution of the 3D printer before creating FE mesh, the post-processing after topology optimization can be avoided and the solutions can be sent directly to additive manufacture.

REFERENCES

[1] M.P. Bendsøe, K. Noboru, Generating optimal topologies in structural design using a homogenization method. *Computer methods in applied mechanics and engineering*, 71, 197-224, 1988.

[2] M.P. Bendsøe, Optimal shape design as a material distribution problem. *Struct. Optim.* 1, 193-202, 1989.

[3] M. Zhou, G.I.N. Rozvany, The COG algorithm, Part II: Topological, geometrical and general shape optimisation. *Comp. Meth. Appl. Mech. Eng*., 89, 309-336, 1991.

[4] Y.M. Xie, G.P. Steven, A simple evolutionary procedure for structural optimization. *Computers & Structures*, 49, 885-896, 1993.

[5] Y.M. Xie, G.P. Steven, *Evolutionary structural optimization*, London, 1997.

[6] M.Y. Wang, X. Wang, D. Guo, A level set method for structural topology optimisation. *Comput. Meth .Appl. Eng.*, 192, 227-46, 2003.

[7] G. Allaire, F. Jouve and A.M. Toader, Structural optimisation using sensitivity analysis and a level set method. *J. Comp. Phys.*, 194, 363-393, 2004.

[8] M. Jakiela, C. Chapman, J. Duda, A. Adewuya, K. Saitou, Continuum structural topology design with genetic algorithms. *Comput Methods Appl Mech Eng*, 186(2–4), 339-356, 2000.

[9] M. Abdi, R.D. Wildman, I.A. Ashcroft, Evolutionary topology optimization using extended finite element method and isolines. *Engineering Optimization*, 46(5), 628-647, 2014.

[10] M. Abdi, I.A. Ashcroft, R.D. Wildman, An X-FEM based approach for topology optimization of continuum structures. *Advances in Intelligent Systems and Computing*, 256, 277-289, 2014.

[11] M. Abdi, I.A. Ashcroft, R.D. Wildman, High Resolution Topology Design with Iso-XFEM, *Proceedings of the Solid Freeform Fabrication Symposium*, Austin, TX, 1288-1303, 2015.

[12] M. Victoria, P. Martı´, O.M. Querin, Topology design of two-dimensional continuum structures using isolines. *Computer and Structures*, 87, 101-109, 2009.

[13] M. Victoria, O.M. Querin, P. Martı´, Topology design for multiple loading conditions of continuum structures using isolines and isosurfaces. *Finite Elements in Analysis and Design*, 46, 229-237, 2010.

[14] N. Sukumar, D.L. Chopp, N. Moës, T. Belytschko, Modeling Holes and Inclusions by Level Sets in the Extended Finite Element Method. *Computer Methods in Applied Mechanics and Engineering*, 190, 6183-6200, 2001.

[15] X. Huang, M. Xie, Evolutionary topology optimization of continuum structures: methods and applications. John Wiley & Sons, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

LAMINATE ELEMENT METHOD AND ITS APPLICATION TO THE STUDY OF GUIDED WAVE RESONANCE PHENOMENA IN LAYERED ELASTIC STRUCTURES WITH DEFECTS

Evgeny Glushkov[1], **Natalia Glushkova**[1], **Artem Eremin**[1], **and Rolf Lammering**[2]

[1]Institute for Mathematics, Mechanics and Informatics, Kuban State University
Krasnodar 350040, Russia
e-mail: evg@math.kubsu.ru

[2] Institute of Mechanics, Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg
D-22043 Hamburg, Germany
e-mail: rolf.lammering@hsu-hh.de

Keywords: elastic guided waves; semi-analytical integral approach; laminate element method; resonance diffraction

Abstract. *A semi-analytical laminate element technique developed for the investigation of guided wave propagation and diffraction in layered structures with local inhomogeneities is presented. The approach is based on the use of fundamental solutions for the pristine layered structure as a basis functions for the scattered field approximation. As an example, its application to the investigation of resonance guided wave interaction with deep surface notches in elastic layer, which gradually change shape from rectangular to elliptical, has been considered to estimate the effect of notch geometry on resonance frequencies. The obtained numerical results are validated by the finite element simulation and experimentally confirmed on the basis of laser Doppler vibrometry.*

1 INTRODUCTION

Active structural health monitoring of plate-like units is commonly based on the propagation of elastic guided waves (GW) for long distances and their interaction with local inhomogeneities (defects) of various types [1]. The reflection, refraction and mode conversion phenomena serve as distinct indications of faults or defects. Computer simulation aims to clarify such a complex ultrasonic wave motion. General-purpose and specific finite-element computational codes have proven their efficiency for calculating wave propagation in elastic structures with complex geometry [2]. At the same time, with lengthy laminate waveguides, analytically based methods, such as boundary integral equation (BIE) technique and its derivations, e.g., boundary element method (BEM) or method of fundamental solutions (MFS), may serve as an efficient alternative [3]. They allow one to reduce the problem's dimension and obtain the results in a physically clear form of GW asymptotic expressions.

The laminate element method (LEM) [4] is a kind of boundary element technique specifically adjusted to simulate the elastodynamic behavior of lengthy layered structures. It is based on the boundary integral representations of elastic wave fields with the kernels in the form of fundamental solutions for the intact structure as a whole. Such basis functions called laminate elements (LEs) satisfy identically the governing equations in the sub-layers and interface boundary conditions among them as well as homogeneous conditions on the exterior plane-parallel surfaces. Therefore, only the integration over the obstacle's surface is necessary for the LE approximation of scattered wave fields.

In the current contribution, we present and discuss the LEM application to the investigation and parametric analysis of GW resonance phenomena in layered structures with obstacles. The resonance trapping mode effect [5] is featured by the capturing of incident wave energy and its prolonged localization in the defect's vicinity in the form of weakly decaying oscillations at the resonance frequencies [6]. The latter coincide with the spectral points of the corresponding boundary value problem (BVP) in the complex frequency plane. These frequencies strongly depend on the defect's size, shape and depth, making resonance effects a potentially useful tool for the enhancement and automation of the monitoring processes. In the calculations, they are approximated by the roots of the determinant of the linear algebraic system, to which the original BVP is reduced within the LEM approach.

A LEM based analysis of resonance wave phenomena is demonstrated on the examples of GW propagation and diffraction in aluminium specimens with surface notches of varying shape. The evaluated wave patterns and eigenfrequencies are compared with the results of FEM simulation and experimental data acquired using a contactless laser Doppler vibrometry allowing refined wave propagation sensing and visualization. The experimental and FEM-based results confirm the predicted values of resonance frequencies, revealing the trapping mode effect with a long standing-wave-type oscillation near the obstacles.

2 MATHEMATICAL MODEL

Let us consider a plane-strain time-harmonic oscillation $\mathbf{u}e^{-i\omega t},\ \mathbf{u} = \{u_x,\ u_z\}$, of a layered isotropic linear-elastic waveguide of thickness H governed by the Lamé equations. In the Cartesian coordinate system $\mathbf{x} = (x, z)$, the structure occupies the domain $D = \{|x| < \infty,\ -H < z < 0\}$ with surface and/or internal inhomogeneities (Fig. 1). The exterior plane-parallel surfaces $z = 0$ and $z = -H$ are stress-free except, possibly, at a local source zone Ω to which a load $\mathbf{q}_0$ is applied. Surface and internal defects are located to the right from this area, and their boundaries denoted by S are also traction-free.

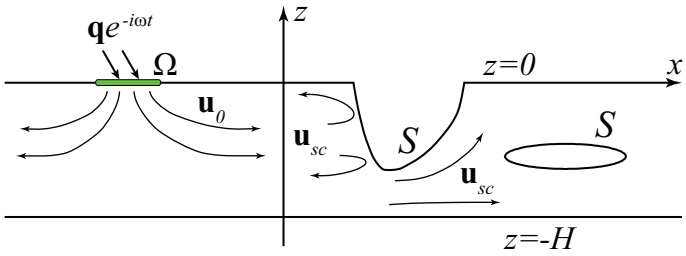


Figure 1: Schematic geometry of the problem.

The interaction of load-induced GWs $\mathbf{u}_0$ with obstacles results in the scattered field $\mathbf{u}_{sc}$. The total wave field $\mathbf{u}$ is a sum of the known, i.e., analytically expressed field $\mathbf{u}_0$ generated by the source in the pristine layer, and unknown scattered (reflected and transmitted) field $\mathbf{u}_{sc}$: $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}_{sc}$.

Within the geometry considered, the incident field $\mathbf{u}_0(\mathbf{x})$ can be represented as a path integral of the inverse Fourier transform $\mathcal{F}_x^{-1}$ with respect to the horizontal coordinate x, which is further reduced to the far-field GW asymptotics using the residue technique:

$$\mathbf{u}_0(\mathbf{x}) = \tfrac{1}{2\pi}\int\limits_{\Gamma} K^+(\alpha,z)\mathbf{Q}_0(\alpha)e^{-i\alpha x}d\alpha \approx \sum_{m=1}^{M}\mathbf{a}_m(z)e^{i\zeta_m x}, \qquad (1)$$
$$\mathbf{a}_m(z) = -i\mathrm{res}K^+(\alpha,z)\mathbf{Q}_0(\alpha)|_{\alpha=-\zeta_m}$$

Here $K^+(\alpha,z) = \mathcal{F}_x[k^+(\mathbf{x})]$ and $\mathbf{Q}_0(\alpha) = \mathcal{F}_x[\mathbf{q}_0(\mathbf{x})]$ are Fourier symbols of the waveguide Green's matrix k^+ and load vector-function $\mathbf{q}_0$; the sign "+" indicates that the traction is applied to the upper surface $z = 0$; M is the number of real poles ζ_m of the Green's matrix $K^+(\alpha,z)$, the residues from which yield travelling waves. The sum in Eq. 1 is valid for the travelling GWs propagating to the right from the loading area Ω; the integration path Γ bypasses real poles ζ_m according to the principle of limiting absorption.

Within the indirect BIE formalism, the unknown field $\mathbf{u}_{sc}$ is sought for in terms of 2×2 LE matrix $l(\mathbf{x},\boldsymbol{\xi})$ integrated over the defect's boundary S together with an unknown source density $\mathbf{c}(\mathbf{x})$:

$$\mathbf{u}_{sc}(\mathbf{x}) = \int\limits_S l(\mathbf{x},\boldsymbol{\xi})\mathbf{c}(\boldsymbol{\xi})d\boldsymbol{\xi} \qquad (2)$$

In contrast to the classical fundamental-matrix solution $g(\mathbf{x}-\boldsymbol{\xi})$ for an infinite homogeneous elastic space [3], the matrix $l(\mathbf{x},\boldsymbol{\xi})$ is derived for the pristine infinite layered structure as a whole. Their columns $\mathbf{l}_j$ are displacement vectors $\mathbf{u}_j(\mathbf{x},\boldsymbol{\xi})$ associated with the point sources $\delta(\mathbf{x}-\boldsymbol{\xi})\mathbf{i}_j$ directed along the coordinate unit vectors $\mathbf{i}_1$ and $\mathbf{i}_3$ taken to be parallel to the axes x and z, respectively; $\boldsymbol{\xi} = (\xi_1,\xi_3)$ is the point of LE source location.

Being straightforward and computationally efficient, representation (1) inspires to derive the scattered field (2) in an analogous way. For this purpose the matrix $l(\mathbf{x},\boldsymbol{\xi})$ is constructed as a composition of the conventional fundamental solution matrix $g(\mathbf{x}-\boldsymbol{\xi})$ and non-singular matrix $v(\mathbf{x})$ that accounts for the fields reflected from the external sides of the structure and provides the required homogeneous boundary conditions at these plane-parallel boundaries: $l = g + v$. The matrix v, in its turn, is decomposed into two terms; each of them, being derived via the Green's matrices k^+ and k^- for the considered layer with non-zero stresses q^+ and q^- at the

upper and lower sufraces $z = 0$ and $z = -H$, respectively. With the tractions $q^{\pm}$ caused by the direct source field $g(\mathbf{x} - \boldsymbol{\xi})$, it takes the form::

$$v(\mathbf{x}) = \int_{\Gamma} [K^{+}(\alpha, z)Q^{+}(\alpha) + K^{-}(\alpha, z)Q^{-}(\alpha)]e^{-i\alpha x} d\alpha \tag{3}$$

where $Q^{\pm}(\alpha) = \mathcal{F}_x[-Tg(\mathbf{x} - \boldsymbol{\xi})]$ and T is the matrix of the stress operator for the external plane-parallel boundaries.

At the next step, the vector factor $\mathbf{c}(\boldsymbol{\xi})$ is discretized the same way as in the BEM, e.g., assuming a piecewise constant approximation over S. It gives rise to the following approximate representation for the scattered field:

$$\mathbf{u}_{sc} \approx \sum_{j=1}^{N} \mathbf{u}_{sc,j},\ \mathbf{u}_{sc,j}(\mathbf{x}) = \int_{S_j} l(\mathbf{x}, \boldsymbol{\xi})\mathbf{c}_j d\boldsymbol{\xi} = \mathbf{u}^{inf}_{sc,j}(\mathbf{x}) + \mathbf{u}^{c}_{sc,j}(\mathbf{x}) \tag{4}$$

$$\mathbf{u}^{inf}_{sc,j}(\mathbf{x}) = \int_{S_j} g(\mathbf{x} - \boldsymbol{\xi})\mathbf{c}_j d\boldsymbol{\xi}$$

Here S_j are segments of a polygon approximation of the smooth boundary S; $\mathbf{c}_j$ are 2×1 vectors of unknown constants. The non-singular component of the scattered field $\mathbf{u}^{c}_{sc,j}(\mathbf{x})$ can be represented in the way similar to expression (3):

$$\mathbf{u}^{c}_{sc,j}(\mathbf{x}) = \int_{\Gamma} [K^{+}(\alpha, z)\mathbf{Q}^{+}_{j}(\alpha) + K^{-}(\alpha, z)\mathbf{Q}^{-}_{j}(\alpha)]e^{-i\alpha x} d\alpha \tag{5}$$

$$\mathbf{Q}^{+}_{j} = \mathcal{F}_x[-T\mathbf{u}^{inf}_{sc,j}(x, 0)],\ \ \mathbf{Q}^{-}_{j} = \mathcal{F}_x[-T\mathbf{u}^{inf}_{sc,j}(x, -H)]$$

The derivation of scattered travelling waves is essentially based on the analytical evaluation of the vector-functions $\mathbf{Q}^{\pm}_{j}(\alpha)$. It is achieved using the trick, which is based on the specific rotation in the Fourier parameter space [7].

The unknown constants $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N)$ are obtained from the system of linear algebraic equations $A\mathbf{c} = \mathbf{f}$, which arises from the substitution of relation (3) into the boundary conditions on S and further implementation of the Galerkin projection scheme. Resonance frequencies $\hat{\omega}_n$ of the considered BVP are approximated by the roots ω_n of the characteristic equation

$$\Delta(\omega) = \det A(\omega) = 0 \tag{6}$$

This equation is numerically solved using either Müller method (if a good initial guess of ω_n is known in advance) or with the argument principle approach.

3 RESULTS AND DISCUSSION

Recent theoretical and experimental investigations of the GW diffraction by deep rectangular notches in an elastic waveguide have revealed the trapping mode effect [8, 9]. Mathematically it is associated with the spectral points $\hat{\omega}_n = 2\pi\hat{f}_n$ of the BVP considered that are located in the complex frequency plane close to the real axis. While in a transient field this effect is featured by the wave energy localization near the obstacle, in the time-harmonic field it manifests itself in a strong resonance transmission of the incident wave at the frequencies $f_n \approx \mathrm{Re}\,\hat{f}_n$. Considering the pitting corrosion damage detection as a goal, it is of particular practical interest to further investigate such resonance phenomena for various notch shapes.

As an example, Fig. 2 demonstrates the A_0 mode transmission coefficient $\kappa^{+}(f)$ for various surface notches. Their shapes vary from the rectangle one (dashed curve No. 1, same as in Ref.

[8]) to the elliptical form (solid magenta curve No. 6). The coefficient κ^+ is introduced as the ratio of the time-averaged wave energy E_A^+ carried by the A_0 mode behind the obstacle to the incident wave energy E_0: $\kappa^+ = E_A^+/E_0$.

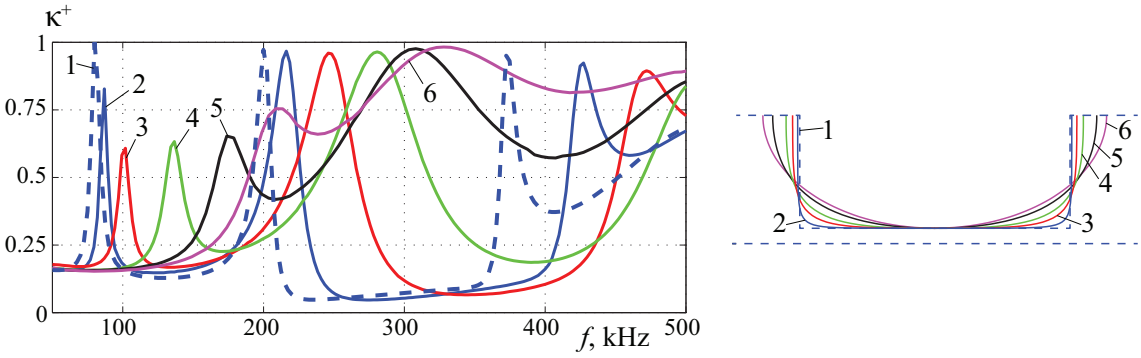


Figure 2: A_0 mode transmission coefficient κ^+ vs frequency f; curve 1: rectangular notch, 2: $p = 8$ in Eq. 7, 3: $p = 6$, 4: $p = 5$, 5: $p = 4$, 6: $p = 1$.

The input parameters for the calculations correspond to the aluminium plate samples used in the experiments: $Y = 70$ GPa, $\nu = 0.34$, $\rho = 2700$ kg/m^3; $H = 2$ mm; the initial obstacle is the surface sawed rectangular notch (width $a = 2$ mm, depth $d = 1.76$ mm), while the boundaries of other defects are constructed using the parametric equations

$$x = a_1 \cos^p(t), \quad y = b_1 \sin^p(t), \tag{7}$$

where $b_1 \equiv d$ and the parameter a_1 varies to provide the defect's square equal to $a \cdot d$; an elliptic surface occurs with $p = 1$.

It is clear from Fig. 2 that though the damage severity (its square and depth) remains the same, the defect's shape strongly influences on the peak transmission frequencies f_n. These frequencies f_n were taken as initial guesses for the $\hat{\omega}_n$ evaluation from Eq. 6. The first two complex eigenvalues of the considered BVP are summarized in Table 1 (in [kHz]).

Obstacle type	$\hat{f}_1$	$\hat{f}_2$	f_1^{FEM}	f_2^{FEM}
rectangular notch	81(1-i0.05)	200(1-i0.03)	81	216
$p = 8$	82(1-i0.04)	217(1-i0.05)	89	216
$p = 6$	97(1-i0.05)	246(1-i0.09)	115	255
$p = 5$	130(1-i0.06)	274(1-i0.14)	140	299
$p = 4$	182(1-i0.08)	306(1-i0.21)	181	316
elliptical notch	203(1-i0.11)	313(1-i0.27)	199	315

Table 1: The first two complex eigenvalues of the notched plates.

The validity of the obtained natural frequencies for a rectangular notch has been experimentally confirmed on the basis of laser Doppler vibrometry measurements with $f_1^{exp} = 85$ kHz and $f_2^{exp} = 215$ kHz [8]. However, the manufacturing of other notch types is not so straightforward and, therefore, FEM simulation with COMSOL Femlab 5.0 has been also used for the verification. For this purpose, following the strategy proposed in Ref. [10], eigenvalues of a finite

specimen, which is bounded by the defect's surface S from above and by the plane $z = -H$ from below with fixed side edges, are evaluated and summarized in the last two columns of Table 1, being in a considerable coincidence with the LEM-based results. At the same time, the FEM simulation does not account for the wave energy outflow to infinity. That is why it yields only real spectral points (resonance frequencies f_n^{FEM}) without imaginary parts that specify the damping rate of resonance oscillations.

Resonance GW interaction with defects is typically followed by a strong localization of oscillations in the vicinity of the obstacle. Examples of such localization are shown in Fig. 3, depicting spatial distributions of the amplitude $|u_z(x,z)|$ of the vertical displacement component at the frequencies $f_n \approx \mathrm{Re}\,\hat{f}_n$ for the rectangular and elliptical ($p = 1$) notches. The lower images are for the validating FEM results obtained for the aforementioned limited cuts from the specimens at the corresponding eigenfrequencies f_1^{FEM} and f_2^{FEM}. These eigenforms $u_z^{FEM}(x,z)$ are very close to the LEM-based obtained ones that are shown in the upper images. One can notice that in the case of elliptical notch the difference between the oscillation amplitudes inside and outside the defected region is not so contrast as for the rectangular obstacle due to the sufficiently bigger imaginary part in $\hat{f}_2$.

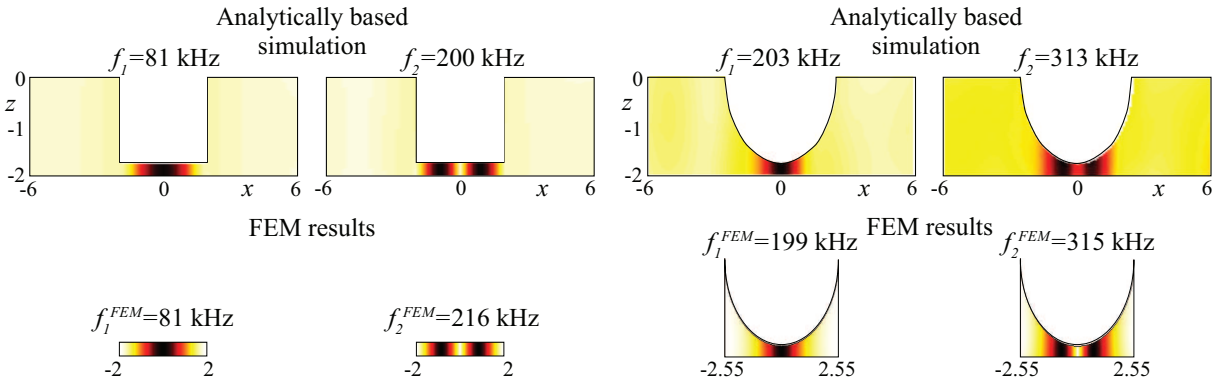


Figure 3: The patterns of resonance wave localization in the infinite waveguides (top plots) and corresponding eigenforms of the finite specimens (bottom plots) (dark regions correspond to higher amplitudes).

In the transient GW diffraction, natural frequencies $\hat{f}_n$ pronounce themselves with prolonged oscillations being localized at the obstacle region (trapping modes) [6]. The duration of such motion depends on the imaginary part of the corresponding eigenfrequency $\delta_n = 2\pi \mathrm{Im}\hat{f}_n$ through the factor $\exp(-\delta_n t)$. Though the real part of $\hat{f}_n$ is well approximated by FEM modelling, a LEM-based simulation for the infinite waveguide allows predicting both the resonance frequency and "quality" through the complete evaluation of complex $\hat{f}_n$.

In order to illustrate the influence of $\mathrm{Im}\hat{f}_n$ on the transient GW interaction with notches, the B-scans, which are out-of-plane surface velocity $v_z(x,t)$ along the x-axis depicted as functions of x and t, are shown in Fig. 4. The incident wave packets are generated by a remote surface-bonded piezoactuator excited with a Hann-modulated five-cycle bursts with central frequencies f_c close to the ones in the first and last rows of Table 1. As expected, in the case of rectangular defect, an intense and prolonged localization of wave energy at the notch is observed both theoretically and experimentally [8, 9]. At the same time, the growth of $\mathrm{Im}\hat{f}_n$ results in fast attenuation of oscillations in the defected area (two right images in Fig. 4). It is especially pronounced for the second eigenfrequency $\hat{f}_2$ of the elliptical notch. The motion at the obstacle vicinity stops just after the incident wavefield leaves this area.

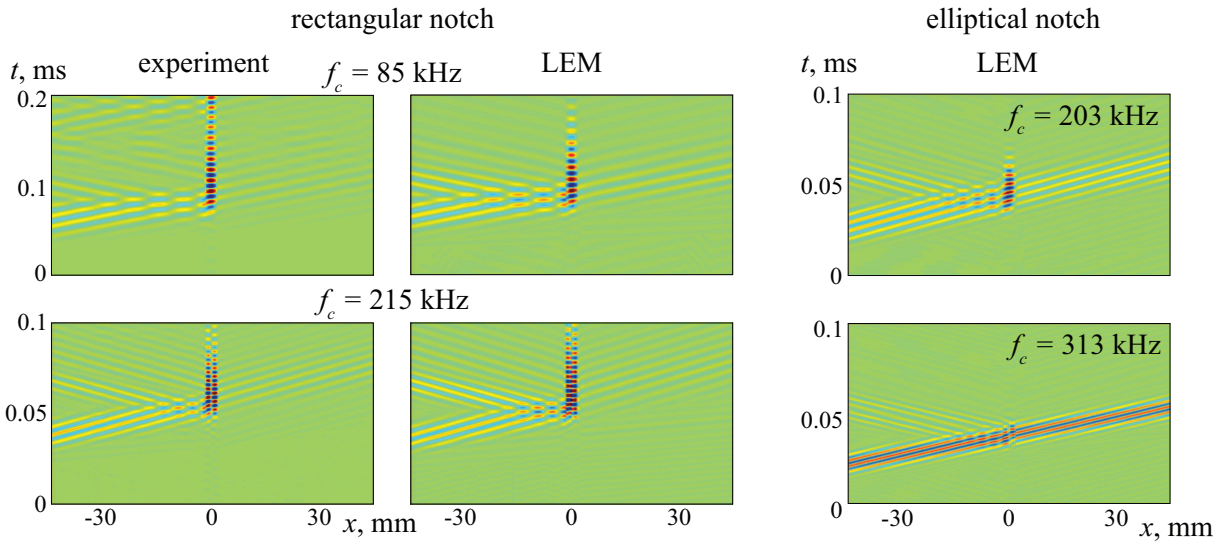


Figure 4: Localization of transient oscillations in the obstacle vicinity at the resonance frequency excitation that is ceased in the last case due to a high damping rate induced by comparatively large $\mathrm{Im}\hat{f}_n$.

4 CONCLUSIONS

A semi-analytical LEM-based approach is developed for the investigation of resonance diffraction of time-harmonic and transient GWs by a deep notch in a metallic plate. It allows revealing a strong dependence of the specimen's complex eigenfrequencies on the defect's shape and to study the influence of the eigenfrequency imaginary part on the intensity of the trapping mode effect.

The work is supported by the Russian Foundation for Basic Research (project No. 14-08-00370) and by the Russian Ministry of Science and Education (project No. 1.189.2014K).

REFERENCES

[1] T. Wandowski, P. Malinowski, W. Ostachowicz, M. Rawski, P. Tomaszewicz, T. Luba, G. Borowik, Embedded damage localization subsystem based on elastic wave propagation. *Computer-Aided Civil and Infrastructure Engineering*, **30**, 654–665, 2015.

[2] C. Willberg, S. Duczek, J.M. Vivar-Perez, Z.A.B. Ahmad, Simulation methods for guided wave-based structural health monitoring: a review. *Applied Mechanics Reviews*, **67**, 010803, 2015.

[3] M.H. Aliabadi, *The boundary element method. Volume 2: Applications in solids and structures*, John Willy & Sons LTD, 2002.

[4] Ye.V. Glushkov, N.V. Glushkova, A.A. Yeremin, V.V. Mikhas'kiv, The layered element method in the dynamic theory of elasticity. *Journal of Applied Mathematics and Mechanics*, **73**, 449-456, 2009.

[5] F. Ursell, Trapping modes in the theory of surface waves. *Mathematical Proceedings of the Cambridge Philosophical Society*, **47**, 347–358, 1951.

[6] E. Glushkov, N. Glushkova, M.V. Golub, J. Moll, C.-P. Fritzen, Wave energy trapping and localization in a plate with a delamination. *Smart Materials and Structures*, **21**, 125001, 2012.

[7] E. Glushkov, N. Glushkova, A. Ekhlakov, E. Shapar, An analytically based computer model for surface measurements in ultrasonic crack detection. *Wave Motion*, **43**, 458–473, 2006.

[8] E. Glushkov, N. Glushkova, A. Eremin, R. Lammering, Trapped mode effects in notched plate-like structures. *Journal of Sound and Vibration*, **358**, 142–151, 2015.

[9] E.V. Glushkov, N.V. Glushkova, A.A. Eremin, R. Lammering, Guided wave propagation and diffraction in plates with obstacles: resonance transmission and trapping mode effects. *Physics Procedia*, **70**, 447–450, 2015.

[10] I. Solodov, J. Bai, G. Busse, Resonant ultrasound spectroscopy of defects: Case study of flat-bottomed holes. *Journal of Applied Physics*, **113**, 223512, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

INVESTIGATIONS OF MECHANICAL GUIDED WAVES PROPAGATION IN PIPES REPAIRED LOCALLY BY COMPOSITE PATCHES

S. Yaacoubi[1], W. Ke[1], F. Dahmene[1], and M. El Mountassir[1]

[1] Institut de Soudure, Plateforme RDI CND, 4 Bvd Henri Becquerel, Espace Cormontaigne 57970 Yutz, France

e-mail: {s.yaacoubi, w.ke}@isgroupe.com

Keywords: Structural Health Monitoring, Composites, pipes reparation, Ultrasonic guided waves, Numerical simulations.

Abstract. *Mechanical guided waves are frequently used to non-destructively test and/or monitor large structures, such pipes. These pipes, which are used to transport fluids, are vulnerable to structure defects, such as corrosion and crack. For mainly economic concerns (save time and cost), the areas where defects had occurred are sometimes repaired. The material used for reparation is often composites, which cause a hug attenuation of guided waves. This attenuation can limit drastically the distance of inspection (i. e. the distance between the transducer and the smallest detectable defect). The current work aims to investigate numerically, through finite element method, the influence of composite reparations on guided waves propagation. Various parameters such as the shape of the reparation and the operating frequency are studied. The final purpose of this work is to propose some recommendations on guided waves setup to ensure an efficient use of this nondestructive testing technique.*

1 INTRODUCTION

The need to pipelines for transporting liquids, gases and any chemically stable substance is in continuous increase [1]. The material of most pipelines is Carbone steel. Due to its cost, this material is preferred to stainless steel. However, it is much more vulnerable to corrosions. In order to save maintenance cost and avoid interrupting production, composite repair (see an example given in Figure 1) is used to reinforce the strength of the segments of pipes where corrosions had occurred. Using composite repair, in lieu of replacing pipes or at least the damaged areas especially when a leak is assessed to be imminent, can yield significant economic and environmental benefits [2]. Note that the panel of defect types is large and includes internal and external corrosion, external damage such as dents, gouges, and cracks, as well as manufacturing defects. The problem is that these defects continue to progress even after reparation, but maybe with lower velocities. Assessment of the evolution of the said defects is not a straightforward item since it depends of various factors such as geometry and material characteristics of the repair, the tube as well as the transported fluid, the environmental and operational conditions. Consequently, the remaining lifetime is unpredictable and so, non-destructive testing techniques, which should be adequate with this kind of structures, are needed.

Figure 1 : Photography of a composite pipe repair example

Ultrasonic guided waves (UGW) technique seems to be a good candidate to reach this aim (*i. e.* testing periodically this type of structure). Indeed, UGW technique well-adapted for tubular and large structures [3], and showed its efficiency in detecting defects especially in bare pipes. The application of UGW technique on composite repaired pipe segments should however not be trivial because the propagation of these waves is closely linked to the characteristics of the surrounding medium, more particularly the damping. The aim of the project, from which the current study is extracted, aimed to study the feasibility of the monitoring of the said structure. In other words, is UGW technique able to follow the evolution of the "repaired" defect? If a new defect occurs after reparation, is UGW able to detect it?

The problem of attenuation is accentuated by the lack of other required information such as thickness, composite modulus, filler materials, fiber orientation, shape of the repair, etc. The aim of this paper is to investigate numerically the influence of the main parameters which are the material of the repair, its shape and the frequency of excitation.

The paper consists of 3 sections except the first one which is the introduction. Section two concerns the models built in this study. Section three is devoted to present the main obtained results and their discussions. The last section deals with the conclusions gained from this work.

2 MODELS CONSTRUCTION, AND PROBLEM STATEMENT

2.1 Model setting

One among softwares which are available in our laboratory, and which permits to simulate complex configurations is Comsol Multiphysics [4], which is based on Finite Element Method. This software allows simulating wave's propagation in:

- two dimensions structures (2D) in Cartesian coordinates,
- axisymmetric structures (which is 2D, but in cylindrical coordinate), and
- three dimensions structures (3D)

In all cases, simulations can be carried out in:

- frequency domain, or
- time domain

A post-processing, based mainly on Fourier Transform, is then used to analyse results in the suited domain. There is always preference for 3D simulations for its advantage to mimic the true-life phenomenon. However, the models size is too big, which needs high memory calculators, and the computing time is large. 2D models (whatever in Cartesian of Cylindrical coordinates) are preferred, when it is possible, to overcome this problem. In the present case, 2D calculation permits obtaining quick access to representative results, as well as the ability to run high exciting frequency cases (i. e. small mesh size).

In the present study,

- the piping element possesses an axisymmetric geometry,
- homogeneous material
- the sensor (to receive waves) is axisymmetric, and
- the actuator (to generate waves) is axisymmetric.

Hence, the modelling can be performed in 2D axisymmetric medium. However, we cannot use 2D axisymmetric model by exciting the displacement of T(0,1) mode, which is u_θ-displacement component (the radial and axial displacement are null). The 2D axisymmetric dimension is defined by these displacements.

The solution is to build 2D model combined with unique variable wave equation (similar to the case of wave propagation in liquid, where the acoustic pressure p is the only variable). The plane of propagation is x-y plane, while the only displacement component has distribution dependent on x and y, but supposed uniform along z. Besides, the material properties linked to the shear SH wave (T(0,1) mode in circular tubular structures) should also be carefully set, especially those of anisotropic materials (as different elastic matrix component should be taken into account for wave propagating in different directions). For this reason, the detailed wave equation should be developed to check which parameters should be set to study the influence from composite repair.

2.2 Equations formulation

Based on the Hooke's law which can be described by:

$$\sigma_{kl} = C_{ijkl} e_{ij} \tag{1}$$

Associated with the law of conservation of momentum, the general wave equation is written as below:

$$\rho \frac{\partial^2 u_i}{\partial t^2} - c_{ijkl} \frac{\partial^2 u_l}{\partial x_j \partial x_k} = f_i \tag{2}$$

So for x-y plane SH wave (with z-displacement component) propagating, the wave equation concerning z displacement component is:

$$\rho\frac{\partial^2 u_z}{\partial t^2}-\left(\frac{\partial \sigma_{zz}}{\partial z}+\frac{\partial \sigma_{xz}}{\partial x}+\frac{\partial \sigma_{yz}}{\partial y}\right)=f_z \tag{3}$$

As it is assumed that it is uniform along z direction, so $\frac{\partial \sigma_{zz}}{\partial z}=0$, the equation can be simplified as:

$$\rho\frac{\partial^2 u_z}{\partial t^2}-\left(\frac{\partial \sigma_{xz}}{\partial x}+\frac{\partial \sigma_{yz}}{\partial y}\right)=f_z \tag{4}$$

The nonzero stress components σ_{xz} and σ_{yz} can be calculated with the Hooke's law, associated with material anisotropy. It can be simplified as following:

$$\begin{pmatrix}\sigma_{yz}\\ \sigma_{xz}\end{pmatrix}=\begin{pmatrix}C_{44} & \\ & C_{55}\end{pmatrix}\begin{pmatrix}e_{yz}\\ e_{xz}\end{pmatrix} \tag{5}$$

So the final wave equation is written as:

$$\rho\frac{\partial^2 u_z}{\partial t^2}-\frac{1}{2}\left(C_{44}\frac{\partial^2 u_z}{\partial y^2}+C_{55}\frac{\partial^2 u_z}{\partial x^2}\right)=f_z \tag{6}$$

C_{44} et C_{55} describe the materials to used hereafter.

2.3 Results extraction

We choose to show, as results, the circumferential displacement u_θ and its spatial waveform at different frequencies. The former gives an immediate idea about the distribution of the displacement through the thickness of the pipe as well as the repair. It is a fictitious C-Scan in the axial (z) – radial (r) plan. It is an easy way to show qualitatively the influence of the repair on wave's propagation. The spatial waveforms are however collected in the outer lateral surface to show clearly the u_θ displacement amplitude, as shown by Figure 2. Both representations are complementary. Reflection coefficient will also be plotted.

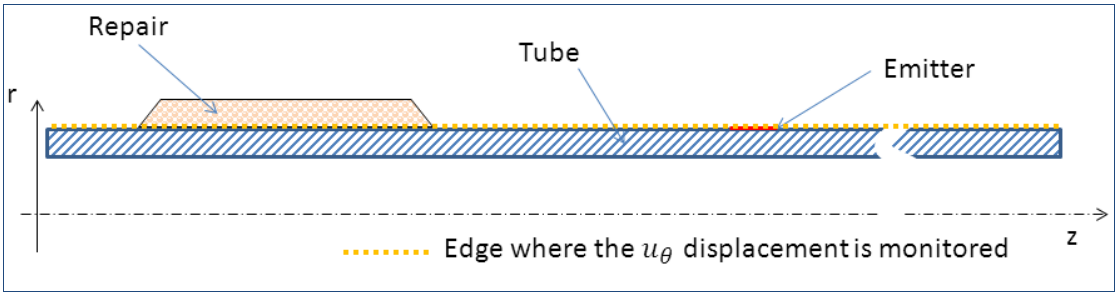


Figure 2 : A not scaled schematic showing the area where the shear displacement u_θ is monitored (along the tube including the repair/tube interface)

3 RESULTS AND DISCUSSIONS

3.1 Displacement monitoring

The first model was a bare pipe. This model simulates also a limit case of pipe with repair (thickness = 0). The analysis of this case will be revisited afterwards, but for the moment it serves to validate the calculations.

After validating the model, we focus now on the investigation of the T(0,1) mode propagation in repaired pipe. As explained previously the shape of the repair is trapezoidal. To understand better the effects of this repair on T(0,1), and be able to analyse properly the obtained results, we add in first step a steel repair. Note that in this case, we make only one change, which is the geometry of the waveguide. The main results are given by Figure 3 and Figure 4.

For the ease of the reader, we recalled the drawing of the repaired tube, as it can be seen in Figure 4. The monitored displacement contains 7 regions, which correspond to:

- (1) ➔ Set of absorbing region, near field and reflected waves by the repair. By comparison with the case of the bare pipe, which can be token as benchmark (not shown here for the sake of brevity), we can see that:
  - the amplitude is slightly different,
  - the amplitude of the reflected signal is not proportional to the other zones, where the displacement amplitude decreases with frequency,
  - ⇨ This confirms that the current repair caused a reflection but it is not too big. Note that reflection coefficient depends at least of:
    - Acoustic impedance of the reflector with regard to that of the main waveguide (tube wall in our case), and
    - Reflector shape.

    In the current case, the acoustic impedance doesn't have any impact, since the repair has the same material as the pipe. So, the reflection should be driven only by the shape, and more particularly by the angle of this conical form ($\emptyset$ in Figure 3). To verify this ascertainment, we studied the case where the repair is rectangular ($\emptyset = 90°$). The correspondent result is not shown here, for the sake of brevity. As it can be seen effortlessly, the reflection coefficient is too big, as what is expected. The other extreme case is where $\emptyset = 0°$, which corresponds naturally to a bare pipe.

- (2) ➔ zone of the pipe just before the repair. The displacement amplitude is constant in whole this zone. Waves reflection is too weak as explained before, otherwise this amplitude should be bigger,
- (3), (4), and (5) ➔ displacement amplitude at the repaired region. They are lower in these zones. However, the transmitted wave amplitude remains almost the same as the incident one. In this case, a little energy was reflected.
  The phenomenon of the decrease of the amplitude at these zones can be explained by the energy flow density (or Poynting Vector). Indeed, the same energy flow density travel in the waveguide (pipe + repair). When the surface increases, the flow decreases.
- (6) ➔ zone of the pipe just after the repair and just before zone (7). Amplitude displacements are constant along this zone, and approximately equal to that of zone (2). This is normal because:
  - waves reflection is quasi-absent, and
  - The medium is non-absorbing.

- (7) ➔ absorbing region

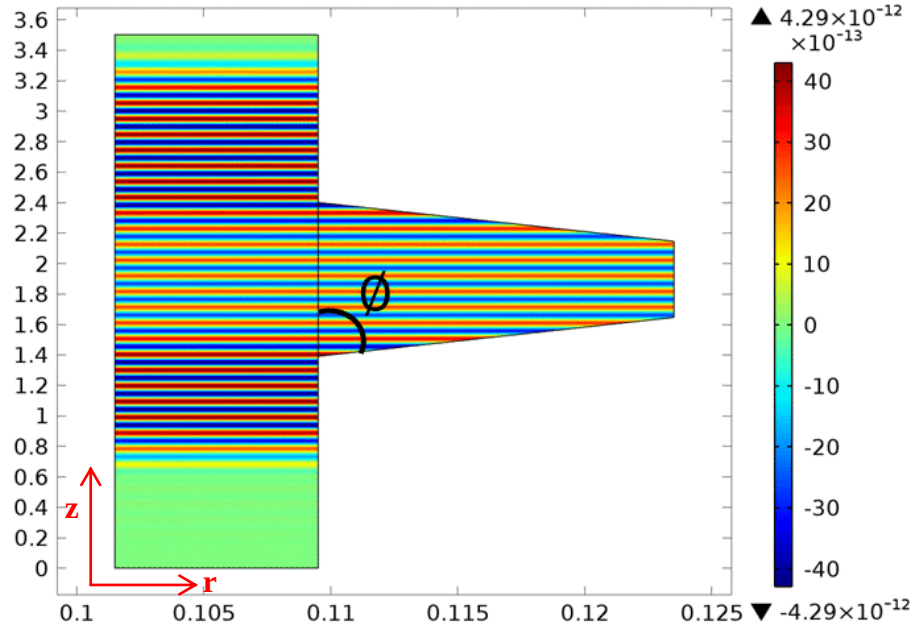


Figure 3 : distribution of u_θ-displacement, at frequency of excitation = 30 kHz, not-scaled figure.

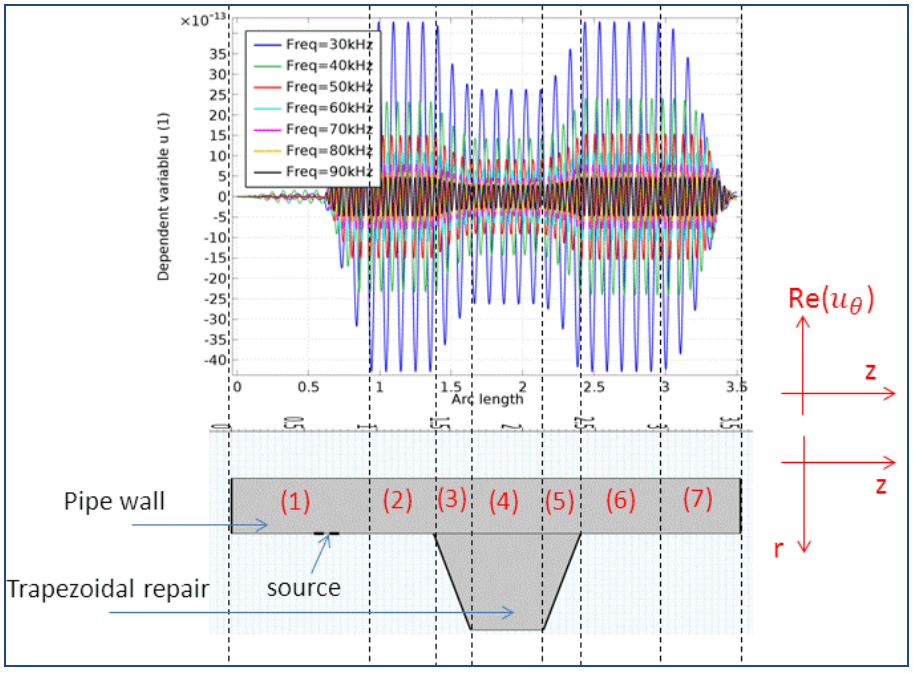


Figure 4 : u_θ-displacement monitored at outer surface of the pipe, for different frequencies (top), and drawing of the repaired pipe (bottom).

Up to this stage of the numerical calculations, the cases studied are: bare pipe and pipe with rectangular and trapezoidal repair shapes. Both the pipes and the repair were steel. From numerical simulation point of view, the material is not absorbing. So, if there is any attenuation, it should be linked to reflection. Hence, and as it was remarked, the waves reflection by the trapezoidal are too weak in front of that caused by the rectangular repair. This was ex-

plained by the repair shape, which is gradually increased in one case (trapezoidal shape) and sudden in the other one (rectangular shape).
Once these studied are achieved, we spent to investigate the case of viscoelastic composite repair. The pipe wall covered by a composite repair will make a bilayer structure. So the relative wave properties will be changed, such as dispersion and attenuation. Concerning the dispersion properties, what will be taken into account are:

- Cut-off frequency for higher order mode
- Mode shape affecting the displacement distribution through thickness (of detected signal)

While the attenuation is mainly linked to the viscosity of the composite material, especially the resin of the fibre reinforced material. The viscosity can be pretty low or extreme high.
To understand the influence of the material on T(0,1) waves propagation, and achieve a more reliable analysis, the investigation will be performed in two steps:

i. pure elastic material, that means non-absorbing material but different from steel. It is the lower limit of damping (viscosity component = 0),
ii. viscoelastic material ➔ absorbing material. The damping is supposed equal *0.15 * 408MPa*, which represents high viscosity.

For the sake of brevity, the obtained results are not shown in the current paper. In the displacement distribution, the maximal amplitude in the elastic case is around 5 times that in the viscoelastic case. This argues the expected influence of the viscoelastic material.

3.2 Reflection coefficient

The final results are gathered in Figure 5. These results confirm what it was already concluded previously. As it can be seen:

- reflections are small in the case of:
  - steel repair with gradually increasing thickness (trapezoidal shape), as well as
  - trapezoidal high damping composite repair, while
- reflections are large where:
  - the composite repair is purely elastic, even with trapezoidal shape
  - and rectangular steel repair,

In all cases, reflections vary greatly with the exciting frequency. It can be easily remarked where they have large coefficient.

In conclusion, reflections depend on:

- material of the repair,
- shape of the repair,
- excitation time frequency.

4 CONCLUSIONS

A numerical study was performed in healthy waveguide (i. e. without any defect). The aim was to investigate the T(0,1) wave mode in a steel pipe coated locally by a composite repair. Due to lack of some characteristics of this material, which were needed as input of the built models, some assumptions were considered:

i. pure elastic repair : this is the lower limit of damping,
ii. viscoelastic repair with high damping.
iii. for both cases (i and ii), the density was token, basing on the literature, the density was token (1000 kg/m3). However, this value depends critically on fiber/matrix ratio, and fiber and matrix types.

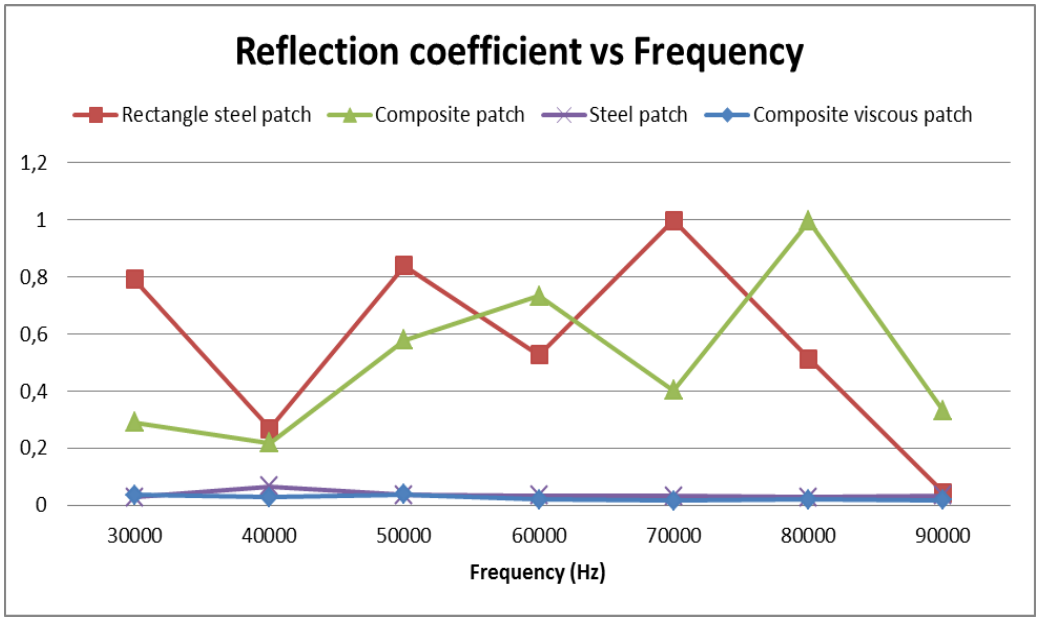


Figure 5 : Reflection coefficient versus frequency for different studied cases : rectangular steel repair, trapezoidal pure elastic repair, trapezoidal steel repair, and trapezoidal high viscoelastic repair.

The repair material, used in this current study, should be between that described in case (i) and in case (ii).

To help verify the results, additional models were built and run:

- Bare pipe
- Pipe with trapezoidal steel repair
- Pipe with steel repair in rectangular form

The main conclusions were found concerning the reflection coefficients based on different models:

- For pipe with pure elastic composite repair (case i), the reflection coefficient may be big as 1 at some exciting frequencies,
- For pipe with high viscoelastic composite repair (case ii), there is few reflections.

The results cited from literature [] are consistent with our results. Note that parameters of this study are different from ours, such as:

- dimensions and the material of the pipe, and
- dimensions, material and shape of the composite repair.

Even both studies are achieved with different parameters, the main conclusion is unique and common: reflections (and consequently transmitted waves) depend considerably and non-linearly (as one can think) of the frequency. Reflection coefficient versus the frequency consists of alternating bands: pass-band where the reflected waves are weak and forbidden-band where they are large. Attention should be paid to this point, since it impacts the attenuation and so, the sensitivity. For future study, it is necessary to have more information about the material properties (elastic or viscoelastic), to perform simulation in smaller frequency step if need.

REFERENCES

[1] http://www.pipeline101.com/why-do-we-need-pipelines/what-do-pipelines-transport

[2] https://www3.epa.gov/gasstar/documents/ll_compwrap.pdf

[3] P. Cawley, M. J. S. Lowe, D. N. Alleyne, B. Pavlakovic and P. Wilcox, "Practical long range guided wave testing: applications to pipes and rail" *Mater. Eval*. 61 66–74, 2003

[4] https://www.comsol.fr/

[5] M. Castaings, and C. Bacon, “Finite element modeling of torsional wave modes along pipes with absorbing materials”, J. Acoust. Soc. Am. 6, 119, 3741 – 3751, 2006

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TWO-STEPS SHAPE OPTIMIZATION ALGORITHM IMPROVING HYDRODYNAMICS STABILITY

Takashi NAKAZAWA

Graduate School of Information Science, Tohoku University
Aoba-ku, Aramaki Aza, Aoba 6-3, Sendai, Miyagi, Japan 980-8579
e-mail: nakazawa@math.is.tohoku.ac.jp

Keywords: Shape Optimization Problem, Hydrodynamics Stability

Abstract. *This proceeding suggests the two-steps shape optimization algorithm improving Hydrodynamics stability, by constructing the existing shape optimization problems Problem 1 and Problem 2. In Problem 1, a dissipation energy is defined as a cost function and the stationary Navier–Stokes problem is used as a main problem. In Problem 2, the maximum value of a real part of the leading eigenvalues is defined as a cost function and the stationary Navier–Stokes problem and its eigenvalue problem are used as main problems. The initial domain is a two dimensional cavity flow Ω_0, where side walls and bottom walls are used as the design boundaries as the cost functions decrease. First, Problem 1 is solved to obtain an optimal domain Ω_1. Second, Problem 2 is demonstrated by using Ω_1 as the initial domain, and an optimal domain Ω_2 is obtained. Finally, eigenvalue problems are solved in Ω_0 and Ω_1, Ω_2 to depict a linear neutral curve and spectrum. As a result, it is confirmed numerically that the critical Reynolds numbers are increased throughout Problem 1 and Problem 2.*

1 INTRODUCTION

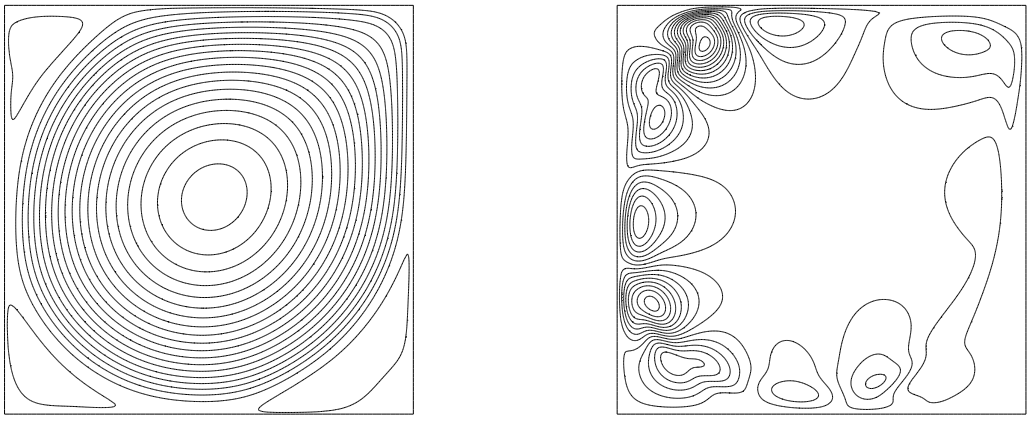
This proceeding suggests a shape optimization method improving Hydrodynamics stability with a regularization technique on the initial domain.

A shape optimization problem is generalized as a problem of optimizing the boundary design of a domain in which a boundary value problem of partial differential equations is defined, where the domain topology is fixed as that of the initial domain. In this problem, cost functions are defined as functionals of the domain and the solution of the boundary value problem. Along with computer development, numerical examinations of shape optimization problems have been conducted. Based on results of these trials, it was noted that the direct application of the gradient method often results in oscillating shapes. However, the oscillation was able to be suppressed by reduction of the degrees of freedom on the boundary.

Subsequently, it was clear that such the oscillation is caused by a lack of smoothness of the first variation of functional. To avoid oscillation without reducing the degrees of freedom, a method using the Laplace operator as smoother was proposed by H. Azegami et al. [1]. This method was called the H^1 gradient method. In the H^1 gradient method, domain variation that minimizes the objective functional is obtained as a solution to a boundary value problem of a linear elastic continuum defined in the design domain.

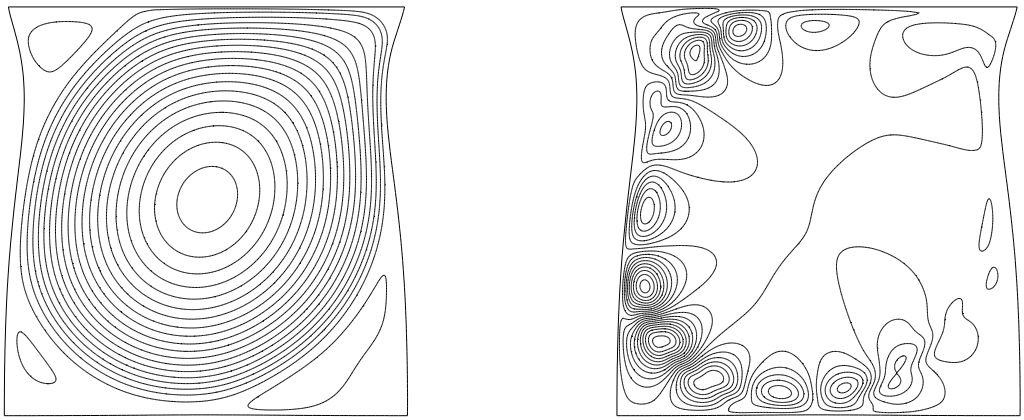
In fluid dynamics, E. Katamine et al. [2] demonstrated the applicability of the H^1 gradient method to a shape optimization problem of an isolated body in a domain defined the stationary Navier–Stokes equations as the governing equations. However, in [2], the stabilities of fluid flows in the initial domain and the optima domain were not discussed. Wherein, T. Nakazawa [3, 4] reported the possibility that the the stability could be controlled by the shape optimization problem. In particular, on the two dimensional Cavity flow Ω_0, using the stationary Navier–Stokes problem as the main problem, the maximizing and minimizing problems of the energy dissipation can make the critical Reynolds number increase and decrease, respectively.

Especially, on an optimal shape Ω_1 in the case of the minimizing problem (Problem 1), figs. 1 and 2 show stream lines of stationary flows and a real part of the leading eigenvalue on Ω_0 and Ω_1 at $\mathrm{Re} = 11500$. Fig. 3 depicts a linear neutral curve on Ω_0 and Ω_1 and a spectrum at $\mathrm{Re} = 11500$, and it is observed that the minimizing problem can make the critical Reynolds number larger.



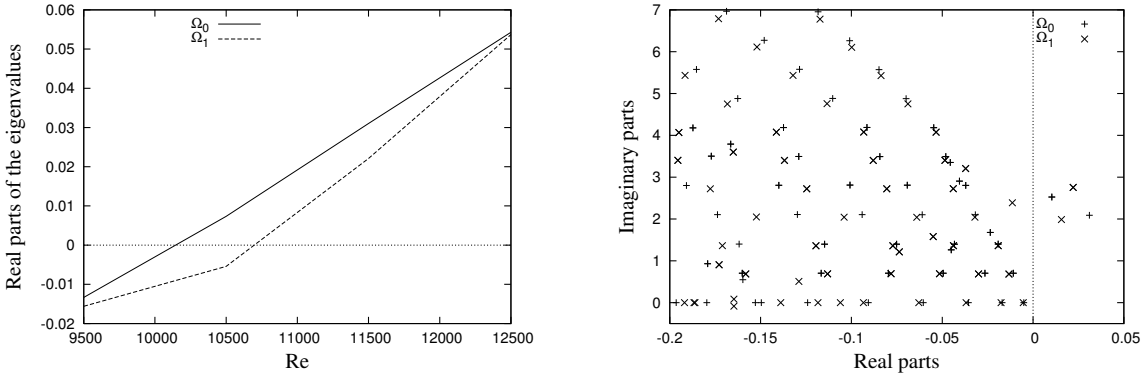
(a) (b)

Figure 1: Stream lines of (a) stationary flows and (b) a part of the leading eigenvalue on Ω_0 at $\mathrm{Re} = 11500$.



(a) (b)

Figure 2: Stream lines of (a) stationary flows and (b) a real part of the leading eigenvalue on Ω_1 at $\mathrm{Re} = 11500$.



(a) (b)

Figure 3: (a) a linear neutral curve on Ω_0 and Ω_1 and (b) a spectrum at $\mathrm{Re} = 11500$.

The results of [3, 4] claim that the shape optimization problem might be able to control Hydrodynamics stability. For the aim to control Hydrodynamics stability more directly, T. Nakazawa and H. Azegami [5] suggested a new pioneering shape optimization method to make the disturbances more stable directly, in which the real parts of the eigenvalue is used as the cost function and the stationary Navier-Stokes problem and its eigenvalue problem are defined as the main problems (Problem 2). In [5], an initial domain is Poissolle flow with a sudden expansion, and the critical Reynolds number increases. In this proceeding, Problem 2 is demonstrated on Ω_0 and Ω_1 at $\mathrm{Re} = 11500$ and each linear stabilities are compared.

2 Formulation of problem

2.1 Initial Domain

A Cartesian coordinate system is used and a position vector is generally denoted by $\boldsymbol{x} = (x,y) \in \mathbb{R}^2$. An initial domain $\Omega_0 = \left\{(x,y) \in \mathbb{R}^2 \mid ([0,1] \times [0,1])\right\}$ is considered, and the top boundary and the wall boundary are set as $\Gamma_{\mathrm{top}} = \{(x,y) \mid 0 \leq x \leq 1,\ y = 1\}$ and $\Gamma_{\mathrm{wall}} = \partial\Omega \backslash \Gamma_{\mathrm{top}}$.

2.2 Domain Variation

Using the initial domain, the domain variation is defined in the following way. Let $\boldsymbol{i} + \boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^2$ be a bi-Lipschitz transform and D be the set of $\boldsymbol{\phi}$, where $\boldsymbol{i}$ denotes identity mapping. For a $\boldsymbol{\phi} \in D$, let a varied domain $\Omega(\boldsymbol{\phi})$ and boundary $\Gamma_{\mathrm{wall}}(\boldsymbol{\phi})$ be defined respectively as $(\boldsymbol{i}+\boldsymbol{\phi})(\Omega_0)$ and $(\boldsymbol{i}+\boldsymbol{\phi})(\Gamma_{\mathrm{wall}})$. The function space of the Fréchet derivative with respect to arbitrary domain variation $\boldsymbol{\phi}$ is defined as

$$X = \left\{\boldsymbol{\phi} \in H^1(\mathbb{R}^2;\mathbb{R}^2) \mid \boldsymbol{\phi} = \boldsymbol{0} \text{ on } \Gamma_{\mathrm{top}}\right\}.$$

The shape derivatives of functionals are obtained as follows. Let ζ be a real-valued function of $\varphi \in C^1(D;H^2(\mathbb{R}^2;\mathbb{R}))$ and $\nabla\varphi(\boldsymbol{\phi})$, and

$$L(\boldsymbol{\phi}) = \int_{\Omega(\boldsymbol{\phi})} \zeta(\varphi(\boldsymbol{\phi}), \nabla\varphi(\boldsymbol{\phi}))dx.$$

Using the shape derivative ζ' of ζ, the shape derivative JL of L is given as

$$JL(\boldsymbol{\phi}) = \int_{\Omega(\boldsymbol{\phi})} \zeta' dx + \int_{\Gamma_{\mathrm{wall}}(\boldsymbol{\phi})} \zeta \boldsymbol{\nu} \cdot \boldsymbol{\psi} d\gamma,$$

where the deformed boundary is denote by $\Gamma_{\mathrm{wall}}(\boldsymbol{\phi}) = \partial\Omega(\boldsymbol{\phi}) \backslash \Gamma_{\mathrm{top}}$ and where $\boldsymbol{\nu}$ denotes an outward unit normal vector on the boundary ([1] Eq. (18), p.274).

2.3 Main problems

The stationary Navier–Stokes equation and the equation of continuity in non-dimensional form are considered, where $(\hat{\boldsymbol{u}}, \hat{p}) \in \hat{U} \times \hat{Q}$ denote non dimensional stationary velocity and pressure and

$$\hat{U} = \left\{\hat{\boldsymbol{u}} = (\hat{u}, \hat{v}) \in H^1(\Omega(\boldsymbol{\phi});\mathbb{R}^2) \mid \hat{\boldsymbol{u}} = \boldsymbol{u}_{\mathrm{D}} \text{ on } \partial\Omega(\boldsymbol{\phi})\right\},$$

$$\hat{Q} = \left\{\hat{q} \in L^2(\Omega(\boldsymbol{\phi});\mathbb{R}) \mid \int_{\Omega(\boldsymbol{\phi})} \hat{q} dx = 0\right\}.$$

The weak form is written as

$$\int_{\Omega(\boldsymbol{\phi})} \left\{ ((\hat{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\hat{\boldsymbol{u}})\cdot\hat{\boldsymbol{w}} - \hat{p}\boldsymbol{\nabla}\cdot\hat{\boldsymbol{w}} + \frac{1}{\mathrm{Re}}(\boldsymbol{\nabla}\hat{\boldsymbol{u}}^{\mathrm{T}})\cdot(\boldsymbol{\nabla}\hat{\boldsymbol{w}}^{\mathrm{T}}) - \hat{q}\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}} \right\} dx = 0,$$

for all $(\hat{\boldsymbol{w}},\hat{q}) \in \hat{W}\times\hat{Q}$, where $(\hat{\boldsymbol{w}},\hat{q})$ represents trial functions for velocity $\hat{\boldsymbol{u}}$ and pressure $\hat{p}$, and

$$\hat{W} = \left\{ \hat{\boldsymbol{w}} \in H^1(\Omega(\boldsymbol{\phi});\mathbb{R}^2) \mid \hat{\boldsymbol{w}} = \boldsymbol{0} \text{ on } \partial\Omega(\boldsymbol{\phi}) \right\}.$$

A set of velocity fields $\bar{\boldsymbol{u}} \in \bar{U}$ and pressure $\bar{p} \in \bar{Q}$ for the disturbance

$$\bar{U} = \left\{ \bar{\boldsymbol{u}} \in H^1(\Omega(\boldsymbol{\phi});\mathbb{C}^2) \mid \bar{\boldsymbol{u}} = \boldsymbol{0} \text{ on } \partial\Omega(\phi) \right\},$$
$$\bar{Q} = \left\{ \bar{q} \in L^2(\Omega(\boldsymbol{\phi});\mathbb{C}) \mid \int_{\Omega(\boldsymbol{\phi})} \bar{q}dx = 0 \right\},$$

are introduced. Based on the linear stability theory, for an eigenvalue $\lambda \in \mathbb{C}$, an eigenvalue equation in weak form is written as

$$\int_{\Omega(\boldsymbol{\phi})} \left\{ \lambda\bar{\boldsymbol{u}}\cdot\bar{\boldsymbol{w}}^c + ((\hat{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\bar{\boldsymbol{u}})\cdot\bar{\boldsymbol{w}} + ((\bar{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\hat{\boldsymbol{u}})\cdot\bar{\boldsymbol{w}}^c - \bar{p}\boldsymbol{\nabla}\cdot\bar{\boldsymbol{w}} + \frac{1}{\mathrm{Re}}(\boldsymbol{\nabla}\bar{\boldsymbol{u}}^{\mathrm{T}})\cdot(\boldsymbol{\nabla}\bar{\boldsymbol{w}}^{c\mathrm{T}}) - \bar{q}^c\boldsymbol{\nabla}\cdot\bar{\boldsymbol{u}} \right\} dx = 0,$$

for all $(\bar{\boldsymbol{w}},\bar{q}) \in \bar{W}\times\bar{Q}$, where $(\bar{\boldsymbol{w}},\bar{q})$ represents trial functions for $\bar{\boldsymbol{u}}$ and $\bar{p}$, and

$$\bar{W} = \left\{ \bar{\boldsymbol{w}} \in H^1(\Omega(\boldsymbol{\phi});\mathbb{C}^2) \mid \bar{\boldsymbol{w}} = \boldsymbol{0} \text{ on } \partial\Omega(\boldsymbol{\phi}) \right\}.$$

3 Shape optimization problem

The minimization problem of the maximum value of the real parts of the leading eigenvalues (Problem 2) is formulated as

Find Ω_2
that minimizes $f = 2\mathrm{Real}[\lambda]$
subject to $(\hat{\boldsymbol{u}},\hat{p},\bar{\boldsymbol{u}},\bar{p}) \in \hat{U}\times\hat{Q}\times\bar{U}\times\bar{Q}$ such that Eqs. (1) and (1) ,

where $\lambda,\bar{\boldsymbol{u}}$ depict an eigenvalue and an eigenfunction for the disturbance with the maximum value of the real parts of the leading eigenvalues. The shape derivative of the Lagrange function L is evaluated by application of the Lagrange multiplier method, where L is written as

$$\begin{aligned} L(\hat{\boldsymbol{u}},\hat{p},\hat{\boldsymbol{w}},\hat{q},\lambda,\bar{\boldsymbol{u}},\bar{p},\bar{\boldsymbol{w}},\bar{q}) = & \\ & 2\mathrm{Real}[\lambda] \\ & - \int_{\Omega(\boldsymbol{\phi})} \left\{ ((\hat{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\hat{\boldsymbol{u}})\cdot\hat{\boldsymbol{w}} - \hat{p}\boldsymbol{\nabla}\cdot\hat{\boldsymbol{w}} + \frac{1}{\mathrm{Re}}(\boldsymbol{\nabla}\hat{\boldsymbol{u}}^{\mathrm{T}})\cdot(\boldsymbol{\nabla}\hat{\boldsymbol{w}}^{\mathrm{T}}) - \hat{q}\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}} \right\} dx \\ & - \int_{\Omega(\boldsymbol{\phi})} \left\{ h(\hat{\boldsymbol{u}},\bar{\boldsymbol{u}},\bar{p},\bar{\boldsymbol{w}},\bar{q}) + h(\hat{\boldsymbol{u}},\bar{\boldsymbol{u}},\bar{p},\bar{\boldsymbol{w}},\bar{q})^c \right\} dx, \end{aligned}$$

where

$$\begin{aligned} h(\hat{\boldsymbol{u}},\bar{\boldsymbol{u}},\bar{p},\bar{\boldsymbol{w}},\bar{q}) = & \lambda\bar{\boldsymbol{u}}\cdot\bar{\boldsymbol{w}} + ((\hat{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\bar{\boldsymbol{u}})\cdot\bar{\boldsymbol{w}} + ((\bar{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\hat{\boldsymbol{u}})\cdot\bar{\boldsymbol{w}} \\ & - \bar{p}\boldsymbol{\nabla}\cdot\bar{\boldsymbol{w}} + \frac{}{\mathrm{e}}(\boldsymbol{\nabla}\bar{\boldsymbol{u}}^{\mathrm{T}})\cdot(\boldsymbol{\nabla}\bar{\boldsymbol{w}}^{c\mathrm{T}}) - \bar{q}^c\boldsymbol{\nabla}\cdot\bar{\boldsymbol{u}}. \end{aligned}$$

The shape derivative JL for L is taken to obtain a sensitivity and main problem (Eqs. (1) and (1)) and its adjoint problems considering KKT condition, and adjoint equations for eq. (1) is obtained as

$$\int_{\Omega(\boldsymbol{\phi})} \Big\{ -\lambda \bar{\boldsymbol{u}}' \cdot \bar{\boldsymbol{w}}^c - ((\bar{\boldsymbol{w}}^c \cdot \boldsymbol{\nabla})\hat{\boldsymbol{u}}) \cdot \bar{\boldsymbol{u}}' + \left((\nabla \hat{\boldsymbol{u}}^{\mathrm{T}})\bar{\boldsymbol{w}}^c\right) \cdot \bar{\boldsymbol{u}}' - \bar{p}' \boldsymbol{\nabla} \cdot \bar{\boldsymbol{w}}^c + \frac{1}{\mathrm{Re}} (\boldsymbol{\nabla} \bar{\boldsymbol{u}}'^{\mathrm{T}}) \cdot (\boldsymbol{\nabla} \bar{\boldsymbol{w}}^{c\mathrm{T}}) - \bar{q}^c \boldsymbol{\nabla} \cdot \bar{\boldsymbol{u}}' \Big\} dx = 0,$$

for all $(\bar{\boldsymbol{u}}', \bar{p}') \in \bar{U} \times \bar{Q}$, where $(\cdot)'$ represents shape derivative. And adjoint equations for eq. (1) is written as

$$\int_{\Omega(\boldsymbol{\phi})} \Big\{ -((\hat{\boldsymbol{u}} \cdot \boldsymbol{\nabla})\hat{\boldsymbol{w}}) \cdot \hat{\boldsymbol{u}}' + \left((\nabla \hat{\boldsymbol{u}}^{\mathrm{T}})\hat{\boldsymbol{w}}\right) \cdot \hat{\boldsymbol{u}}' - \hat{p}' \boldsymbol{\nabla} \cdot \hat{\boldsymbol{w}} + \frac{1}{\mathrm{Re}} (\boldsymbol{\nabla} \hat{\boldsymbol{u}}'^{\mathrm{T}}) \cdot (\boldsymbol{\nabla} \hat{\boldsymbol{w}}^{\mathrm{T}}) - \hat{q} \boldsymbol{\nabla} \cdot \hat{\boldsymbol{u}}' + 2\mathrm{Real}\left[-((\bar{\boldsymbol{u}} \cdot \boldsymbol{\nabla})\bar{\boldsymbol{w}}) \cdot \hat{\boldsymbol{u}}' + \left((\nabla \bar{\boldsymbol{u}}^{\mathrm{T}})\bar{\boldsymbol{w}}\right) \cdot \hat{\boldsymbol{u}}' \right] \Big\} dx = 0,$$

for all $(\hat{\boldsymbol{u}}', \hat{p}') \in \hat{U} \times \hat{Q}$. Solving these problems, substitute main variables $(\hat{\boldsymbol{u}}, \hat{p}, \bar{\boldsymbol{u}}, \bar{p})$ and adjoint variables $(\hat{\boldsymbol{w}}, \hat{q}, \bar{\boldsymbol{w}}, \bar{q})$ into JL, JL is reduced to

$$JL = \int_{\Gamma_{wall}(\boldsymbol{\phi})} G \boldsymbol{\psi} \cdot \boldsymbol{\nu} d\gamma,$$

where G represents the sensitivity for Problem 2

$$G = -\frac{1}{\mathrm{Re}} (\boldsymbol{\nabla} \hat{\boldsymbol{u}}'^{\mathrm{T}}) \cdot (\boldsymbol{\nabla} \hat{\boldsymbol{w}}^{\mathrm{T}}) + 2\mathrm{Real}\left[-\frac{1}{\mathrm{Re}} (\boldsymbol{\nabla} \bar{\boldsymbol{u}}'^{\mathrm{T}}) \cdot (\boldsymbol{\nabla} \bar{\boldsymbol{w}}^{\mathrm{T}}) \right].$$

See more detail in Section 3 of T. Nakazawa and H. Azegami [5]. For reshaping a domain, H^1 gradient method is used .

4 RESULTS and CONCLUSIONS

Problem 2 is addressed on Ω_0 and Ω_1. Fig. 4 depicts cost function f with reshaping steps, and Ω_{real} and Ω_2 represent f for using Ω_0 and Ω_1 as initial domains. From fig. 4, in the case of using Ω_0 as an initial domain, f is not decreasing, and on the other hand f using Ω_1 as an initial domain is minimizing. As a result, Problem 1 might regularize a geometrical shape of Ω_0 for Problem 2.

Fig. 5 shows stream lines of stationary flows and a real part of the leading eigenvalue on Ω_2 at $\mathrm{Re} = 11500$, and fig. 6 a spectrum at $\mathrm{Re} = 11500$. Finally, a linear neutral curve on Ω_0 and Ω_1, Ω_2 is shown in fig. 7, from which it is confirmed numerically that the critical Reynolds numbers are increased throughout Problem 1 and Problem 2.

By the way, the sensitivity derived in Problem 2 is evaluated by using the only disturbance with a maximum value of real parts of the leading eigenvalues, and after the second bifurcation, Problem 2 ignores all the disturbances except the most unstable one, which is not reasonable in the sense of Hydrodynamics stability. Therefore, it is needed to construct a shape optimization problem considering all the unstable disturbances as one of future works.

Acknowledgment The present study was supported by JSPS KAKENHI（26310203）.

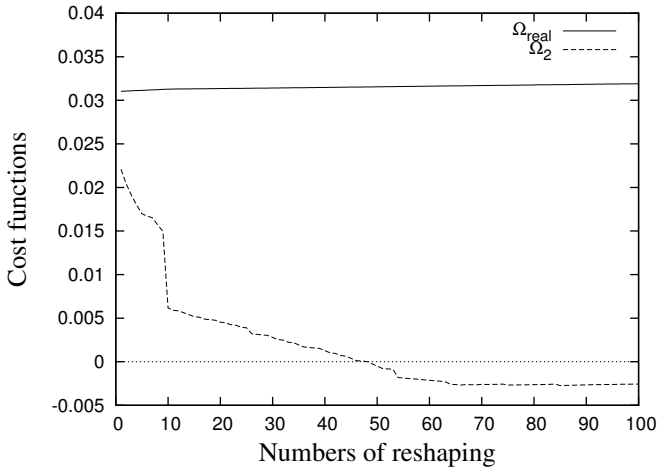


Figure 4: Cost functions with reshaping steps on Ω_0 and Ω_1 at $\mathrm{Re} = 11500$.

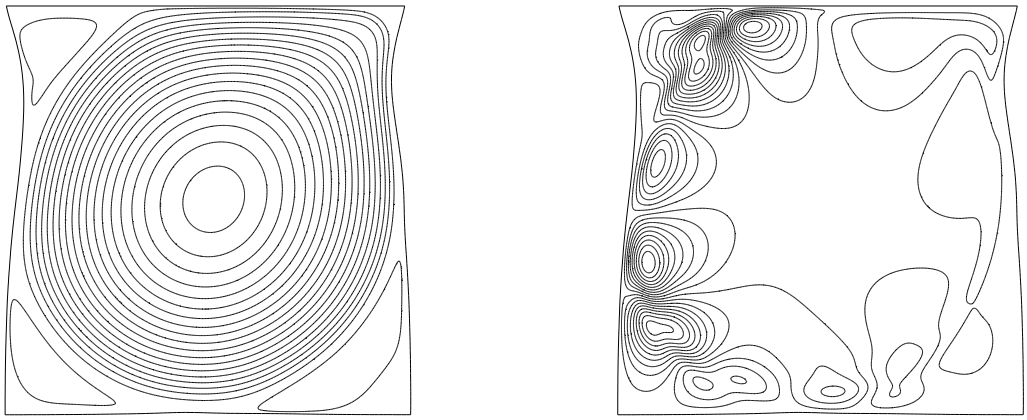


Figure 5: (a) Stream lines of (a) stationary flows and (b) a real part of the leading eigenvalue on Ω_2 at $\mathrm{Re} = 11500$.

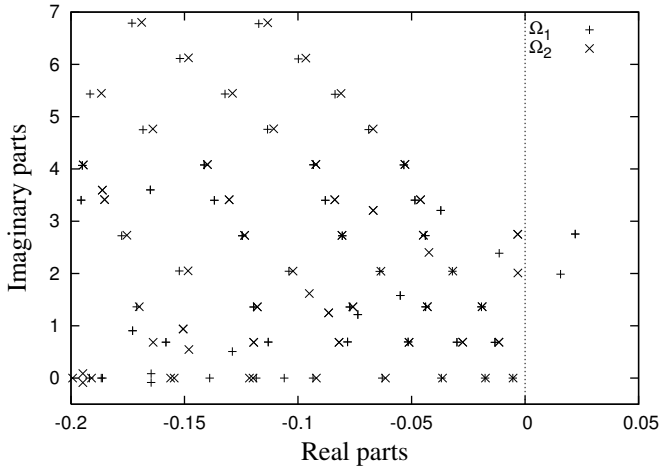


Figure 6: Spectrum on and Ω_2 at $\mathrm{Re} = 11500$.

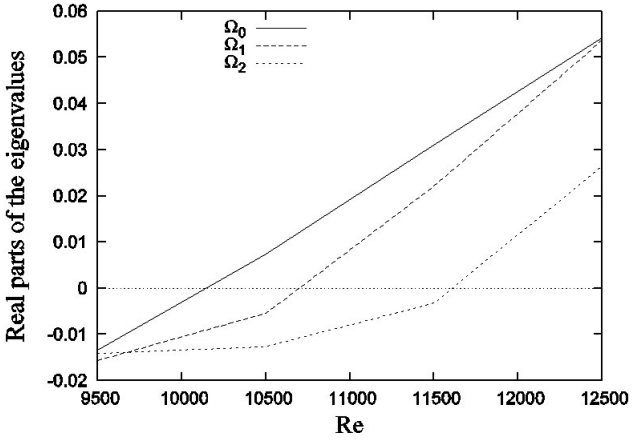


Figure 7: Cost functions with reshaping steps on Ω_0, Ω_1 and Ω_2 at Re $= 11500$.

REFERENCES

[1] Azegami, H., Wu, Z.:Domain Optimization Analysis in Linear Elastic Problems: Approach Using Traction Method. JSME Int. J. Ser. A **39**, 272-278 (1996)

[2] Katamine, E., Nagatomo, Y., Azegami, H.: Shape optimization of 3D viscous flow fields. Inverse Prob. in Sci. and Eng. **17**, 105–114 (2009)

[3] Nakazawa, T.: Shape Optimization of the maximizing problem of the dissipation energy and its effect on Hydrodynamic stability. the 15th European Turbulence Conference 2015, Proceedings, **477**, (2015)

[4] Nakazawa, T.: Increasing the Critical Reynolds number by maximizing dissipation energy problem. the Int. J. of heat and fluid flow, Special Issue, 2015.(in print)

[5] Nakazawa, T., Azegami, H.:Shape Optimization Method improving Hydrodynamic Stability. Jpn. J. of Indust. and Appl. Math, **33**, 167-181 (2015)

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON THE CORRECT APPLICATION OF AD CHECKPOINTING TO ADJOINT MPI-PARALLEL PROGRAMS

A.Taftaf, L. Hascoët

INRIA, Sophia-Antipolis, France, {Elaa.Teftef, Laurent.Hascoet}@inria.fr

Keywords: Algorithmic Differentiation, Adjoint, Checkpointing, Message Passing, MPI

Abstract. *Checkpointing is a classical technique to mitigate the overhead of adjoint Algorithmic Differentiation (AD). In the context of source transformation AD with the Store-All approach, checkpointing reduces the peak memory consumption of the adjoint, at the cost of duplicate runs of selected pieces of the code. Checkpointing is vital for long run-time codes, which is the case of most MPI parallel applications. However, the presence of MPI communications seriously restricts application of checkpointing.*
In most attempts to apply checkpointing to adjoint MPI codes (the "popular" approach), a number of restrictions apply on the form of communications that occur in the checkpointed piece of code. In many works, these restrictions are not explicit, and an application that does not respect these restrictions may produce erroneous code.
We propose techniques to apply checkpointing to adjoint MPI codes, that either do not suppose these restrictions, or explicit them so that the end users can verify their applicability. These techniques rely on both adapting the snapshot mechanism of checkpointing and on modifying the behavior of communication calls.
One technique is based on logging the values received, so that the duplicated communications need not take place. Although this technique completely lifts restrictions on checkpointing MPI codes, message logging makes it more costly than the popular approach. However, we can refine this technique to blend message logging and communications duplication whenever it is possible, so that the refined technique now encompasses the popular approach. We provide elements of proof of correction of our refined technique, i.e. that it preserves the semantics of the adjoint code and that it doesn't introduce deadlocks.

1 INTRODUCTION

Adjoint algorithms, and in particular those obtained through the adjoint mode of Automatic Differentiation (AD) [1], are probably the most efficient way to obtain the gradient of a numerical simulation. Given a piece of code P, adjoint AD (with the "Store-All" approach) consists of two successive pieces of code. The first one, the "forward sweep" $\overrightarrow{P}$ computes the original values and stores in memory the overwritten variables needed to compute the gradients. The second one, the "backward sweep" $\overleftarrow{P}$, computes the gradients, using the intermediate values stored as needed. Primarily, i.e. before any form of program optimisation, the adjoint program is simply $\overrightarrow{P}$ followed by a $\overleftarrow{P}$.

Many large-scale computational science applications are parallel programs based on Message-Passing, implemented for instance by using the MPI message passing library. We will call them "MPI programs". MPI programs consist of one or more threads (called "MPI processes") that communicate through message exchanges.

In most attempts to apply checkpointing to adjoint MPI codes (the "popular" approach), a number of restrictions apply on the form of communications that occur in the checkpointed piece of code. In many works, these restrictions are not explicit, and an application that does not respect these restrictions may produce erroneous code.

We propose techniques to apply checkpointing to adjoint MPI codes, that either do not suppose these restrictions, or explicit them so that the end users can verify their applicability. These techniques rely on both adapting the snapshot mechanism of checkpointing and on modifying the behavior of communication calls.

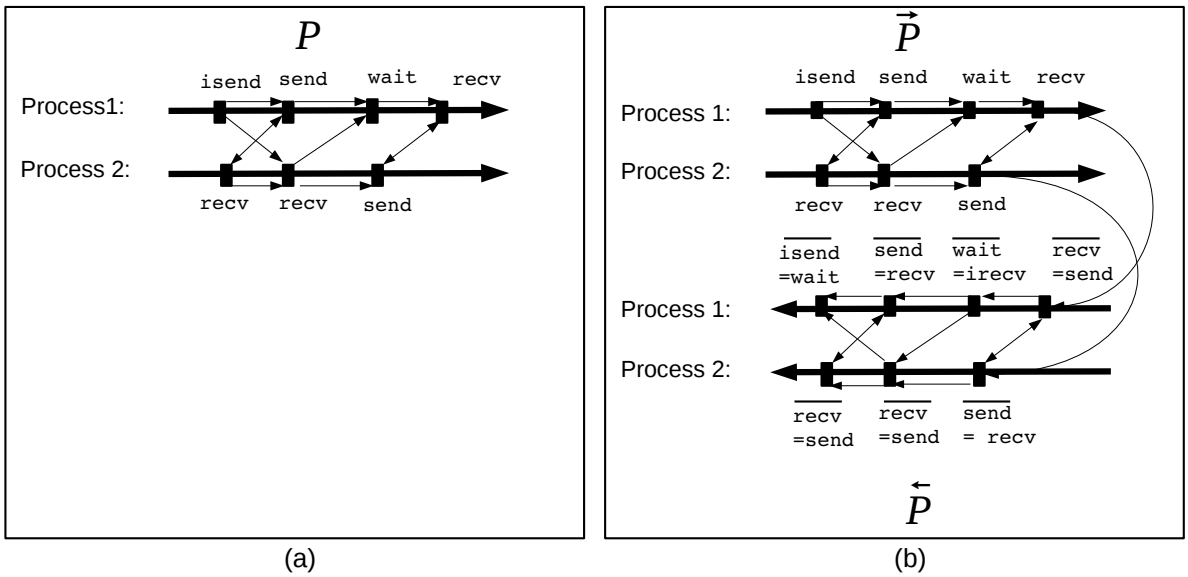


Figure 1: (a) Communications graph of an MPI parallel program with two processes. Thin arrows represent the edges of the communications graph and thick arrows represent the propagation of the original values by the processes. (b) Communications graph of the corresponding adjoint MPI parallel program. The two thick arrows in the top represent the forward sweep, propagating the values in the same order as the original program, and the two thick arrows in the bottom represent the backward sweep, propagating the gradients in the reverse order of the computation of the original values.

1.1 Communications graph of adjoint MPI programs

One commonly used model to study message-passing is the communications graph [[2], pp. 399403], which is a directed graph (see figure 1 (a)) in which the nodes are the MPI communication calls and the arrows are the dependencies between these calls. For simplicity, we omit the `mpi_` prefix from subroutine names and omit parameters that are not essential in our context. Calls may be dependent because they have to be executed in sequence by a same process, or because they are matching `send` and `recv` calls in different processes.

- The arrow from each `send` to the matching `recv` (or to the `wait` of the matching `isend`) reflects that the `recv` (or the `wait`) cannot complete until the `send` is done. Similarly, the arrow from each `recv` to the matching `send` (or to the `wait` of the matching `irecv`) reflects that the `send` will block until the `recv` is done.

- The arrows between two successive MPI calls within the same process reflect the dependency due to the program execution order, i.e. instructions are executed sequentially. In the sequel, we will not show these arrows.

A central issue for correct MPI programs is to be deadlock free. Deadlocks are cycles in the communications graph.

There have been several works on the adjoint of MPI parallel programs [3],[4], [5], [6], [7]. When the original code performs an MPI communication call, the adjoint code must perform another MPI call, which we will call an "adjoint MPI call".

- For instance the adjoint for a receiving call `recv(`b`)` is a *send* of the corresponding adjoint value $\overline{b}$. In practice, this will write as `send(`$\overline{b}$`)` ; $\overline{b} = 0$.

- Symmetrically the adjoint for a sending call `send(`a`)` performs a *receive* of the corresponding adjoint value. In practice this will write as `recv(`tmp`)` ; $\overline{a} += temp$.

This way, the adjoint code will perform a communication of the adjoint value (called "adjoint communication") in the opposite direction of the communication of the primal value, which is what should be done according to the AD model. This creates in $\overleftarrow{P}$ a new graph of communications (see figure 1 (b)), that has the same shape as the communications graph of the original program, except the inversion of the direction of arrows. This implies that if the communications graph of the original program is acyclic, then the communications graph of $\overleftarrow{P}$ is also acyclic. Since $\overrightarrow{P}$ is essentially a copy of P with the same communications structure, the communications graphs of $\overrightarrow{P}$ and $\overleftarrow{P}$ are acyclic if the communications graph of P is acyclic. Since we observe in addition that there is no communication from $\overrightarrow{P}$ to $\overleftarrow{P}$, we conclude that if P is deadlock free, then $\overline{P} = \overrightarrow{P}; \overleftarrow{P}$ is also deadlock free.

1.2 Checkpointing

Storing all intermediate values in $\overrightarrow{P}$ consumes a lot of memory space. In the case of serial programs, the most popular solution is the "checkpointing" mechanism [8] (see figure 2). Checkpointing is best described as a transformation applied with respect to a piece of the original code (a "checkpointed part"). For instance figure 2 (a) and (b) illustrate checkpointing applied to the piece C of a code, consequently written as $U; C; D$.
On the adjoint code of $U; C; D$ (see figure 2 (a)), checkpointing C means in the forward sweep **not** storing the intermediate values during the execution of C. As a consequence, the backward

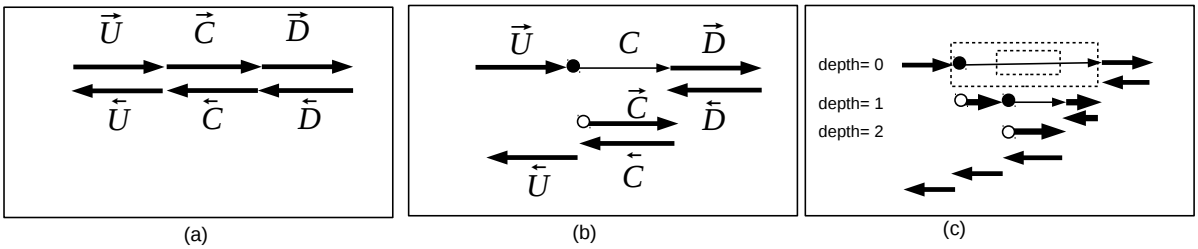


Figure 2: (a) A sequential adjoint program without checkpointing. (b) The same adjoint program with checkpointing applied to the part of code C. The thin arrow reflects that the first execution of the checkpointed code C does not store the intermediate values in the stack. (c) Application of the checkpointing mechanism on two nested checkpointed parts. The checkpointed parts are represented by dashed rectangles.

sweep can execute $\overleftarrow{D}$ but lacks the stored values necessary to execute $\overleftarrow{C}$. To cope with that, the code after checkpointing (see figure 2 (b)) runs the checkpointed piece again, this time storing the intermediate values. The backward sweep can then resume, with $\overleftarrow{C}$ then $\overleftarrow{U}$. In order to execute C twice (actually C and later $\overrightarrow{C}$), one must store (a sufficient part of) the memory state before C and restore it before $\overleftarrow{C}$. This storage is called a *snapshot*, which we represent on figures as a $\bullet$ for taking a snapshot and as a $\circ$ for restoring it. Taking a snapshot "$\bullet$" and restoring it "$\circ$" have the effect of resetting a part of the machine state after "$\circ$" to what it was immediately before "$\bullet$". We will formalize and use this property in the demonstrations that follow. To summarize, for original code $U; C; D$, whose adjoint is $\overrightarrow{U}; \overrightarrow{C}; \overrightarrow{D}; \overleftarrow{D}; \overleftarrow{C}; \overleftarrow{U}$, checkpointing C transforms the adjoint into $\overrightarrow{U}; \bullet; C; \overrightarrow{D}; \overleftarrow{D}; \circ; \overrightarrow{C}; \overleftarrow{C}; \overleftarrow{U}$.
The benefit of checkpointing is to reduce the peak size of the stack in which intermediate values are stored: without checkpointing, this peak size is attained at the end of the forward sweep, where the stack contains $k_U \oplus k_C \oplus k_D$, where k_X is the values stored by code X. In contrast, the checkpointed code reaches two maximums $k_U \oplus k_D$ after $\overrightarrow{D}$ and $k_U \oplus k_C$ after $\overrightarrow{C}$. The cost of checkpointing is twofold: the snapshot must be stored, generally on the same stack, but its size is in general much smaller than k_C. The othor part of the cost is that C is executed twice, thus increasing run time.

1.3 Checkpointing on MPI adjoints

Checkpointing MPI parallel programs is restricted due to MPI communications. In previous works, the "popular" checkpointing approach has been applied in such a way that a checkpointed piece of code always contains both ends of each communication it performs. In other words, no MPI call inside the checkpointed part may communicate with an MPI call which is outside. Furthermore, non-blocking communication calls and their corresponding waits must be both inside or both outside of the checkpointed part. This restriction is often not explicitly mentioned. However, if only one end of a point to point communication is in the checkpointed part, then the above method will produce erroneous code. Consider the example of figure 3 (a), in which only the `send` is contained in the checkpointed part. The checkpointing mechanism duplicates the checkpointed part and thus duplicates the `send`. As the matching `recv` is not duplicated, the second `send` is blocked. The same problem arises if only the `recv` is contained in the checkpointed part (see figure 3 (b)). The duplicated `recv` is blocked. Figure 3 (c) shows the case of a non-blocking communication followed by its `wait`, and only the `wait` is contained in the checkpointed part. This code fails because the repeated `wait` does not correspond to any pending communication.

We propose techniques that adapt checkpointing to MPI programs with point-to-point com-

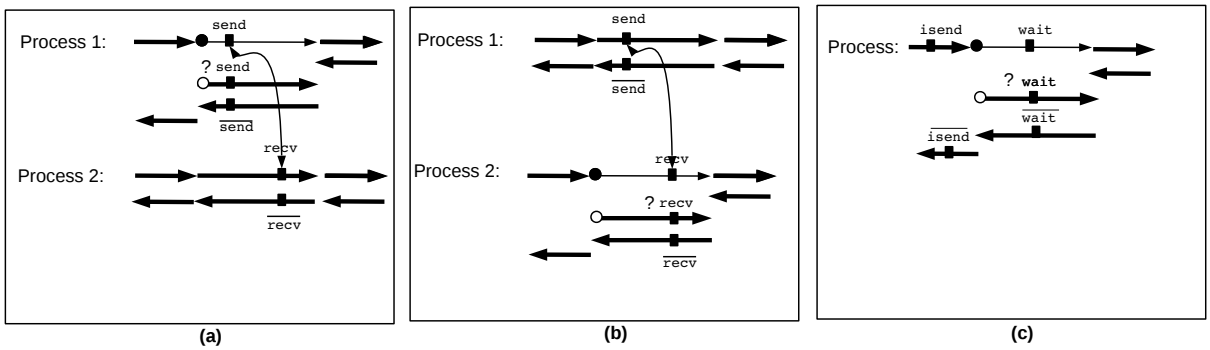


Figure 3: Three examples of careless application of checkpointing to MPI programs, leading to wrong code. For clarity, we separated processes: process 1 on top and process 2 at the bottom. In (a), an adjoint program after checkpointing a piece of code containing only the *send* part of point-to-point communication. In (b), an adjoint program after checkpointing a piece of code containing only the *recv* part of point-to-point communication. In (c), an adjoint program after checkpointing a piece of code containing a `wait` without its corresponding non blocking routine `isend`.

munications. These techniques either do not suppose restrictions on the form of communications that occur in the checkpointed code, or explicit them so that the end user can verify their applicability. One technique is based on logging the values received, so that the duplicated communications need not take place. Although this technique completely lifts restrictions on checkpointing MPI codes, message logging makes it more costly than the popular approach. However, we can refine this technique to replace message logging with communications duplication whenever it is possible, so that the refined technique now encompasses the popular approach. In section 2, we give a proof framework for correction of checkpointed MPI codes, that will give some sufficient conditions on the MPI adapted checkpointing technique so that the checkpointed code is correct. In section 3 , we introduce our MPI adapted checkpointing technique based on message logging. We prove that this technique respects the assumptions of section 2 and thus that it preserves the semantics of the adjoint code. In section 3, we show how this technique may be refined in order to reduce the number of values stored in memory. We prove that the refinement we propose respects the assumptions of section 2 and thus that it preserves the semantics of the adjoint code as well.

2 ELEMENTS OF PROOF

We propose adaptations of the checkpointing method to MPI adjoint codes, so that it provably preserves the semantics of the resulting adjoint code for any choice of the checkpointed part. To this end, we will first give a proof framework of correction of checkpointed MPI codes, that relies on some sufficient conditions on the MPI adapted checkpointing method so that the checkpointed code is correct.

On large codes, checkpointed codes are nested (see figure 2 (c)) , with a nesting level often as deep as the depth of the call tree. Still, nested checkpointed parts are obtained by repeated application of the simple pattern described in figure 2 (b). Specifically, checkpointing applies to any sequence of forward, then backward code (e.g. $\overrightarrow{C}; \overleftarrow{C}$ on figure 2 (b)) independently of the surrounding code. Therefore, it suffices to prove correctness of one elementary application of checkpointing to obtain correctness for every pattern of nested checkpointed parts.

To compare the semantics of the adjoint codes without and with checkpointing, we define the effect $\mathcal{E}$ of a program P as a function that, given an initial machine state σ, produces a

new machine state $\sigma_{new} = \mathcal{E}(P, \sigma)$. The function $\mathcal{E}$ describes the semantics of P. It describes the dependency of the program execution upon all of its inputs and specifies all the program execution results. The function $\mathcal{E}$ is naturally defined on the composition of programs by :
$\mathcal{E}((P_1; P_2), \sigma) = \mathcal{E}(P_2, \mathcal{E}(P_1, \sigma))$.

When P is in fact a parallel program, it consists of several processes p_i run in parallel. Each p_i may execute point-to-point communication calls. We will define the effect $\mathcal{E}$ of one process p. To this end, we need to specify more precisely the contents of the execution state σ for a given process, to represent the messages being sent and received by p. We will call "R" the (partly ordered) collection of messages that will be received (i.e. are expected) during the execution of p. Therefore R is a part of the state σ which is input to the execution of p, and it will be consumed by p. It may well be the case that R is in fact not available at the beginning of p. In real execution, messages will accumulate as they are being sent by other processes. However, we consider R as a part of the input state σ as it represents the communications that are expected by p. Symmetrically, we will call "S" the collection of messages that will be sent during the execution of p. Therefore, S is a part of the state σ_{new} which is output by execution of p and it is produced by p.
We must adapt the definition of $\mathcal{E}$ for the composition of programs accordingly. We explicit the components of σ as follows. The state σ contains:

- W, the values of variables
- R, the collection of messages expected, or "to be received" by p
- S, the collection of messages emitted by p

With this shape of σ, the form of the semantic function $\mathcal{E}$ and the rule of the composition of programs become more complex. Definition of $\mathcal{E}$ on one process p imposes the prefix R_p of R (the messages to be received) that is required by p and that will be consumed by p. Therefore, the function $\mathcal{E}$ applies pattern matching on its R argument to isolate this "expected" part. Whatever remains in R is propagated to the output R. Similarly, S_P denotes the suffix set of messages emitted by p, to be added to S. Formally, we will write this as:
$\mathcal{E}(p, \langle W, R_P \oplus R, S\rangle) = \langle W', R, S \oplus S_P\rangle$
To explicit the rule of code sequence, suppose that p runs pieces of code C and D in sequence, with C expecting incoming received messages R_C and D expecting incoming received messages R_D. Assuming that the effect of C on the state is:
$\mathcal{E}(C, \langle W, R_C \oplus R, S\rangle) = \langle W', R, S \oplus S_C\rangle$
and the effect of D on the state is:
$\mathcal{E}(D, \langle W', R_D \oplus R, S\rangle) = \langle W'', R, S \oplus S_D\rangle$,
then $C; D$ expects received messages $R_C \oplus R_D$ (for the appropriate concatenation operator $\oplus$) and its effect on the state is:
$\mathcal{E}(C; D, \langle W, R_C \oplus R_D \oplus R, S\rangle) = \langle W'', R, S \oplus S_C \oplus S_D\rangle$.

Adjoint programs operate on two kinds of variables. On one hand, the variables of the original primal code are copied in the adjoint code. In the state σ, we will note their values "V". On the other hand, the adjoint code introduces new adjoint variables to hold the derivatives. In the state σ, we will denote their values "$\overline{V}$".
Moreover, adjoint computations with the store-all approach use a stack to hold the intermediate values that are computed and pushed during the forward sweep $\overrightarrow{P}$ and that are popped and used during the backward sweep $\overleftarrow{P}$. We will denote the stack as "k". In the sequel, we will use a

fundamental property of the stack mechanism of AD adjoints, which is that when a piece of code has the shape $\overrightarrow{P}; \overleftarrow{P}$, then the stack is the same before and after this piece of code. To be complete, the state should also describe the sent and received messages corresponding to adjoint values (see section 1.1). As these parts of the state play a very minor role in the proofs, we will omit them. Therefore, we will finally split states σ of a given process as: $\sigma = \langle V, \overline{V}, k, R, S\rangle$.
For our needs, we formalize some classical semantic properties of adjoint programs. These properties can be proved in general, but this is beyond the scope of this paper. We will consider these properties as axioms.

- Any "copied" piece of code X (for instance C) that occurs in the adjoint code operates only on the primal values V and on the R and S communication sets, but not on $\overline{V}$ nor on the stack. Formally, we will write:
  $\mathcal{E}(X, \langle V, \overline{V}, k, R_X \oplus R, S\rangle) = \langle V_{new}, \overline{V}, k, R, S \oplus S_X\rangle$, with the output V_{new} and S_X depending only on V and on R_X.

- Any "forward sweep" piece of code $\overrightarrow{X}$ (for instance $\overrightarrow{U}$, $\overrightarrow{C}$ or $\overrightarrow{D}$) works in the same manner as the original or copied piece X, except that it also pushes on the stack new values noted δk_X, which only depend on V and R_X. Formally, we will write:
  $\mathcal{E}(\overrightarrow{X}, \langle V, \overline{V}, k, R_X \oplus R, S\rangle) = \langle V_{new}, \overline{V}, k \oplus \delta k_X, R, S \oplus S_X\rangle$

- Any "backward sweep" piece of code $\overleftarrow{X}$ (for instance $\overleftarrow{U}$, $\overleftarrow{C}$ or $\overleftarrow{D}$), on one hand operates on the adjoint variables $\overline{V}$ and, on the other hand, uses exactly the top part of the stack δk_X that was pushed by $\overrightarrow{X}$. In the simplest AD model, δk_X is used to restore the values V that were held by the primal variables immediately before the corresponding forward sweep $\overrightarrow{X}$. There exists a popular improvement in the AD model in which this restoration is only partial, restoring only a subset of V to their values before $\overrightarrow{X}$. This improvement (called TBR) guarantees that the non-restored variables have no influence on the following adjoint computations and therefore need not be stored. The advantage of TBR is to reduce the size of the stack. Without loss of generality, we will assume in the sequel that the full restoration is used, i.e. no TBR is used. With the TBR mechanism, the semantics of the checkpointed program are preserved at least for the output $\overline{V}$ so that this proof is still valid. Formally, we will write:
  $\mathcal{E}(\overleftarrow{X}, \langle V, \overline{V}, k \oplus \delta k_X, R, S\rangle) = \langle V_{new}, \overline{V}_{new}, k, R, S\rangle$, where V_{new} is equal to the value V before running $\overrightarrow{X}$ (which is achieved by using δk_X and V) and $\overline{V}_{new}$ depends only on V, $\overline{V}$ and δk_X.

- A "take snapshot" operation "$\bullet$" for a checkpointed piece C does not modify V nor $\overline{V}$, expects no received messages, and produces no sent messages. It adds into the stack enough values Snp_C to permit a later re-execution of the checkpointed part. Formally, we will write :
  $\mathcal{E}(\bullet, \langle V, \overline{V}, k, R, S\rangle) = \langle V, \overline{V}, k \oplus Snp_C, R, S\rangle$, where Snp_C is a subset of the values in V, thus depending on only V.

- A "restore snapshot" operation "$\circ$" of a checkpointed piece C does not modify $\overline{V}$, expects no received messages and produces no sent messages. It pops from the stack the same set of values Snp_C that the "take snapshot" operation pushed "onto" the stack. This modifies V so that it holds the same values as before the "take snapshot" operation.
  We introduce here the additional assumption that restoring the snapshot may (at least conceptually) add some messages to the output value of R. In particular:

Assumption 1. *The duplicated $recvs$ in the checkpointed part will produce the same values as their original calls.*

Formally, we will write:
$\mathcal{E}(\circ, \langle V, \overline{V}, k \oplus Snp_C, R, S\rangle) = \langle V_{new}, \overline{V}, k, R_C \oplus R, S\rangle$ where V_{new} is the same as V from the state input to the take snapshot.

Our goal is to demonstrate that the checkpointing mechanism preserves the semantics i.e.:

Theorem 1. *For any individual process p, for any checkpointed part C of p, (so that p = $\{U;C:D\}$), for any state σ and for any checkpointing method that respects the* **Assumption 1***:*

$$\mathcal{E}(\{\overrightarrow{U}; \overrightarrow{C}; \overrightarrow{D}; \overleftarrow{D}; \overleftarrow{C}; \overleftarrow{U}\}, \sigma) = \mathcal{E}(\{\overrightarrow{U}, \bullet, C, \overrightarrow{D}, \overleftarrow{D}, \circ, \overrightarrow{C}, \overleftarrow{C}, \overleftarrow{U}\}, \sigma)$$

Proof. We observe that the non-checkpointed adjoint and the checkpointed adjoint share a common prefix $\overrightarrow{U}$ and also share a common suffix $\overleftarrow{C}$;$\overleftarrow{U}$. Therefore, as far as semantics equivalence is concerned, it suffices to compare $\overrightarrow{C}; \overrightarrow{D}; \overleftarrow{D}$ with $\bullet, C, \overrightarrow{D}, \overleftarrow{D}, \circ, \overrightarrow{C}$.
Therefore, we want to show that for any initial state σ_0 :

$$\mathcal{E}(\{\overrightarrow{C}; \overrightarrow{D}; \overleftarrow{D}\}, \sigma_0) = \mathcal{E}(\{\bullet, C, \overrightarrow{D}, \overleftarrow{D}, \circ, \overrightarrow{C}\}, \sigma_0)$$

Since the semantic function $\mathcal{E}$ performs pattern matching on the R_0 part of its σ_0 argument, and the non-checkpointed code has the shape $\{\overrightarrow{C}; \overrightarrow{D}; \overleftarrow{D}\}$, R_0 matches the pattern $R_C \oplus R_D \oplus R$. Therefore, what we need to show writes as:

$$\mathcal{E}(\{\overrightarrow{C}; \overrightarrow{D}; \overleftarrow{D}\}, \langle V_0, \overline{V}_0, k_0, R_C \oplus R_D \oplus R, S_0\rangle) =$$
$$\mathcal{E}(\{\bullet, C, \overrightarrow{D}, \overleftarrow{D}, \circ, \overrightarrow{C}\}, \langle V_0, \overline{V}_0, k_0, R_C \oplus R_D \oplus R, S_0\rangle)$$

We will call σ_2, σ_3 and σ_6 the intermediate states produced by the non-checkpointed code (see

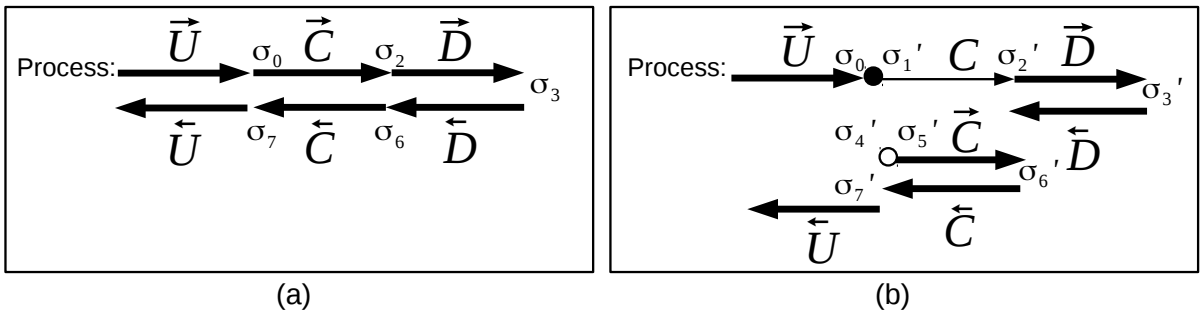


Figure 4: (a) An adjoint program run by one process. (b) The same adjoint after applying checkpointing to C. The figures show the locations (times) in the execution for the successive states σ_i and σ'_i.

figure 4 (a)). Similarly, we call σ'_1, σ'_2, σ'_3, σ'_4, σ'_5, σ'_6 the intermediate states of the checkpointed code (see figure 4 (b)). In other words: $\sigma_2 = \mathcal{E}(\overrightarrow{C}, \sigma_0)$; $\sigma_3 = \mathcal{E}(\overrightarrow{D}, \sigma_2)$; $\sigma_6 = \mathcal{E}(\overleftarrow{D}, \sigma_3)$ and similarly $\sigma'_1 = \mathcal{E}(\bullet, \sigma_0)$; $\sigma'_2 = \mathcal{E}(C, \sigma'_1)$; $\sigma'_3 = \mathcal{E}(\overrightarrow{D}, \sigma'_2)$; $\sigma'_4 = \mathcal{E}(\overleftarrow{D}, \sigma'_3)$; $\sigma'_5 = \mathcal{E}(\circ, \sigma'_4)$; $\sigma'_6 = \mathcal{E}(\overrightarrow{C}, \sigma'_5)$.

Our goal is to show that $\sigma'_6 = \sigma_6$. Considering first the non-checkpointed code, we propagate the state σ by using the axioms already introduced:

$$\begin{aligned}\sigma_2 \doteqdot \mathcal{E}(\overrightarrow{C}, \sigma_0) &= \mathcal{E}(\overrightarrow{C}, \langle V_0, \overline{V_0}, k_0, R_C \oplus R_D \oplus R, S_0\rangle) \\ &= \langle V_2, \overline{V_0}, k_0 \oplus \delta k_C, R_D \oplus R, S_0 \oplus S_C\rangle\end{aligned}$$

with V_2, S_C and δk_C depending only on V_0 and R_C

$$\begin{aligned}\sigma_3 \doteqdot \mathcal{E}(\overrightarrow{D}, \sigma_2) &= \mathcal{E}(\overrightarrow{D}, \langle V_2, \overline{V_0}, k_0 \oplus \delta k_C, R_D \oplus R, S_0 \oplus S_C\rangle) \\ &= \langle V_3, \overline{V_0}, k_0 \oplus \delta k_C \oplus \delta k_D, R, S_0 \oplus S_C \oplus S_D\rangle\end{aligned}$$

with V_3, S_D and δk_D depending only on V_2 and R_D

$$\begin{aligned}\sigma_6 \doteqdot \mathcal{E}(\overleftarrow{D}, \sigma_3) &= \mathcal{E}(\overleftarrow{D}, \langle V_3, \overline{V_0}, k_0 \oplus \delta k_C \oplus \delta k_D, R, S_0 \oplus S_C \oplus S_D\rangle) \\ &= \langle V_2, \overline{V_6}, k_0 \oplus \delta k_C, R, S_0 \oplus S_C \oplus S_D\rangle\end{aligned}$$

with V_2 and $\overline{V}_6$ depending only on V_3, $\overline{V_0}$ and δk_D

Considering now the checkpointed code, we propagate the state σ', starting from $\sigma'_0 = \sigma_0$ by using the axioms already introduced:

$$\sigma'_1 \doteqdot \mathcal{E}(\bullet, \sigma_0) = \mathcal{E}(\bullet, \langle V_0, \overline{V_0}, k_0, R_C \oplus R_D \oplus R, S_0\rangle)$$

The snapshot-taking operation $\bullet$ stores a subset of the original values V_0 in the stack "Snp_C".

$$\sigma'_1 = \langle V_0, \overline{V_0}, k_0 \oplus Snp_C, R_C \oplus R_D \oplus R, S_0\rangle$$

$$\sigma'_2 \doteqdot \mathcal{E}(C, \sigma'_1) = \mathcal{E}(C, \langle V_0, \overline{V_0}, k_0 \oplus Snp_C, R_C \oplus R_D \oplus R, S_0\rangle)$$

The forward sweep of the checkpointed code $\overrightarrow{C}$ is essentially a copy of the checkpointed code C. As the only difference between the two states σ'_1 and σ_0 is the stack k and both C and $\overrightarrow{C}$ don't need the stack during run time ($\overrightarrow{C}$ stores values in the stack, but doesn't use it), the effect of C on the state σ'_1 produces exactly the same output values V_2 and the same collection of sent values S_C as the effect of $\overrightarrow{C}$ on the state σ_0 .

$$\sigma'_2 = \langle V_2, \overline{V_0}, k_0 \oplus Snp_C, R_D \oplus R, S_0 \oplus S_C\rangle$$

The next step is to run $\overrightarrow{D}$:

$$\sigma'_3 \doteqdot \mathcal{E}(\overrightarrow{D}, \sigma'_2) = \mathcal{E}(\overrightarrow{D}, \langle V_2, \overline{V_0}, k_0 \oplus Snp_C, R_D \oplus R, S_0 \oplus S_C\rangle$$

The output state of $\overrightarrow{D}$ uses only the input state's original values V and received values R. As V and R are the same in both σ'_2 and σ_2, the effect of $\overrightarrow{D}$ on the state σ'_2 produces the same variables values V_3, the same collection of messages sent through MPI communications S_D and the same set of values stored in the stack δk_D as the effect of of $\overrightarrow{D}$ on the state σ_2.

$$\sigma'_3 = \langle V_3, \overline{V_0}, k_0 \oplus Snp_C \oplus \delta k_D, R, S_0 \oplus S_C \oplus S_D\rangle$$

Then, the backward sweep starts with the backward sweep of D.

$$\sigma'_4 \doteqdot \mathcal{E}(\overleftarrow{D}, \sigma'_3) \quad = \quad \mathcal{E}(\overleftarrow{D}, \langle V_3, \overline{V_0}, k_0 \oplus Snp_C \oplus \delta k_D, R, S_0 \oplus S_C \oplus S_D \rangle$$

The output state of $\overleftarrow{D}$ uses only its input state's original values V, the values of the adjoint variables $\overline{V}$ and the values stored in the top of the stack δk_D. As V, $\overline{V}$ and δk_D are the same in both σ'_3 and σ_3, the effect of $\overleftarrow{D}$ on the state σ'_3 produces exactly the same variables values V_2 and the same values of adjoint variables $\overline{V}_6$ as the effect of $\overleftarrow{D}$ on the state σ_3.

$$\sigma'_4 \quad = \quad \langle V_2, \overline{V_6}, k_0 \oplus Snp_C, R, S_0 \oplus S_C \oplus S_D \rangle$$

$$\sigma'_5 \doteqdot \mathcal{E}(\circ, \sigma'_4) \quad = \quad \mathcal{E}(\circ, \langle V_2, \overline{V_6}, k_0 \oplus Snp_C, R, S_0 \oplus S_C \oplus S_D \rangle$$

The snapshot-reading operation $\circ$ overwrites V_2 by restoring the original values V_0. According to **Assumption 1**, the snapshot-reading $\circ$ conceptually also restores the collection of values that have been received during the first execution of the checkpointed part R_C.

$$\sigma'_5 \quad = \quad \langle V_0, \overline{V_6}, k_0, R_C \oplus R, S_0 \oplus S_C \oplus S_D \rangle$$

$$\sigma'_6 \doteqdot \mathcal{E}(\overrightarrow{C}, \sigma'_5) \quad = \quad \mathcal{E}(\overrightarrow{C}, \langle V_0, \overline{V_6}, k_0, R_C \oplus R, S_0 \oplus S_C \oplus S_D \rangle$$

The output state after $\overrightarrow{C}$ uses only on the input state's values V and the received values R. As V and R are the same in both σ'_5 and σ_0, the effect of $\overrightarrow{C}$ on the state σ'_5 produces the same original values V_2 and the same set of values stored in the stack δk_C as the effect of $\overrightarrow{C}$ on the state σ_0.

$$\sigma'_6 \quad = \quad \langle V_2, \overline{V_6}, k_0 \oplus \delta k_C, R, S_0 \oplus S_C \oplus S_D \rangle$$

Finally we have $\sigma'_6 = \sigma_6$. □

We have shown the preservation of the semantics at the level of one particular process p_i. The semantics preservation at the level of the complete parallel program P requires to show in addition that the collection of messages sent by all individual processes p_i matches the collection of messages expected by all the p_i. At the level of the complete parallel code, the messages expected by one process will originate from other processes and therefore will be in the messages emitted by other processes.

This matching of emitted and received messages depends on the particular parallel communication library used (e.g. MPI) and is driven by specifying communications, tags, etc. Observing the non-checkpointed code first, we have identified the expected *receives* and produced *sends* $S_U \oplus S_C \oplus S_D$ of each process. Since the non-checkpointed code is assumed correct, the collection of $S_U \oplus S_C \oplus S_D$ for all processes p_i matches the collection of $R_U \oplus R_C \oplus R_D$ for all process p_i.

The study of the checkpointed code for process p_i has shown that it can run with the same expected *receives* $R_U \oplus R_C \oplus R_D$ and produces at the end the same sent values $S_U \oplus S_C \oplus S_D$. This shows that the collected *sends* of the checkpointed version of P matches its collected expected *receives*.

However, matching *sends* with expected *receives* is a necessary but not sufficient condition for correctness. Consider the example of figure 5, in which we have two communications between two processes ("comm A" and "comm B"):

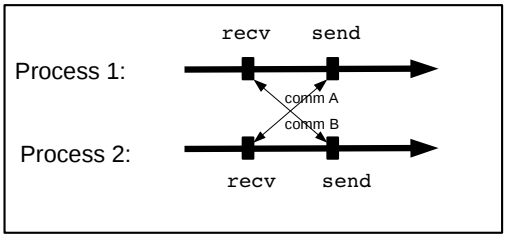


Figure 5: Example illustrating the risk of deadlock if *send* and *receive* sets are only tested for equality.

- The set of messages that process 1 expects to receive R= {comm B}. The set of messages that it will send is S= {comm A}.
- The set of messages that process 2 expects to receive R= {comm A}. The set of messages that it will send is S= {comm B}.

The above required property that the collection of *sends* {comm A, comm B} matches the collection of *receives* {comm A, comm B} is verified. However, this code will fall into a deadlock.
Semantic equivalence between two parallel programs requires not only that collected *sends* match collected *receives* but also that there is no deadlock. If, conversely:

Assumption 2. *the resulting checkpointed code is deadlock free,*

then, the semantics of the checkpointed code is the same as that of its non-checkpointed version.

To sum up, a checkpointing adjoint method adapted to MPI programs is correct if it respects these two assumptions:

Assumption 1. *The duplicated* recvs *in the checkpointed part will collect the same values as their original calls.*

Assumption 2. *The checkpointed code is deadlock free.*

For instance, the "popular" checkpointing approach that we find in most previous works is correct because the checkpointed part which is duplicated is self-contained regarding communications. Therefore, it has always been assumed that the *receive* operations in that duplicated part receive the same value as their original instances. In addition, the duplicated part, being a complete copy of a part of the original code that does not communicate with the rest, is clearly deadlock free.
We believe, however, that this constraint of a self-contained checkpointed part can be alleviated. We will propose a checkpointing approach that respects our two assumptions for any checkpointed piece of code. We will then study a frequent special case where the cost of our proposed checkpointing approach can be reduced.

3 A GENERAL MPI-ADJOINT CHECKPOINTING METHOD

We introduce here a general technique that adapts the checkpointing to the case of MPI parallel programs and that can be applied to any checkpointed piece of code. This technique is

basically inspired by the works that have been done in the context of resilience [9]. Therefore, before detailing this general technique, we will start with a small analogy between the checkpointing in the context of resilience "Resilience-checkpointing" and the checkpointing in the context of AD-Adjoints. In both mechanisms, processes take snapshots of the values they are computing to be able to restart from these snapshots when it is needed. The difference is the reason why taking these snapshots. In the case of "Resilience-checkpointing", the reason is to recover the system from failure, whereas in the case of AD-adjoint, the reason is mostly the reduction of the peak of memory used. Also, the snapshots are called "checkpoints" in the case of "Resilience-checkpointing". Clearly the checkpoints in the context of Adjoint-AD are different from the checkpoints in the context of resilience. We recall that the checkpoints (or also the checkpointed parts) in the case Adjoint-AD are rather intervals of computation that are re-executed when it is needed.
There are two types of checkpointing for resilience: the non-coordinated checkpointing, in which every process takes its own checkpoint independently from the other processes and the coordinated checkpointing in which every process has to coordinate with other process before taking its own checkpoint. We are interested rather by the non-coordinated checkpointing, more precisely by the non-coordinated checkpointing coupled with Message logging. To cope with failure, every process saves in a remote storage checkpoints , i.e. complete images of the process memory. Also, every process saves the messages it receives and every `send` or `recv` event that it performs. In case of failure, only the failed process restarts from its last checkpoint. The other non-failed processes continue their executions normally. The restarted process runs exactly in the same way as before the failure, except that it does not perform any `send` call already done before the failure. The restarted process does not perform either any `recv` call already done, but retrieves instead the value that has been received and stored by the `recv` before the failure.

By analogy, we propose an adaptation of the checkpointing technique to MPI adjoint codes. This adapted technique (we call it "receive-logging") relies on logging every message at the time when it is received.

- During the first execution of the checkpointed part, every communication call is executed normally. However, every *receive* call (in fact its *wait* in the case of non-blocking communication) stores the value it receives into some location local to the process. Calls to *send* are not modified.

- During the duplicated execution of the checkpointed part, every *send* operation does nothing (it is "deactivated"). Every *receive* operation, instead of calling any communication primitive, reads the previously received value from where it has been stored during the first execution.

- The type of storage used to store the received values is First-In-First-Out. This is different from the stack used by the adjoint to store the trajectory.

In the case of nested checkpointed parts, this strategy can either reuse the storage prepared for enclosing checkpointed parts, or free it at the level of the enclosing checkpointed part and re-allocate it at the time of the enclosed checkpoint. This can be managed using the knowledge of the nesting depth of the current checkpointed part.

Notice that this management of storage and retrieval of received values, triggered at the time of the `recv`'s or the `wait`'s, together with nesting depth management, can be implemented by a specialized wrapper around MPI calls, for instance inside the AMPI library [7].

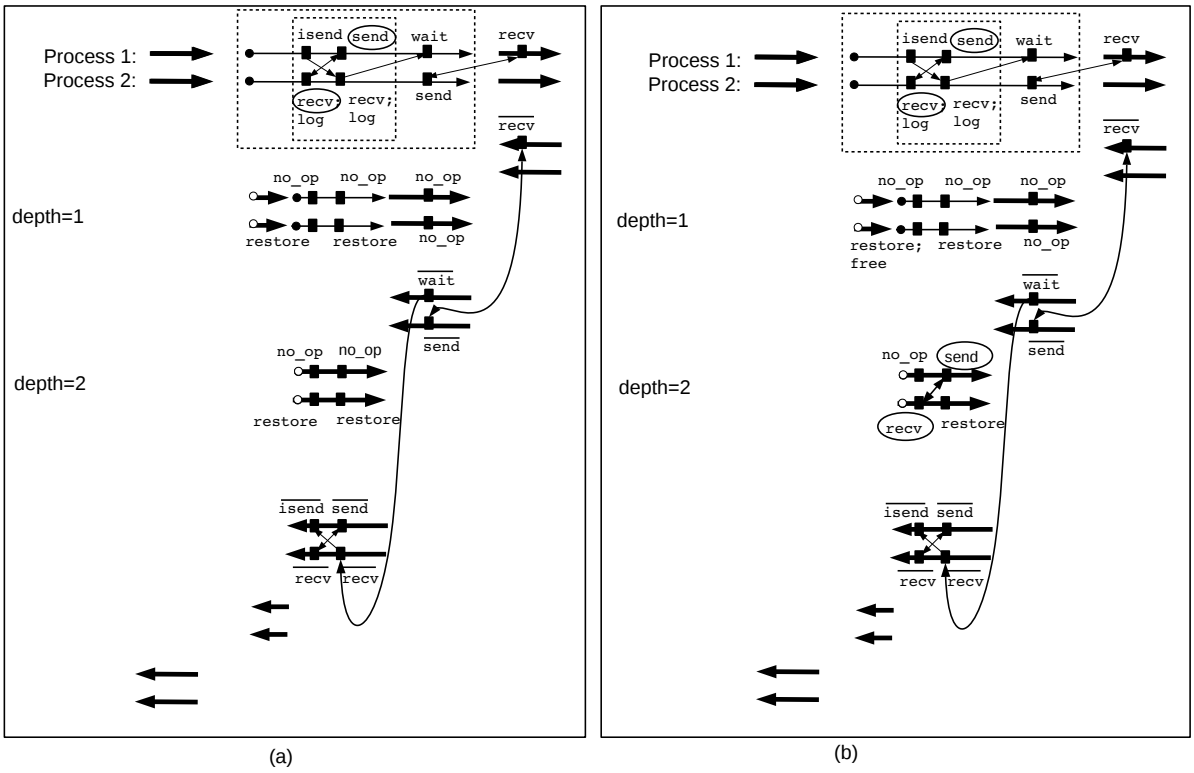


Figure 6: (a) Checkpointing a parallel adjoint program on two nested checkpointed parts by using the receive-logging method. (b) Refinement of the checkpointed code by applying the message re-sending to a `send-recv` pair with respect to the inner checkpointed code which is right-tight

Figure 6 (a) shows the example of two nested checkpointed parts together with an arbitrary communication pattern that straddles across the boundaries of the checkpointed parts.
During execution of the duplicated checkpointed parts, no communication call is made and *receive* operations read from the local storage instead. We can see that the communication pattern of the forward sweep is preserved by checkpointing, the communication pattern of the backward sweep is also preserved, and no communication takes place during duplicated parts.

To show that this strategy is correct, we will check that it verifies the two assumptions of section 2.

3.1 Correctness

By construction, this strategy respects **Assumption 1** because the duplicated *receives* read what the initial *receives* have received and stored.
To verify **Assumption 2** about the absence of deadlocks, it suffices to consider one elementary application of checkpointing, shown in the top part of figure 7. Communications in the checkpointed code occur only in $\overrightarrow{U}$, $\overrightarrow{C}$, $\overrightarrow{D}$ (about primal values) on one hand, and in $\overleftarrow{D}$, $\overleftarrow{C}$, $\overleftarrow{U}$ (about derivatives) on the other hand. The bottom part of the figure 7 shows the communications graph of the checkpointed code, identifying the sub-graphs of each piece of code. Dotted arrows express execution order, and solid arrows express communication dependency. Communications may be arbitrary between $G_{\overrightarrow{U}}$, G_C and $G_{\overrightarrow{D}}$ but the union of these 3 graphs is the same as for the forward sweep of the non-checkpointed code, so it is acyclic by hypothesis.
Similarly, communications may be arbitrary between $G_{\overleftarrow{D}}$, $G_{\overleftarrow{C}}$ and $G_{\overleftarrow{U}}$ but (as $G_{\overrightarrow{C}}$ is by def-

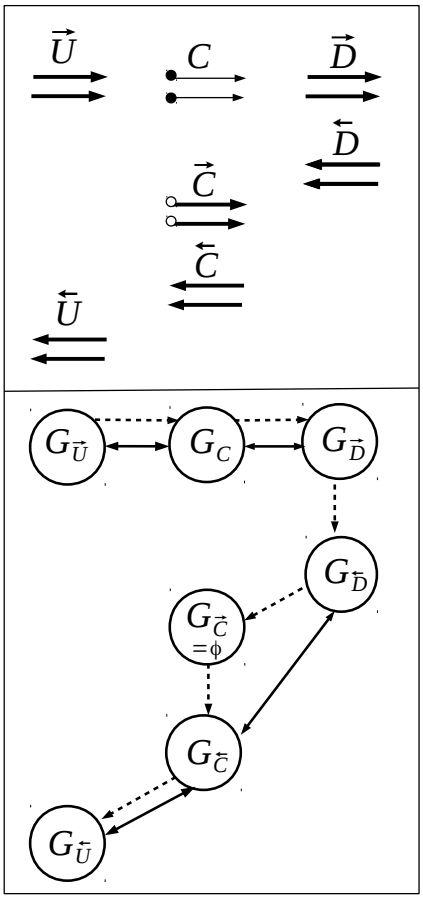


Figure 7: Communications graph of a checkpointed program with pure receive-logging method

inition empty) these graphs are the same as for the non-checkpointed backward sweep. Since we assume that the non-checkpointed code is deadlock free, it follows that the checkpointed code is also deadlock free.

3.2 Discussion

The receive-logging strategy applies for any choice of the checkpointed piece(s). However, it may have a large overhead in memory. At the end of the general forward sweep of the complete program, for every checkpointed part (of level zero) encountered, we have stored all received values, and none of these values has been used and released yet. This is clearly impractical for large codes.
On the other hand, for checkpointed parts deeply nested, the receive-logging has an acceptable cost as stored values are used quickly and their storage space may be released and used by checkpointed parts to come. We need to come up with a strategy that combines the generality of receive-logging with the memory efficiency of an approach based on re-sending.

4 USING MESSAGE RE-SENDING WHENEVER POSSIBLE

We may refine the receive-logging by re-executing communications when possible. The principle is to identify `send-recv` pairs whose ends belong to the same checkpointed part, and to re-execute these communication pairs identically during the duplicated part, thus performing the actual communication twice. Meanwhile, communications with one end not belonging to

the checkpointed part are still treated by receive-logging.

However, the checkpointed part must obey an extra constraint which we will call "right-tight". A checkpointed part is "right-tight" if no communication dependency goes from downstream the checkpointed part back to the checkpointed part, i.e. there is no communication dependency arrow going from D to C in the communications graph of the checkpointed code. For instance, there must no `wait` in the checkpointed part that corresponds with communication call in other process which is downstream (i.e. after) the checkpointed part.

Figure 6 (a) shows an example of two nested checkpointed parts in which the outer checkpointed part is not right-tight, whereas the inner checkpointed part is right-tight since the dependency from the second `recv` of process 2 to the `wait` of the `isend` of process 1 only goes from the checkpoint inside to its outside. In the figure 6 (a), we identify a `send-recv` pair (whose ends are surrounded by circles) that belongs to both nested checkpointed parts. As the outer checkpointed part is not right-tight, we can apply the message re-sending to the `send-recv` pair only with respect to the inner checkpointed part. We see on figure 6 (b) that the `send-recv` pair is re-executed during the execution of the duplicated instance of the inner checkpointed part. As the duplication of the pair `send-recv` is placed between the $\overline{wait}$ of process 1 and the first $\overline{recv}$ of process 2 and since $\overline{wait}$ is a non blocking routine, the duplication of this `send-recv` pair does not create deadlock in the resulting adjoint.

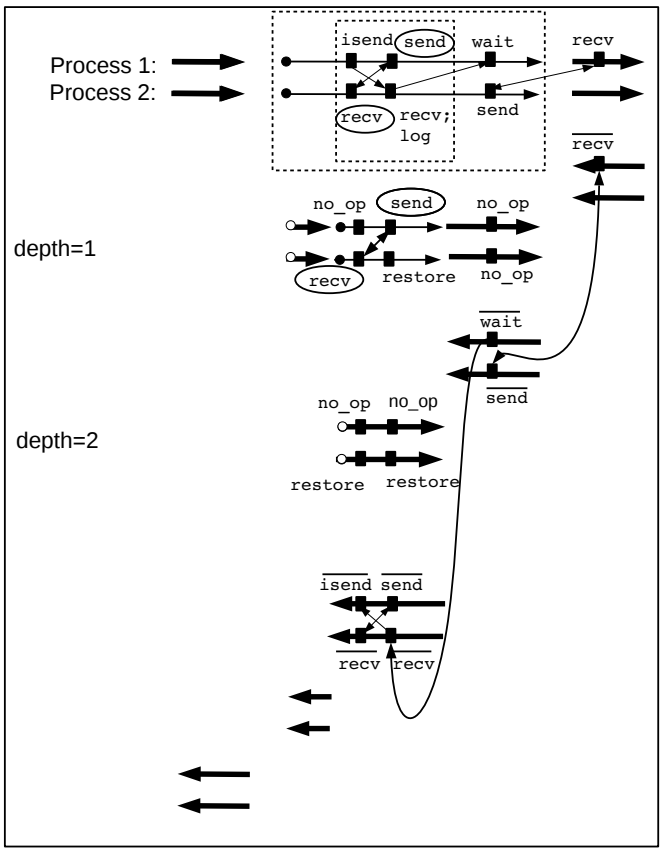


Figure 8: Application of the message re-sending to a `send-recv` pair with respect to the outer checkpointed part which is not right-tight

Figure 8 shows a counterexample, illustrating the danger of applying message re-sending

to a checkpointed part which is not right-tight. We reuse the example of figure 6 (a). Instead of applying the message re-sending to the pair `send-recv` (whose ends are surrounded by circles) with respect to the inner checkpointed code as it is the case in figure 6 (b), we applied the message re-sending to the pair `send-recv` with respect to the outer checkpointed code which is not right-tight. Figure 8 shows the cycle in the communications graph of the resulting adjoint. We see on the figure that, between the $\overline{recv}$ of process 1 and the $\overline{send}$ of process 2 takes place the duplicated run of the outer checkpointed part. In this duplicated run, we find a duplicated `send-recv` pair that causes a synchronization. Execution thus reaches a deadlock, with process 1 blocked on the $\overline{recv}$, and process 2 blocked on the duplicated `recv`.

Only when the checkpointed part is right-tight can we mix message re-sending of communications pairs that are contained in the checkpointed part with receive-logging of the others. The interest is that memory consumption is limited to the (possibly few) logged *receives*. The cost of extra communications is tolerable compared to the gain in memory.

4.1 Correctness

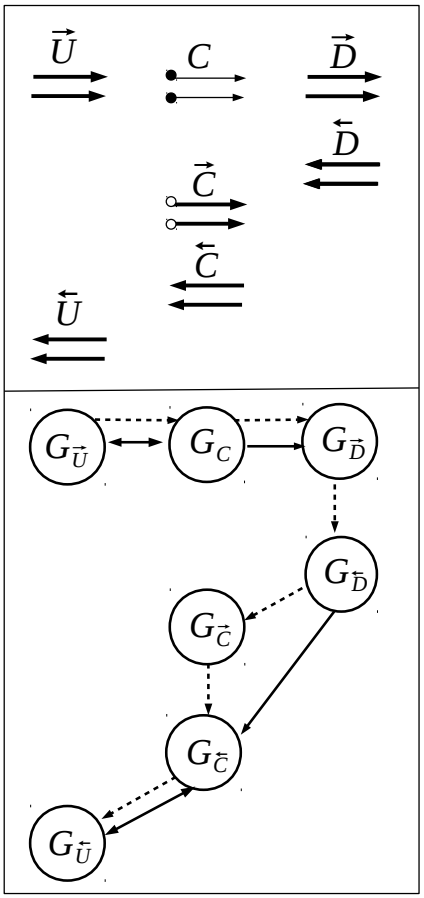


Figure 9: Communications graph of a checkpointed program by using the receive-logging coupled with the message re-sending

The subset of the duplicated *receives* that are treated by receive-logging still receive the same value by construction. Concerning the duplicated `send-recv` pair, the duplicated checkpointed part computes the same values as its original execution (see step from σ'_5 to σ'_6 in section 2). Therefore the duplicated `send` and the duplicated `recv` transfer the same value.

The proof about the absence of deadlocks is illustrated in figure 9. In contrast with the pure receive-logging case, $G_{\overrightarrow{C}}$ is not empty any more because of re-sent communications. $G_{\overrightarrow{C}}$ is a subgraph of G_C and is therefore acyclic. Since the checkpointed part is right-tight, the dependency from G_C to $G_{\overrightarrow{D}}$ and from $G_{\overleftarrow{D}}$ to $G_{\overleftarrow{C}}$ are unidirectional. There is no communication dependency between $G_{\overrightarrow{C}}$ and $G_{\overleftarrow{D}}$ and $G_{\overleftarrow{C}}$ because $G_{\overrightarrow{C}}$ communicates only primal values and $G_{\overleftarrow{D}}$ an $G_{\overleftarrow{C}}$ communicate only derivative values.
Assuming that the communications graph of the non-checkpointed code is acyclic, it follows that:

- Each of $G_{\overrightarrow{U}}$, $G_{\overrightarrow{C}}$, $G_{\overrightarrow{D}}$, $G_{\overleftarrow{D}}$, $G_{\overleftarrow{C}}$ and $G_{\overleftarrow{U}}$ is acyclic.
- Communications may be arbitrary between $G_{\overrightarrow{U}}$ and G_C but since these pieces of code occur in the same order in the non-checkpointed code, and it is acyclic, there is no cycle involved in $(G_{\overrightarrow{U}}; G_C)$. The same argument applies to $(G_{\overleftarrow{C}}; G_{\overleftarrow{U}})$.

Therefore, the complete graph on the bottom of figure 9 is acyclic.

5 DISCUSSION AND FURTHER WORK

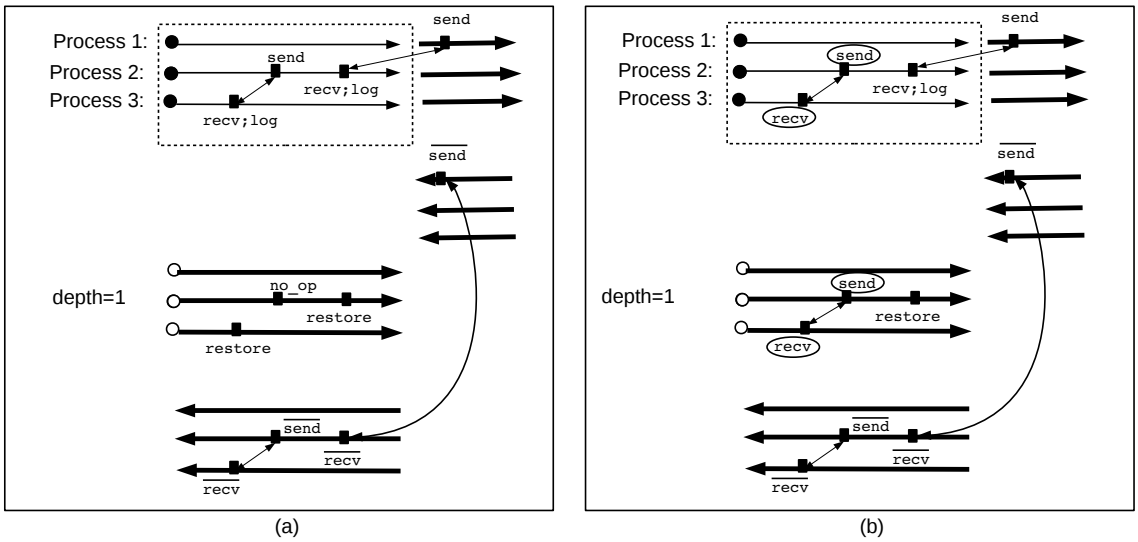


Figure 10: (a) The receive-logging applied to a parallel adjoint program. (b) Application of the message re-sending to a `send-recv` pair with respect to a non-right-tight checkpointed code

We studied checkpointing in the case of MPI parallel programs with point-to-point communications. We proved that any technique that adapts the checkpointing mechanism to MPI parallel programs and that respects some sufficient conditions, is a correct MPI checkpointing technique, in the sense that, the checkpointed code resulting from the application of this MPI checkpointing technique preserves the semantics of the non-checkpointed adjoint code. We introduced, a general MPI checkpointing technique that respects the sufficient conditions for any choice of the checkpointed part. This technique is based on logging the received messages , so that the duplicated communications need not take place. We proposed a refinement that reduces the memory consumption of this general technique by duplicating the communications whenever possible. There are a number of questions that should be studied further:

We imposed a number of restrictions on the checkpointed part in order to apply the refinement. These are sufficient conditions, but it seems they are not completely necessary. Figure 10 shows a checkpointed code which is not right-tight. Still, the application of the message re-sending to a `send-recv` pair (whose ends are surrounded by circles) in this checkpointed part, does not introduce deadlocks in the resulting checkpointed code.

In real codes we may have nested structure of checkpointed parts in which each checkpointed part may be or not right-tight. Applying the message re-sending to only checkpointed parts that are right-tight means that some communication calls will be activated in some levels and deactivated in the other levels. Thus, to implement the refined receive-logging, we need to think about the way we could automatically alternate between these two situations for each communication call. For instance, a *receive* that is deactivated at a level and activated at the level just after has to release its stored value.

Finally, these checkpointing techniques need to be experimented in real codes. It would be interesting to measure the memory consumption of the general checkpointing technique before and after the application of message re-sending.

6 ACKNOWLEDGEMENT

This research is supported by the project "About Flow", funded by the European Commission under FP7-PEOPLE-2012-ITN-317006. See "http://aboutflow.sems.qmul.ac.uk".

REFERENCES

[1] A. Griewank, A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Other Titles in Applied Mathematics, #105. SIAM, 2008.

[2] M. Snir, S. Otto, *The Complete Reference: The MPI Core*. Cambridge, MA, USA: MIT Press, 1998.

[3] P. Hovland, Automatic differentiation of parallel programs *Ph.D. dissertation, University of Illinois at Urbana-Champaign*, 1997.

[4] A. Carle, M. Fagan, Automatically differentiating MPI-1 datatypes: The complete story, in *Automatic Differentiation of Algorithms: From Simulation to Optimization*, ser. Computer and Information Science, G. Corliss, C. Faure, A. Griewank, L. Hascoët, and U. Naumann, Eds. New York: Springer, 2002, ch. 25, pp. 215-222.

[5] P. Heimbach, C. Hill, R. Giering, *An efficient exact adjoint of the parallel MIT general circulation model, generated via automatic differentiation Future Generation Computer Systems 21*. 2005. p. 1356-1371.

[6] U. Naumann, L. Hascoët, C. Hill, P. Hovland, J. Riehme, J. Utke, A Framework for Proving Correctness of Adjoint Message Passing Programs. *In Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Springer. 2008. p. 316-321.

[7] J. Utke, L. Hascoët, P. Heimbach, C. Hill, P. Hovland, U. Naumann, Toward Adjoinable MPI. *Proceedings of the 10th IEEE International Workshop on Parallel and Distributed Scientific and Engineering, PDSEC'09*, 2009.

[8] B. Dauvergne, L. Hascoët, The Data-Flow Equations of Checkpointing in reverse Automatic Differentiation. *International Conference on Computational Science, ICCS 2006, Reading, UK*, 2006.

[9] F. Capello, A. Geist, W. Gropp, S. Kale, B. Kramer, M. Snir, Toward Exascale Resilience: A 2014 Update. *Supercomput.Front. Innovations 1(1) (2014)*.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SPHERICITY: MESH OPTIMIZATION FOR ARBITRARY ELEMENT TOPOLOGY

Pavlos P. Alexias[1,2] **and Eugene De Villiers**[3]

[1]Engys S.R.L
Via del Follatoio 12, Trieste 34148, Italy
e-mail: p.alexias@engys.com

[2] NTUA - Parallel CFD & Optimization Unit (PCOpt)
Heroon Polytechniou 9, Zografou 15780, Greece
e-mail: p.alexias@engys.com

[3] Engys Ltd.
Studio 20, RVPB, John Archer Way, London, SW 18 3SX, U.K
e-mail: e.devilliers@engys.com

Keywords: Sphericity, Mesh Optimization, Constrained Optimization, Shape Optimization, BFGS.

Abstract. *In this paper, a new method for optimizing CFD meshes, based on the usage of a geometric quantity called Sphericity, is introduced. Sphericity is an element-wise property that provides a measure of the shapes deviation from a perfect sphere. When used as an objective in a gradient based framework, it results in a fast, scale invariant, optimization algorithm applicable to arbitrary finite volume element topologies. The methodology is heuristic in nature in that it does not consider known numerical quality metrics explicitly. Our study has however shown it to be exceptionally robust and effective in improving mesh quality, provided that grid topology does not introduce fundamentally concave cells.*
The main focus of this paper is the development of a robust optimization algorithm that can effectively handle complications arising from constraints imposed by surface interactions. Specifically, efficient heuristic and gradient based techniques, for handling difficult mesh topologies with acutely curved surfaces, feature edges and high aspect ratio layer cells are introduced. In combination with the base method, these measures ensure exceptional quality meshes even on industrial scale applications. As a final step, the method is integrated into a general purpose mesh motion framework, allowing the maintenance of high cell quality even during extreme deformation events. Examples and industrial applications in the context of adjoint based shape optimization are presented, illustrating the method's general effectiveness.

1 INTRODUCTION

In the field of computational solutions of partial differential equations (PDEs), the quality of the computational mesh plays a significant role in various aspects of the solution procedure. Important properties such as the rate of convergence of the solution, the fidelity of the results and the robustness of the algorithm are all impacted by the mesh. In this context, the development of an efficient and effective mesh quality optimization algorithm, can provide significant benefits, not just in the context of static meshes, but also for moving mesh problems where degradation of quality is often a severe constraint.

The quality of the mesh can be evaluated using element-wise quality metrics such as, non-inversion, cell skewness, non-orthogonality, cell determinant, volume deviation etc. The mesh optimization algorithm that is presented in this paper aims to calculate a new (optimum) set of mesh vertex positions inside the computational domain that improves these mesh quality metrics. The algorithm does not alter the mesh topology, i.e. no merging, swapping, insertion or subdivision of mesh elements take place. Attempting to directly optimize all important mesh quality metrics simultaneously would lead to an extremely costly and complex multi-objective optimization problem that would require careful calibration. In an attempt to circumvent these issues we introduce without further justification a geometric quantity termed sphericity as the main quality metric and use its maximization as the objective for the mesh optimization. Since sphericity is scale invariant and we only have to optimize for a single objective the process will be much more efficient and robust than would otherwise be the case. It is shown empirically that the maximization of sphericity inside the computational domain results a global improvement to all other quality metrics, provided certain conditions are met.

While this paper deals with the mesh optimization problem specifically for finite volume (FV) meshes, the method can be easily generalized for other areas of applications, such as Finite Elements and surface triangulation problems, with the same efficacy.

2 OPTIMIZATION ALGORITHM

2.1 Sphericity

Sphericity is an element-wise geometric quantity which will be used as the objective function on the mesh optimization problem. Sphericity defines how spherical a geometrical object is and is given by the formula

$$Sph = \frac{\left(\frac{6}{\pi}V_o\right)^{1/3}}{\left(\frac{1}{\pi}S_o\right)^{1/2}} \tag{1}$$

where V_o is the volume and S_o is the surface of the object. In other words sphericity is the ratio between the diameter of a sphere, that has the same volume with the object, and the diameter of a sphere that has the same surface area as the object (considering that for a sphere holds $S = \pi^{1/3}\left(6V\right)^{2/3}$). Given this the sphericity of a sphere is unity by definition and unity is the largest value that any three dimensional object can have. It therefore follows that maximizing this quantity will make elements more spherical. The most important property of spherical elements is isotropy. Highly anisotropic elements can be in general elements with high aspect ratio or/and skewness values that, in a context of computational solutions of PDEs, can affect the quality of the results or even the total convergence of the solution.

In the case of a finite volume mesh, each cell will have a sphericity value calculated as a function of the cell's surface and volume. The goal is to re-position the vertices of the mesh such that the cells will have the maximum sphericity. In order to do so it is necessary to calculate

the derivative of sphericity w.r.t. the position of the points that constitute a cell.

2.2 Sphericity Differentiation

Assuming that a point P is surrounded by M cells, the total sensitivity derivative G_P of that point is the arithmetic mean of the sensitivities from all contributing cells

$$G_P = \frac{1}{M}\sum_{n=1}^{M}\frac{dSph_n}{dP} \tag{2}$$

For each cell that uses point P, the first derivative of sphericity with respect to the position of a point P in the cell is,

$$\frac{dSph_n}{dP} = \frac{\partial\frac{\left(\frac{6}{\pi}V_c\right)^{1/3}}{\left(\frac{1}{\pi}S_c\right)^{1/2}}}{\partial P} = -\frac{6^{1/3}\pi^{1/6}V_c(P)^{1/3}S_c'(P)}{2S_c(P)^{3/2}} + \frac{6^{1/3}\pi^{1/6}V_c'(P)}{3\sqrt{Sc(P)}V_c(P)^{2/3}} \tag{3}$$

where $S_c'(P)$ and $V_c'(P)$ are the partial derivatives of the cell surface and cell volume w.r.t. the position of the point P respectively. Taking into consideration that the cell surface can be written as the sum of the areas of the facets surfaces that constitute the cell, i.e.

$$S_c(P) = \sum_{k=1}^{K} S_f(P)_k \tag{4}$$

where K is the number of faces the cell has. Is it therefore sufficient to calculate the partial derivative of each face w.r.t. the point location. For doing this, the polygon of the face is decomposed into N triangles by defining an approximate face center $C_f(P)$ and forming a triangle from each edge of the polygon. Then it follows that

$$S_f(P) = \frac{1}{2}\sum_{i=1}^{I-1}\sqrt{\left[(P_i - C_f(P)) \times (P_{i+1} - C_f(P))\right]^2} \tag{5}$$

where I is the total number of points of the face polygon. It can be noticed that the face surface is a function of the point position and the face centre which in turn is also a function of the point position. Using the chain rule follows that,

$$\frac{dS_f(P, C_f(P))}{dP} = \frac{\partial S_f}{\partial P} + \frac{\partial S_f}{\partial C_f}\frac{dC_f(P)}{dP} \tag{6}$$

The last step is to write the face center as a function of the points that constitute the face and compute the derivative. In order to avoid cyclical dependencies, the face center is defined as the weighted sum of the edge centers,

$$C_f(P) = \frac{\frac{1}{2}\sum_{i=1}^{I-1}(P_i + P_{i+1})\sqrt{(P_i - P_{i+1})^2}}{\sum_{i=1}^{I-1}\sqrt{(P_i - P_{i+1})^2}} \tag{7}$$

and which can be differentiated to produce,

$$\frac{dC_f(P)}{dP} = \frac{\frac{1}{2}\frac{\partial\sum_{i=1}^{I-1}(P_i + P_{i+1})\sqrt{(P_i - P_{i+1})^2}}{\partial P}}{\sum_{i=1}^{I-1}\sqrt{(P_i - P_{i+1})^2}} - \frac{\frac{1}{2}\sum_{i=1}^{I-1}(P_i + P_{i+1})\sqrt{(P_i - P_{i+1})^2}}{\left[\sum_{i=1}^{I-1}\sqrt{(P_i - P_{i+1})^2}\right]^2} \cdot\frac{\partial\sum_{i=1}^{I-1}\sqrt{(P_i - P_{i+1})^2}}{\partial P} \tag{8}$$

Following the same approach for the derivation of the cell volume, the element is decomposed into K (number of faces) pyramid elements taking as bases the polygonal faces and as the apex the cell center. It thus holds that,

$$V_c(P) = \sum_{k=1}^{K} V_p(P)_k \tag{9}$$

where

$$V_p(P) = \frac{1}{6} \sum_{i=1}^{I-1} (C_f(P) - C_c) \cdot (P_i - C_c) \times (P_{i+1} - C_c)) \tag{10}$$

Since the face center derivative w.r.t. the point locations has already been computed, the calculation of the volume derivative becomes almost straightforward. Having calculated all the component derivatives the total sphericity gradient can now be calculated using equation 3.

2.3 Optimization Method

Having obtained the objective sensitivity derivatives of all the points of the mesh, the next step is to find the new set of positions that maximizes the sphericity value. Thus, for each optimization step, the new position of a point can be calculated as,

$$\vec{P}^{new} = \vec{P}^{old} + \alpha \cdot \vec{G}_P \tag{11}$$

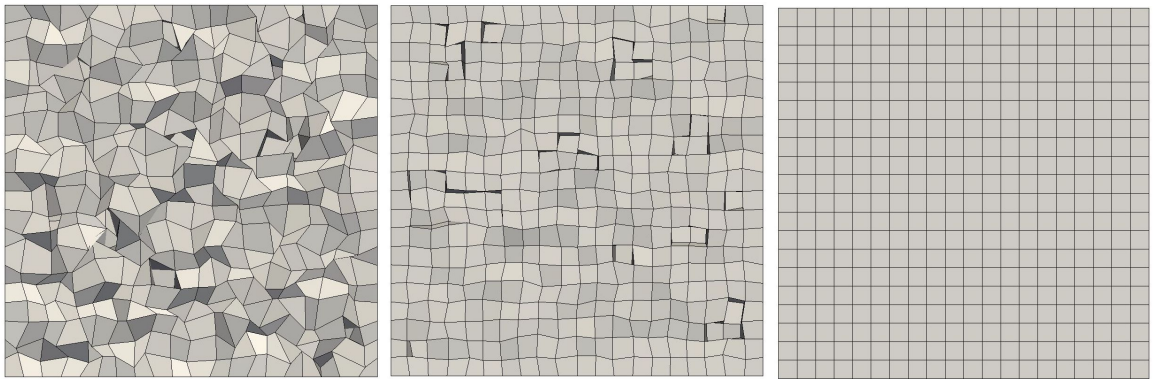
where α is the step length of the optimization cycle. Using a constant value of α is not optimal for the proposed application due to large variations in cell sizes encountered in different CFD meshes (up to 20 orders of magnitude or more). It is therefore necessary to have an adaptive step length. This can be achieved effectively by using second derivatives which also has the added benefit of speeding up the convergence. Based on Nocedal [3] a limited memory BFGS algorithm (LM-BFGS) is constructed for building and updating the inverse Hessian matrix of the second derivatives. Thus, equation 11 can be written as

$$\vec{P}^{new} = \vec{P}^{old} + \lambda H_P^{-1} \cdot \vec{G}_P \tag{12}$$

In general when using Newton and quasi-Newton methods the objective function has to be convex so the optimization algorithm can converge to an optimum value. In other words, the inverse Hessian matrix has to be positive definite. It has been observed that the sphericity function generally behaves in a convex way except from cases where there are negative oriented mesh cells. Therefore, instead of building one Hessian matrix for each point of the mesh which is being optimized, the goal is to create a global Hessian taking into consideration the effect of the displacement of one point, to all other points inside the mesh. Furthermore, the equation 12 can be written as

$$\begin{bmatrix} \vec{P}_1^{new} \\ \vdots \\ \vec{P}_n^{new} \end{bmatrix} = \begin{bmatrix} \vec{P}_1^{old} \\ \vdots \\ \vec{P}_n^{old} \end{bmatrix} + \begin{bmatrix} \lambda_1 & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \lambda_n \end{bmatrix} \cdot \mathbf{H_{global}^{-1}} \cdot \begin{bmatrix} \vec{G}_1 \\ \vdots \\ \vec{G}_n \end{bmatrix} \tag{13}$$

where the H_{global}^{-1} is an $n \times n$ matrix with n to be the total number of points the mesh has. The global matrix is more robust is a sense that is always positive definite which allows the optimization algorithm to reliable converge to a new optimum set of point positions. At last, is



(a) Initial Mesh (b) 2nd Optimization Cycle (c) Optimized Mesh

Figure 1: Mesh optimization based on sphericity value, starting from a low quality mesh (left) and ending up to the optimized one with fully orthogonal cells(right)

fundamental the Wolfe conditions [4] to be satisfied at each and every step of the optimization procedure. This is achieved by assigning a proper step length value λ which will guarantee that the sphericity value, moving towards the direction $\mathbf{H}^{-1}\tilde{\mathbf{G}}$, will actually be increased. Figure 2.3 shows three snapshots of a simple mesh optimization case. Starting from low quality mesh and executing the optimization algorithm, keeping in the same time fixed the boundary points, it results a mesh with fully orthogonal cell elements.

2.4 Constraints

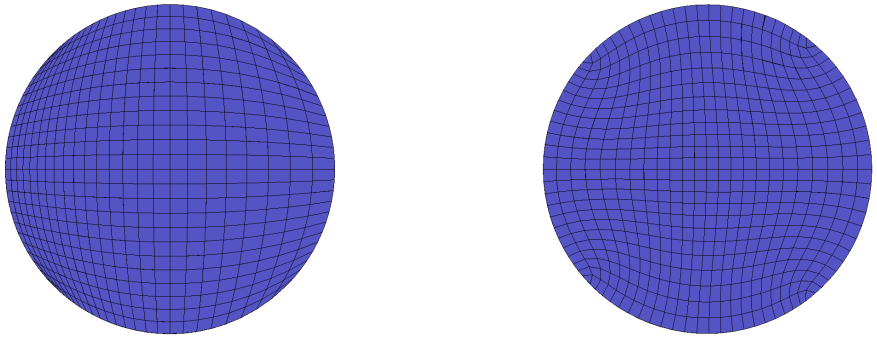
2.4.1 Boundary Points

Interior points are much easier to involve in the optimization as, in general, we want to preserve the original shape of the domain, which means that boundary points have to be constrained to the original surface. As shown on the figure 2.3, if the boundary points are constrained with zero displacement, then the optimization potential of the mesh can be severely curtailed. In order to allow for effective optimization it is necessary to let the boundary points slide on the surface, while respecting essential parts of the geometry like feature edges and feature points. A feature edge in this context is defined as a boundary edge in which, neighbouring faces have normal vectors that differ by more than a threshold value and a feature point is a boundary point that is connected to more than two feature edges.

Supposing that for a boundary point on the mesh the displacement direction vector $\vec{r}$ has been calculated according to the sphericity sensitivity. We decompose this direction into two components, one normal to the surface and the other parallel to the tangent plane of the surface. By eliminating the vertical component, each point on the boundary is allowed to move only in the tangent direction $\vec{r}_t$. Since the surface is often curved the point then has to be projected back onto the surface of the initial geometry. Figure 2.4.1 illustrates an optimization algorithm that is applied in a low quality mesh with curved edges where the boundary nodes are allowed to slide on the surface.

2.4.2 Layer Cells

In order to resolve highly directional anisotropic gradients many applications employ grids which are likewise directionally biased. In CFD analysis this phenomenon often manifests



(a) Initial Mesh (b) Optimized Mesh

Figure 2: Mesh optimization allowing the surface nodes to slide on the surface

itself in the use of stratified layers of cells adjacent to solid boundaries, which are used to capture high wall-normal gradients in this region. Layer cells typically have high aspect ratios, with wall normal dimensions being much smaller that tangential. Figure 3(a) illustrates an example of layers cells close to the boundary wall. Due to their high aspect ratio layer cells have very small sphericity values. Allowing the optimization algorithm to freely deform layer cells will result in their extrusion, as illustrated in figure 3(b), which will significantly reduce wall normal resolution. It is therefore important for the optimization procedure to maintain anisotropic aspect of layer cells. To do this, the displacement of the layer cell points, towards the growing direction of the layer cells, is constrained by eliminating the derivative component in that direction. Since the method is scale invariant the implementation of such a constraint inside the optimization framework can introduce a shrinking cell problem: due to its inability to expand the layers in the normal direction, the algorithm will try to reduce all other dimensions. If left unchecked and if the mesh topology allows, this will result in cells with very small volumes. Figure 3(c) shows the result of a mesh optimization procedure with constrained layer cells. The shrunk cells create problems both in terms of volume deviation and in the skewness of the layer cells.

To overcome this problem, a supplementary objective, in the form of volume uniformity, is introduced inside the optimization framework, to improve the behaviour in cases with high aspect ratio layer cells. Supposing again that a point of the computational domain neighbours M cells, the volume deviation of that point is given by the formula

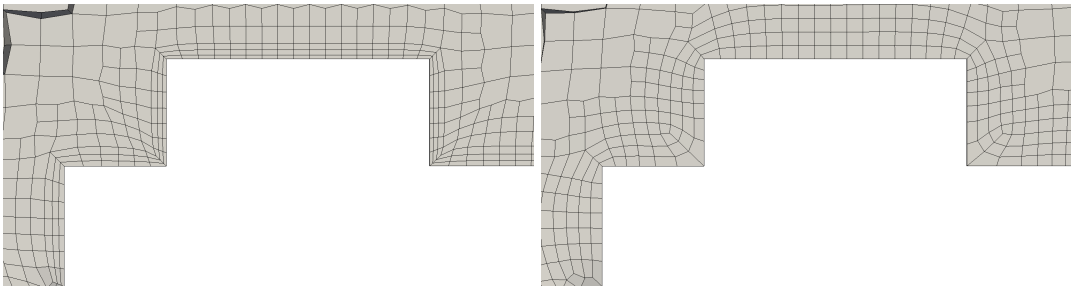
$$V_{dev}(P) = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{V_i(P) - \bar{V}}{\bar{V}} \right)^2 \tag{14}$$

where $\bar{V}$ is the mean volume of the surrounding cells. Its easy to see that the mean volume is independent of the position of the point P. Thus the differentiation of the volume deviation is straightforward,

$$\frac{dV_{dev}(P)}{dP} = \frac{2}{M} \sum_{i=1}^{M} \left[\frac{V_i(P) - \bar{V}}{\bar{V}} \cdot \frac{dV_i(P)}{dP} \right] \tag{15}$$

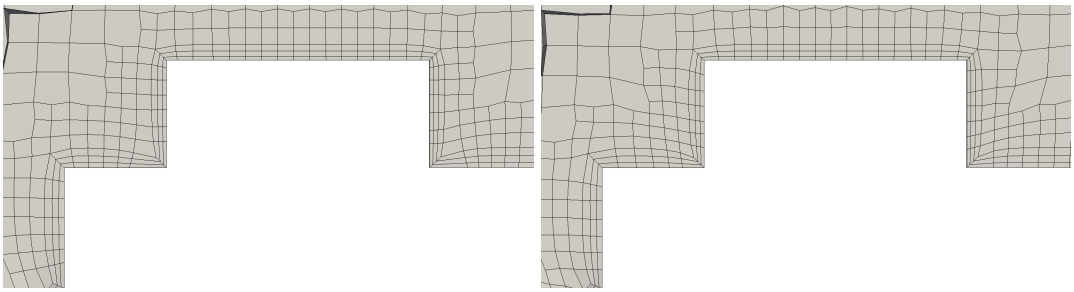
The new optimization function now becomes

$$F_{obj} = Sph \cdot (1 - V_{dev}) \tag{16}$$



(a) Initial mesh with layer cells

(b) Mesh optimization without constrains



(c) Mesh optimization with constrains

(d) Mesh optimization including volume deviation

Figure 3: Mesh optimization procedure starting from the initial mesh (a), firstly without constraints (b), afterwards with layer cells constraints (c) and finally combining constraints with volume deviation derivative (d).

with a derivative of

$$\frac{dF_{obj}}{dP} = \frac{dSph}{dP} \cdot (1 - V_{dev}) - Sph \cdot \frac{dV_{dev}}{dP} \tag{17}$$

Be noted that the addition of the new objective function does not raise the computational cost of the algorithm, since the calculation of the derivative of the volume deviation depends only on the cell volume derivative (which had been already calculated in Eq.10). Figure 3(d) illustrates the optimized computational mesh after the introduction of the volume deviation objective function. As it can be seen the additional derivative maintains the same volume on every stack of layer cells and prevents also the formation of high skewness cells.

2.5 CFD Mesh Optimization Applications

In this section the effectiveness of the method for improving the mesh quality in large scale test cases will be examined. The three quality metrics that will be used to compare the non optimized with the optimized mesh are non-orthogonality, skewness and cell determinant. Non-orthogonality of a face is the angle between the line that connects two cell centres and the normal of their common face. Skewness measures the distance between the intersection of the line connecting two cell centres with their common face and the centre of that face and finally with the determinant we are referring to the determinant of the cell's Jacobian matrix. Both for non-orthogonality and skewness their optimum value is zero, instead the optimum value for the determinant is unity.

Considering the previous example (figure 2.4.2) of the optimization algorithm, table 2.5 contains the maximum, minimum and mean values of those three quality metrics for the meshes that are displayed in figures 3(a), fig. 3(c) and 3(d). It can be noticed that the introduction of the volume deviation affects mostly the maximum/minimum values of non-orthogonality and skewness.

Quality Metric	Max. Value	Min. Value	Mean Value
Non-Orthogonality	56.18^o	0°	8.04^o
Skewness	2.63	0	0.19
Determinant	1	0.243	0.899

Table 1: Quality metrics for the initial mesh with layers (fig. 3(a)).

Quality Metric	Max. Value	Min. Value	Mean Value
Non-Orthogonality	56.32^o	0°	7.14^o
Skewness	3.31	0	0.12
Determinant	1	0.257	0.912

Table 2: Quality metrics for the optimized mesh with layers (fig. 3(c)) without introducing the volume deviation objective.

Quality Metric	Max. Value	Min. Value	Mean Value
Non-Orthogonality	50.16^o	0°	7.02^o
Skewness	3.01	0	0.11
Determinant	1	0.269	0.901

Table 3: Quality metrics for the optimized mesh with layers (fig. 3(d)) introducing also the volume deviation objective.

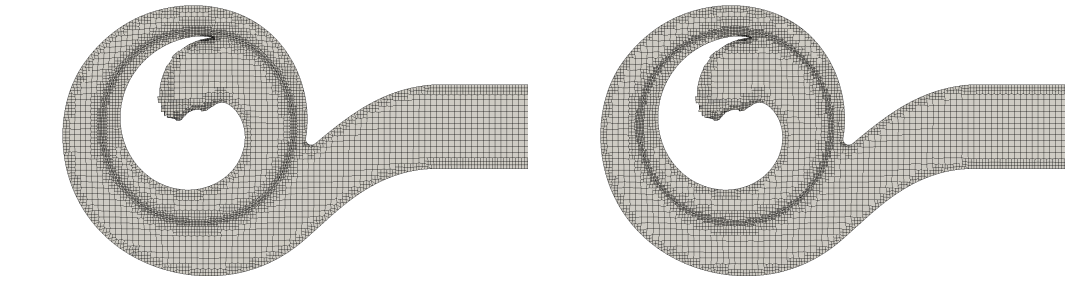
Quality Metrics	Max.Value		Min. Value		Mean Value	
	Before	After	Before	After	Before	After
Non-Orthogonality	69.27^o	68.59^o	0°	0°	7.57^o	7.22^o
Skewness	4.89	3.60	0	0	0.84	0.79
Determinant	1	1	0.621	0.657	0.979	0.984

Table 4: Quality metrics before and after the mesh optimization of the single channel pump.

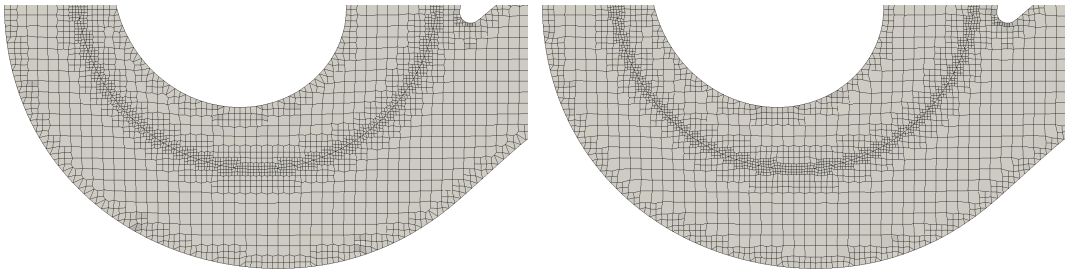
Figure 2.5 shows the mesh optimization results for a single channel pump. The mesh is unstructured, having around 0.5 million cell elements consisting primarily of tetrahedra, prisms and other types of polyhedra. Images 4(a) to 4(d) show the mesh geometry before and after the optimization, while figures 4(e) and 4(f) show the cell determinant. Tables 2.5 show the maximum, minimum and the mean values of the quality metrics before and after the optimization. It can be seen that the maximization of the sphericity leads also to the maximization of the three quantities inside the domain. It should be noted that the algorithm has been calibrated to focus more on improvement of cells near the boundary, where the mesh is finer and the accuracy requirements are more demanding, than those further toward the interior of the domain.

The next case is the Ahmed-Body car (figure 2.5). Table 2.5 contain the quality metric quantities before and after the optimization of the computational mesh (1 million elements). Figures 5(a) and 5(b) show the skewness in the car boundary faces before and after the optimization.

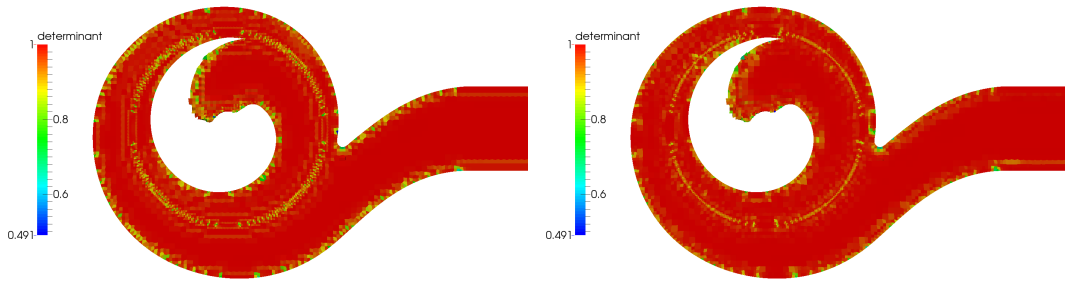
It is important to examine how the improvement of the computational mesh affects the CFD calculations in terms of convergence speed. We solve the flow around the Ahmed-Body using the incompressible OpenFOAM solver (simpleFoam), both for the optimized and the non-



(a) Initial Mesh (b) Final Mesh



(c) Initial Mesh Detailed (d) Final Mesh Detailed

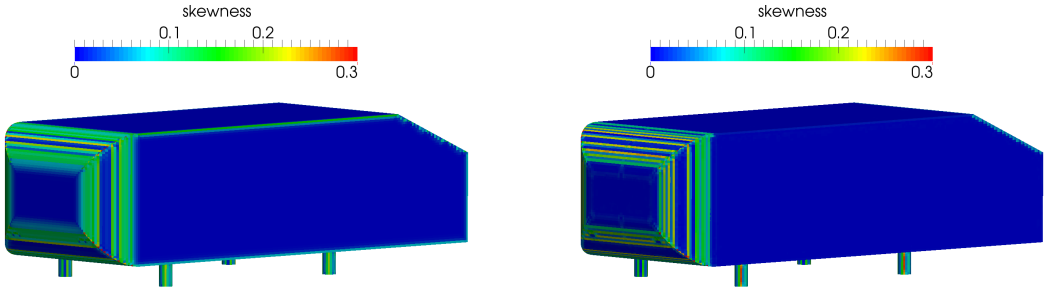


(e) Initial Mesh Cell Determinant (f) Initial Mesh Cell Determinant

Figure 4: Mesh optimization results on a pump with half million cell elements. Figures (a) and (c) illustrate the geometry before optimization while (b) and (d) illustrate the geometry after the optimization. Finally figures (e) and (f) display the cell determinant before and after the optimization.

Quality Metrics	Max.Value		Min. Value		Mean Value	
	Before	After	Before	After	Before	After
Non-Orthogonality	48.68^o	47.23^o	0°	0°	3.93^o	3.88^o
Skewness	3.16	2.23	0	0	0.51	0.48
Determinant	1	1	0.617	0.749	0.990	0.991

Table 5: Quality metrics before and after the mesh optimization of the single Ahmed-Body.



(a) Initial Boundary Face Skewness (b) Final Boundary Face Skewness

Figure 5: Face skewness on the boundary of the S-Bend duct before and after the sphericity optimization.

optimized mesh. In the optimized mesh the solver needs 6.5% less iterations to achieve convergence which reflects a 7.3% reduction on CPU time.

3 MORPHING FRAMEWORK INTEGRATION

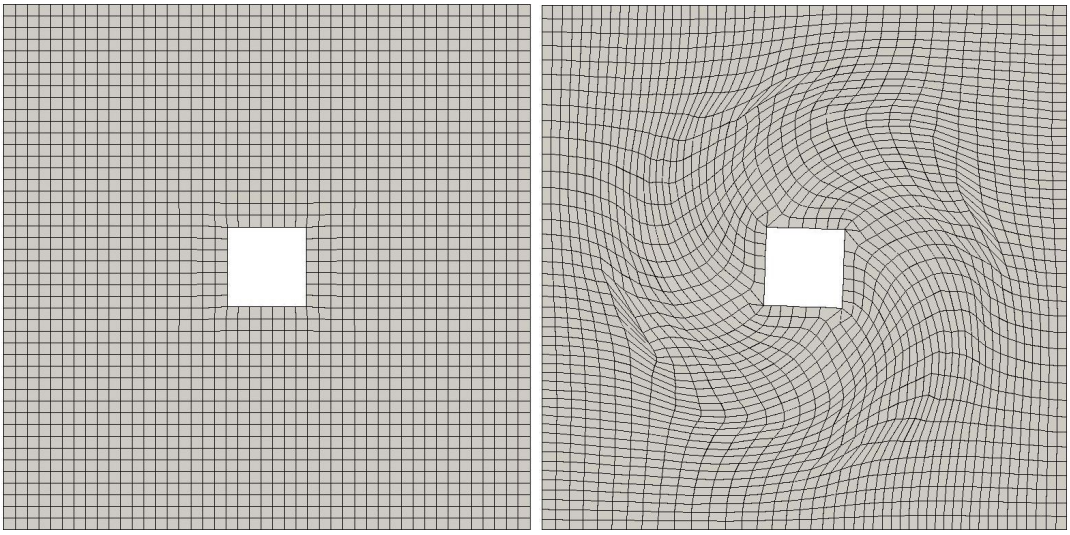
It has been demonstrated that the optimization algorithm can be used for improving the quality metrics on the computational mesh, leading to more accurate and reliable results. However the method can also be employed during dynamic mesh motion to prevent the occurrence of badly deformed elements, resulting in greater robustness and efficacy even during extreme deformation. In the context CFD analysis, a mesh displacement algorithm is indispensable in cases with moving boundaries, like shape optimization, aero-elasticity etc.

Many mesh displacement algorithms have been developed so far following a variety of approaches, like elastic medium analogy [8], [9] spring analogy [10] and Radial Basis Functions (RBF) [11] methods. In this paper the mesh displacement algorithm is based on the existence of an objective function (Sphericity) which quantifies the mesh quality. Solving the optimization problem for the maximization of the sphericity value, results in the new positions of the points inside the mesh.

Figure 3 illustrates the method's ability to maintain high cell quality in extreme deformation conditions such as a 90 degree box rotation. The rotation was made in 45 steps of 2 degrees running 50 optimization cycles for each rotation step. Further rotation of the box (over 93 degrees), will lead to an invalid mesh, primarily due to the fact that the corner vertices are not allowed to slip. Table 3 shows the quality metrics before and after the box rotation. The final mesh even after the extreme deformation maintains very low skewness values. Contrariwise the non-orthogonality has significantly increase, having four faces with non-orthogonality value over 70 degrees (a threshold value over which can be consider that non-orthogonal faces are starting to have impact in the CFD solution).

Quality Metrics	Max.Value		Min. Value		Mean Value	
	Before	After	Before	After	Before	After
Non-Orthogonality	5.77^o	77.18^o	0°	0.01^o	0.23^o	16.54^o
Skewness	0.07	2.11	0	0	~ 0	0.16
Determinant	1	1	0.985	0.041	0.999	0.861

Table 6: Quality metrics before and after the rotation of the box by 90 degrees.



(a) Initial box position (b) 90° degrees rotating box

Figure 6: Mesh displacement around a box using sphericity optimization applying a 90° degrees rotation.

3.1 Adjoint Shape Optimization

In the context of gradient based optimization, the adjoint method can be considered the most computational effective as the gradient calculation of the objective function is independent of the number of the design variables. That allows the creation of a very rich design space which leads to better optimum solutions. The adjoint method was introduced by Pironneau [5] and consequently Jameson [6] and Giles [7] et al. extended the method inside the scope of shape optimization. An in depth analysis of the adjoint method goes beyond the scope of this paper, thus only a brief description of the continuous adjoint formulation will be presented. For a more circumstantial analysis one can refer to [1], [2].

3.1.1 Continuous Adjoint Method

Further on, Einstein's summation convention applies for the lower-case indices, unless declared differently. Assuming a laminar flow of an incompressible fluid, the Navier-Stokes equations can be written as,

$$R^p = -\frac{\partial v_j}{\partial x_j} = 0 \tag{18a}$$

$$R_i^v = v_j \frac{\partial v_i}{\partial x_j} - \frac{\partial}{\partial x_j}\left[\nu\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] + \frac{\partial p}{\partial x_i} = 0 \qquad i = 1,2,3 \tag{18b}$$

where v_i is the velocity in direction of the Cartesian coordinates, p is the static pressure divided by the fluid's density ρ and ν is the kinematic viscosity. Afterwards a general objective function F is defined containing both surface S and volume Ω integrals,

$$F = \int_S F_s dS + \int_\Omega F_\Omega d\Omega \tag{19}$$

then F is augmented by the state equations leading to

$$F_{aug} = F + \int_\Omega u_i R_i^v d\Omega + \int_\Omega q R^p d\Omega \tag{20}$$

note that since the Navier-Stokes equations are satisfied it holds that $F_{aug} = F$.

The differentiating of the augmented objective function F_{aug} arises the adjoint equations for incompressible laminar flow.

$$R^q = \frac{\partial u_i}{\partial x_i} - \frac{\partial F}{\partial p} = 0 \tag{21a}$$

$$R_i^u = -v_j \frac{\partial u_i}{\partial x_j} + u_j \frac{\partial v_j}{\partial x_i} - \nu \frac{\partial}{\partial x_j}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) + \frac{\partial q}{\partial x_i} + \frac{\partial F}{\partial v_i} = 0 \tag{21b}$$

After solving the adjoint equations (which are similar to the Primal ones 18b,18a), the surface sensitivities of the objective function w.r.t. the surface normal of each surface node can be calculated as

$$G = \frac{\delta F_{aug}}{\delta b} = -\int_{S_w} \left[\nu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) - q n_i\right] \frac{\partial v_j}{\partial x_k} \frac{\partial x_k}{\partial b} dS \tag{22}$$

3.1.2 Adjoint Sensitivities Smoothing

The design variables in this shape optimization problem are the surface normal displacements of each boundary node. Any oscillation (noise) in the sensitivity derivatives can lead to impractical, non-manufacturable, shapes which makes it necessary to apply filters that eliminate these oscillations. Based on [12] the filtered distribution of the sensitivity derivatives $\bar{G}$ is given by the convolution

$$\bar{G}(\xi) = \int_{S_w} K(r) G(\xi - r) dr, \qquad \forall \xi \epsilon S_w, \tag{23}$$

where r is the filter radius and ξ is the curved boundary coordinate. The filtered sensitivities can be calculated solving the equation

$$\bar{G} - \frac{\partial}{\partial \xi}\left(\varepsilon \frac{\partial \bar{G}}{\partial \xi}\right) = G \tag{24}$$

where ε is the smoothing intensity.

4 APPLICATIONS

4.1 Optimization of the S-Bend Duct

The final application, is the adjoint-based shape optimization of the S-Bend duct. The aim is to modify the shape of the duct so as to reduce the pressure losses while maintaining the geometry of the inlet and the outlet boundary. The flow is laminar with a fully developed profile

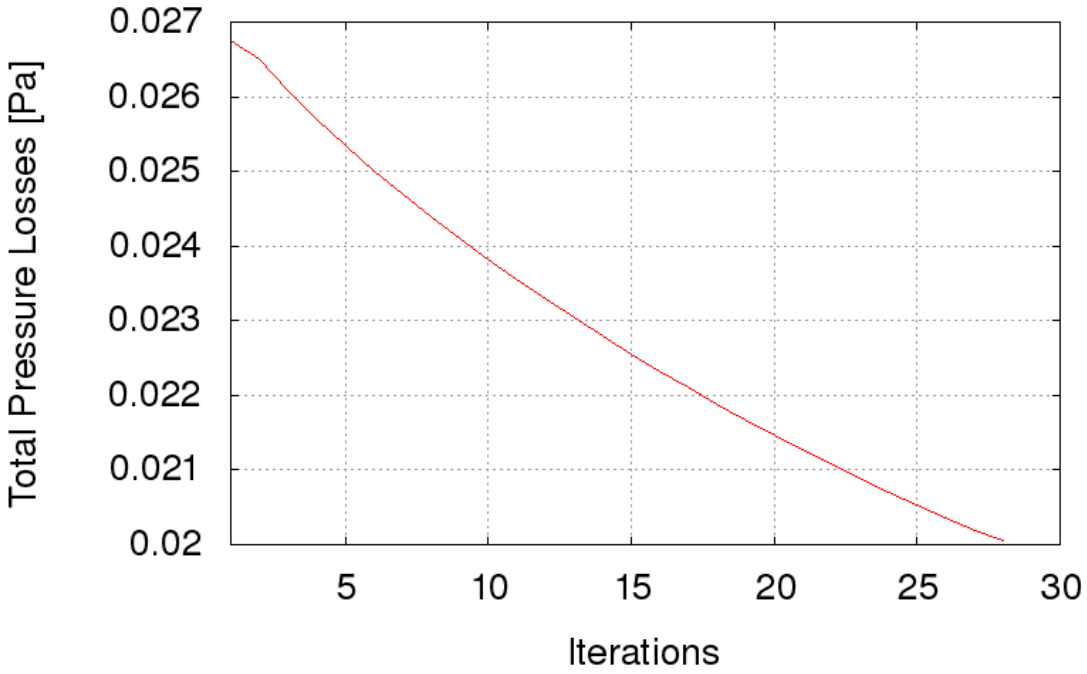
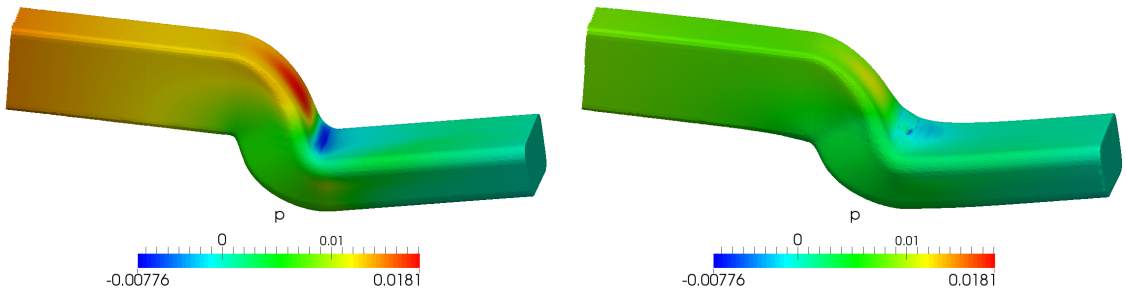


Figure 7: Total pressure losses for each shape optimization step.

at the inlet of the duct. Figure 4.1 illustrates the total pressure losses for every optimization step. After 28 optimization steps, the pressure losses were reduced by 25.01%. Figure 4.1 illustrates the geometry and the pressure distribution before and after the shape optimization of the duct, while figure 4.1 illustrates the difference between initial and smoothed sensitivities.

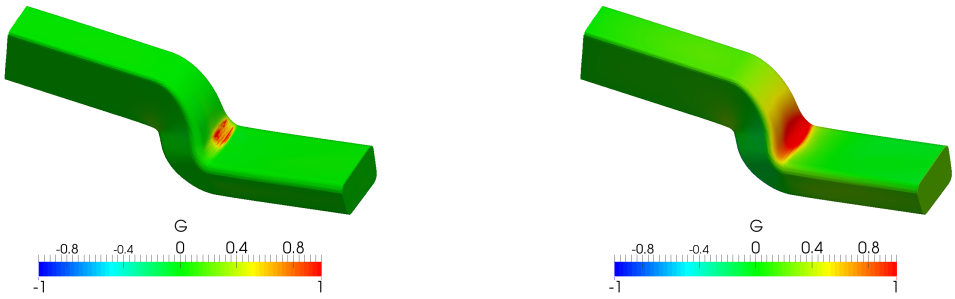
Table 4.1 shows the quality metrics for the initial and the final mesh. A high increase can be noticed at skewness metric having 3 faces over the acceptable skew value (≥ 4). Those faces were also incorrectly oriented preventing us to proceed further in to the optimization procedure.

Using the mesh optimization algorithm as part of a mesh displacement procedure, maintains the quality of the mesh throughout multiple deformation steps. This reduces or removes the need for re-meshing after each optimization cycle, thus speeding up and improving the automation



(a) Initial shape and pressure distribution (b) Optimized shape and pressure distribution

Figure 8: Shape geometry and pressure distribution before and after the adjoint shape optimization.



(a) Initial Sensitivity Map (b) Smoothed Sensitivity Map

Figure 9: Initial and smoothed sensitivity map calculated with the adjoint method.

Quality Metrics	Max.Value		Min. Value		Mean Value	
	Before	After	Before	After	Before	After
Non-Orthogonality	60.09^o	73.59^o	0°	0.01^o	2.47^o	5.26^o
Skewness	2.25	11.23	0	0	0.09	0.13
Determinant	1	1	0.118	0.014	0.999	0.990

Table 7: Quality metrics before and after the shape optimization of the S-Bend duct.

of the procedure.

5 CONCLUSIONS

An algorithm for optimizing finite volume meshes with arbitrary element topology was developed. The algorithm is based on a new heuristic quality metric termed sphericity. The maximization of sphericity inside the computational domain consistently produces improvement of all other quality metrics (skewness, non-orthogonality etc.) during testing, leading to more reliable results and solution times. The algorithm combines hand a derived gradient sphericity with a LM-BFGS method to speed up the solution. Heuristic techniques for limiting cell shrinkage were applied to improve effectiveness on meshes with mixed element topology and highly anisotropic cells. Basic constraints are employed to allow flexibility in the context of boundaries and feature elements. The algorithm was successfully employed in the context of adjoint based shape optimization of an S-bend Duct case.

ACKNOWLEDGEMENTS

The first author is an Early State Researcher participating in ITN *Industrial optimal design using adjoint CFD (IODA)* MSCA-ITN-2014-ETN: Marie Sklodowska-Curie Innovative Training Networks, Grant Agreement No 642959.

REFERENCES

[1] Zymaris, A., Papadimitriou, D., Giannakoglou, K., Othmer, C. Continuous adjoint approach to the Spalart- Allmaras turbulence model for incompressible flows. *Computers & Fluids,* 38(8):15281538, 2009.

[2] Zymaris, A. S., Papadimitriou, D. I., Giannakoglou, K. C. and Othmer C. Adjoint wall

functions: A new concept for use in aerodynamic shape optimization. *J. Comp. Phys.,* 229(13), 52285245., 2010.

[3] J. Nocedal Updating Quasi-Newton Matrices with Limited Storage *Mathematics of Computation,* Vol. 35, No. 151. (1980), pp. 773-782.

[4] P.Wolfe Convergence Conditions for Ascent Methods *SIAM Review 11* (2): 226000

[5] O. Pironneau On optimum design in fluid mechanics. *J. Fluid Mech.*, 64(1):97110, 1974

[6] Jameson, A. Optimum aerodynamic design using CFD and control theory. *AIAA-1995-1729-CP,* 1995.

[7] Giles, M. B., Pierce, N. A. An introduction to the adjoint approach to design. *Flow, Turbul. Combust.*, 65:393415, 2001

[8] T. Baker and P. A. Cavallo. Dynamic adaptation for deforming tetrahedral meshes. *AIAA Paper 1999-3253,* 1999.

[9] Amirante, D. and Hills, N.J. and Barnes, C.J., A moving mesh algorithm for aero-thermo-mechanical modelling in turbomachinery, *International Journal for Numerical Methods in Fluids*, vol. 70, 1118–1138, 2012

[10] J. Batina Unsteady Euler algorithms with unstructured dynamic mesh for complex-aircraft aerodynamic analysis. *AIAA Journal,* 29(3):327-333, 200.

[11] D. Sieger, S. Menzel and M. Botsch. RBF Morphing Techniques for Simulation-based Design Optimization. *Engineering with Computers,* 30(2):161174, 2014

[12] Stuck, A. and Rung, T., Filtered Gradients for Adjoint-based Shape Optimisation *20th AIAA Computational Fluid Dynamics Conference* 27 - 30 June 2011, Honolulu, Hawaii

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

LINKING PARAMETRIC CAD WITH ADJOINT SURFACE SENSITIVITIES

Ilias Vasilopoulos[1,3], Dheeraj Agarwal[2], Marcus Meyer[1], Trevor T. Robinson[2] and Cecil G. Armstrong[2]

[1]Rolls-Royce Deutschland (RRD)
Eschenweg 11, Dahlewitz, D-15827 Blankenfelde-Mahlow, Germany
{ilias.vasilopoulos, marcus.meyer}@rolls-royce.com

[2]School of Mechanical and Aerospace Engineering, Queen's University Belfast (QUB)
Ashby Building, Belfast BT9 5AH, Northern Ireland, UK
{d.agarwal, t.robinson, c.armstrong}@qub.ac.uk

[3]School of Mechanical Engineering, National Technical University of Athens (NTUA)
Iroon Polytecniou 9, 15780 Zografou, Athens, Greece
vasilopoulos@central.ntua.gr

Keywords: CAD, CFD, Adjoint, Design Velocities, Sensitivity Map, Parametric Sensitivities.

Abstract. *The goal of this work is to present an efficient CAD-based adjoint process chain for calculating parametric sensitivities (derivatives of the objective function with respect to the CAD parameters) in timescales acceptable for industrial design processes. The idea is based on linking parametric design velocities (geometric sensitivities computed from the CAD model) with adjoint surface sensitivities. A CAD-based design velocity computation method has been implemented based on distances between discrete representations of perturbed geometries. This approach differs from other methods due to the fact that it works with existing commercial CAD packages (unlike most analytical approaches) and it can cope with the changes in CAD model topology and face labeling. Use of the proposed method allows computation of parametric sensitivities using adjoint data at a computational cost which scales with the number of objective functions being considered, while it is essentially independent of the number of design variables. The gradient computation is demonstrated on test cases for a Nozzle Guide Vane (NGV) model and a Turbine Rotor Blade model. The results are validated against finite difference values and good agreement is shown. This gradient information can be passed to an optimization algorithm, which will use it to update the CAD model parameters.*

1 INTRODUCTION

Throughout industry, engineers use Computer Aided Design (CAD) software packages for design and Computer Aided Engineering (CAE) tools for analysis. The CAD provides the environment for the geometry generation of a model and its parametrisation, whereas the CAE, such as Finite Element Analysis (FEA) and Computational Fluid Dynamics (CFD) solvers, performs an analysis on a computational mesh which approximates the geometry.

The engineer is interested in obtaining the optimum geometry (in terms of some objective function of interest), while satisfying predetermined (e.g. manufacturing) constraints. The use of efficient gradient-based optimization algorithms requires the derivatives of the objective function w.r.t. the design variables. In case of CAD-based optimization, the CAD parameters which define the geometric features can be considered as design variables. The derivative of the model's performance w.r.t. each design parameter can be obtained in a finite difference sense, where the effect of a parameter change is computed by analysing both (baseline and updated) geometries and comparing the results. For each parameter, a perturbed geometry is created in CAD and then fed to CAE for analysis (including geometry healing and mesh generation processes), where the resulting difference in performance enables the derivative calculation. This finite difference approach is computationally expensive, since the cost scales with the number of design parameters, and can thus be prohibitive for industrial applications.

As an alternative and much more efficient way of computing gradient information, adjoint methods have been the subject of considerable research in recent years [1, 2, 3, 4, 5]. The advantage of an adjoint method arises from the fact that the gradient computation becomes independent of the number of design variables. This is achieved by solving a linear problem of computational cost comparable to that of the primal analysis. As a result, most adjoint-based optimization strategies use the model's surface mesh node coordinates as design variables, which represents the richest design space the CFD can evaluate. On the other hand, since all grid surface nodes can move independently, the implementation of a smoothing algorithm is required to prevent the appearance of oscillatory shapes during the optimization process. Moreover, such CAD-free methods produce as their output an optimized mesh, which then has to be converted back to CAD in order to obtain the optimized geometry for further analysis or manufacturing. This mesh-to-CAD step is a non-trivial task and it may require extensive user interaction. For these reasons, CAD-based optimization is considered more desirable and thus a link between adjoint and CAD is of utmost importance.

Very limited work on CAD-based optimization is currently available [6, 7]. Three promising approaches to include the CAD system into the design loop are explored within the EC ITN IODA project: a) NURBS-based parametrisation, b) Parametric CAD sensitivities through finite-differencing and c) CAD derivatives through automatic differentiation (AD).

Non-uniform rational B-splines (NURBS) patches are a generic surface description and are used by all CAD packages (STEP format). Queen Mary University of London (QMUL) has developed a NURBS-based approach [8], where the NURBS control points are considered as design variables and the required derivatives are obtained by AD of the NURBS analytical description. The approach is not limited to a single patch, but allows to enforce continuity constraints across patch interfaces. Recent work has extended this CAD-based parametrisation method to include geometric constraints (such as thickness and trailing edge radius) [9].

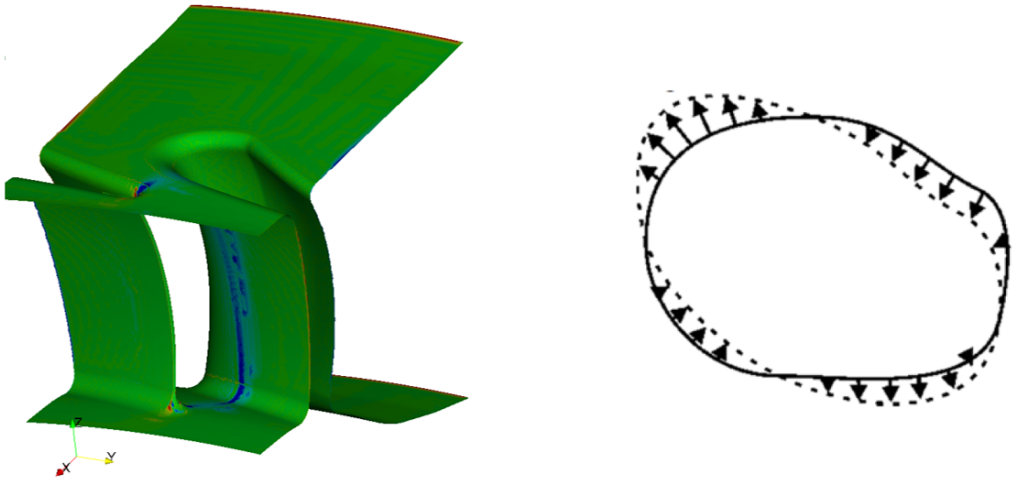
Geometric constraints are best incorporated by a restriction of the design space, which is straightforwardly achieved through the parametrisation of the CAD model. However, if the native CAD model parameters are to be used as design variables, then the derivatives of the

model's surface w.r.t. these parameters are required, since they are not provided by today's commercial CAD packages. One approach to obtain these geometric sensitivities (also found in the literature as design velocities [10]) is to compute finite differences between the CAD model before and after a parameter perturbation. This approach has been identified by some researchers [11, 12, 13] to lack robustness against topology and labelling changes, which can occur even under small parameter perturbations. Nevertheless, Queen's University of Belfast (QUB) has developed an approach based on projections between discrete representations of the CAD geometries [14], which is efficient, robust and has been successfully applied to CFD [15].

To compute the exact geometric derivatives, one can differentiate the CAD kernel, either analytically or via AD tools [16]. Differentiating the CAD kernel is possible only when you have access to the CAD software source code, as in OpenCASCADE [27]. Although this approach is promising, the automatic differentiation of a complete design chain incorporating a CAD tool still remains a challenging task to be tackled.

In this work, the link to CAD is achieved by applying an improved QUB design velocity approach, in which the CAD modeling system is treated as a black-box. The approach uses the generic STEP surface description and thus it can be easily coupled to any feature-based CAD tool. It is immune to topology changes caused by parametric variations and it requires no access to the CAD source code. This makes it suitable for implementation within an industrial context, where commercial CAD packages (such as Siemens NX or CATIA V5) are widely used.

The idea is based on linking two concepts: adjoint surface sensitivities and design velocities. Adjoint surface sensitivities (which form a sensitivity map) can be obtained from an adjoint CFD solution and represent the derivative of the objective function w.r.t. normal perturbations of the boundary mesh at each node. Fig. 1(a) shows the sensitivity map provided by Rolls-Royce Deutschland (RRD) for a Nozzle Guide Vane (NGV) case. Areas of high sensitivity are shown in red or blue, representing areas where the body should be displaced outwards or inwards (respectively) to achieve an increase in objective function, whilst areas of low sensitivity (plotted in green) show that the boundary in this region has no large influence on the objective function.



a) b)

Figure 1: a) Adjoint sensitivity map for NGV test case and b) Simple example of 2D design velocity field.

On the other hand, design velocities are geometric sensitivities computed from the CAD model, representing the displacement of a point on the model boundary due to the perturbation of a CAD model parameter. This work makes use of the normal component of the design velocity, V_n, which is computed as

$$V_n = dp \cdot \hat{n} \tag{1}$$

where dp is the displacement vector of a point on the boundary due to a parameter perturbation and $\hat{n}$ is the outward unit normal of the surface at that point. This measure was initially developed in the context of adjoint-based structural optimization problems [17]. In Fig. 1(b), which is taken from [13], the arrows represent the design velocities as the boundary changes from the solid to the dashed line. Due to the convention adopted throughout this paper that the boundary normals are pointing outwards, a positive design velocity represents an outward movement of the boundary, and negative is inward.

The CAD-based adjoint process chain for calculating parametric sensitivities (derivatives of the objective function w.r.t. the CAD parameters), followed in this work, is outlined in Fig. 2.

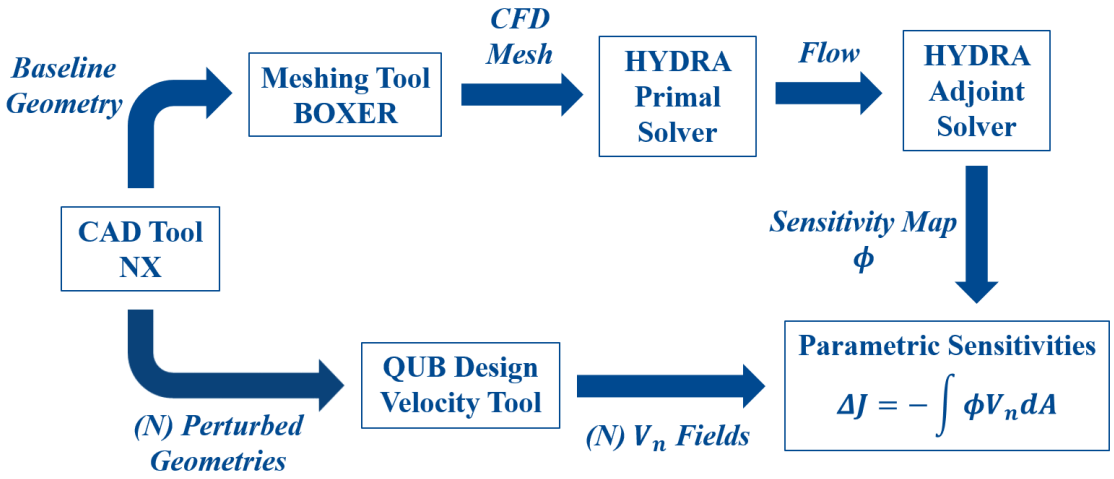


Figure 2: Workflow for parametric sensitivities computation.

Use of the above approach allows the computation of parametric sensitivities at a cost of one adjoint analysis per objective function and one design velocity field evaluation per parameter. Since design velocities can be computed for all of the parameters in a much shorter time than that needed to converge the primal and adjoint solutions, gradient computation is achieved in timescales acceptable for industrial design processes. As shown in the last step of Fig. 2, parametric sensitivities are calculated by applying the chain rule between the sensitivity map ϕ and the design velocities V_n.

The remainder of this paper is structured as follows: Sec. 2 explains the mathematical background of the discrete adjoint solver and how the sensitivity map is derived. Sec. 3 describes the method used for the computation of design velocities. The gradient results obtained for two turbomachinery test cases are presented in Sec. 4, followed by the conclusions in Sec. 5.

2 DISCRETE ADJOINT METHOD

2.1 Mathematical background

In this work, a RANS compressible flow solver and its corresponding discrete adjoint solver are used for computing the flow and sensitivity map. These are part of the Rolls-Royce in-house HYDRA suite of codes, which have been extensively validated and applied to various industrial cases (see for example [18, 19]). More details regarding the underlying theory and implemented algorithm can be found in [3, 20, 21]. The nonlinear flow solver uses a node-based finite-volume discretisation method and the pseudo-time-marching to steady state is accelerated by a block-Jacobi preconditioner and a geometric multigrid technique.

During the convergence of the primal solution, the nonlinear residual for each control volume is driven to zero

$$R(U, X) = 0 \,, \tag{2}$$

where U is the primal solution and X stands for the mesh coordinates. The objective function considered is a function of both the flow solution and the mesh, while the flow solution implicitly depends on the mesh. In addition, the mesh coordinates are defined by the CAD design variables α, which leads to the relation

$$J = J(U(X(\alpha)), X(\alpha)) \,. \tag{3}$$

The parametric sensitivities can be computed using the chain rule as

$$\frac{dJ}{d\alpha} = \frac{dJ}{dX}\frac{dX}{dX_S}\frac{dX_S}{d\alpha} \,, \tag{4}$$

where X_S stands for the surface mesh coordinates. The adjoint solver is responsible for the computation of the volume mesh sensitivities dJ/dX, which is achieved by solving the equations shown below.

The objective function is augmented as

$$J = J + \lambda^T R \,, \tag{5}$$

where λ is the adjoint solution. Differentiating Eq. (5) w.r.t. X leads to

$$\begin{aligned} \frac{dJ}{dX} &= \frac{dJ}{dX} + \lambda^T \frac{dR}{dX} \\ &= \frac{\partial J}{\partial X} + \frac{\partial J}{\partial U}\frac{dU}{dX} + \lambda^T \Big(\frac{\partial R}{\partial X} + \frac{\partial R}{\partial U}\frac{dU}{dX} \Big) \\ &= \frac{\partial J}{\partial X} + \lambda^T \frac{\partial R}{\partial X} + \Big(\frac{\partial J}{\partial U} + \lambda^T \frac{\partial R}{\partial U} \Big) \frac{dU}{dX} \,. \end{aligned} \tag{6}$$

The idea is to eliminate the computationally expensive term dU/dX. To do so, its multiplier in Eq. (6) is set to zero, giving rise to the adjoint equation

$$\frac{\partial R}{\partial U}^T \lambda = -\frac{\partial J}{\partial U}^T \,. \tag{7}$$

Obtaining the adjoint solution λ, by solving the linear Eq. (7), allows for the efficient computation of dJ/dX, given by the remaining terms of Eq. (6) as

$$\frac{dJ}{dX} = \frac{\partial J}{\partial X} + \lambda^T \frac{\partial R}{\partial X} \,. \tag{8}$$

As shown in Eq. (4), the computation of parametric sensitivities also requires the terms dX/dX_S and $dX_S/d\alpha$.

2.2 Sensitivity map

The term dX/dX_S represents the inverse operation of a mesh moving algorithm. It allows the computation of the surface mesh sensitivities dJ/dX_S as

$$\frac{dJ}{dX_S} = \frac{dJ}{dX}\frac{dX}{dX_S} \, . \tag{9}$$

For this projection of sensitivities from the volume to the surface, a simple spring-based mesh deformation algorithm is used. Eq. (9) results in

$$\frac{dJ}{dX_S}^T = \mathbf{K}_{s2v}^T \mathbf{K}^{-T} \frac{dJ}{dX}^T \, , \tag{10}$$

where $\mathbf{K}$ is the stiffness matrix and $\mathbf{K}_{s2v}^T$ is a permutation matrix mapping each volume node to its corresponding surface node if any, and to null if none. The only computationally expensive part is to approximately invert $\mathbf{K}^T$. A Jacobi preconditioned conjugate gradient solver is used in this work, resulting in negligible CPU cost compared to the primal and adjoint solution.

After this step, Eq. (4) is rewritten as

$$\frac{dJ}{d\alpha} = \frac{dJ}{dX_S}\frac{dX_S}{d\alpha} \, . \tag{11}$$

The adjoint sensitivity map is derived by projecting each surface node's sensitivity vector dJ/dX_S onto its corresponding boundary normal. The same is done for the missing term $dX_S/d\alpha$, which represents the design velocities (explained in Sec. 3), and the parametric sensitivities $dJ/d\alpha$ are finally obtained from the inner product of those two components.

3 DESIGN VELOCITY APPROACH

The approach presented in this paper for computing parametric design velocities is based on a finite difference between the CAD model's shape before and after changing the value of a parameter defining the geometry. It is a development of the methodology presented in [14] and is characterized by an increased robustness. It works with existing commercial CAD packages and it can cope with the changes in CAD model topology and face labeling, as discussed in the introduction.

The CAD geometry is represented using a surface mesh of triangular elements. In this work, the mesh representations (for the unperturbed model and for each perturbed model) are created using either Gmsh [28] or CADfix [29] grid generator. The displacement of the model boundary due to a perturbation is approximated by projecting the centroid of each of the elements in the unperturbed model onto the elements in the perturbed model, as shown in Fig. 3. The projection location Pp is defined by the intersection of the surface normal direction $\hat{n}$ at the centroid C of the original model element with the plane the perturbed element exists on. When the correct projection is identified, the displacement vector dp is calculated as the difference between the centroid and its projection. The design velocity is then given by Eq. (1).

The question that arises is which perturbed element should each element centroid in the unperturbed model be projected onto. In order to increase the efficiency of this search, a multidimensional binary search tree known as KD-tree [22] is employed to find the perturbed element with the closest centroid, which is tested first. If the result is an unsuccessful projection, Barycentric coordinates [23] are used to determine which perturbed element should

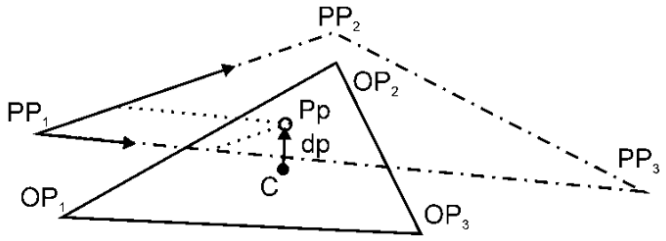


Figure 3: The centroid C of the original model element $OP_1OP_2OP_3$ is projected onto the perturbed element $PP_1PP_2PP_3$.

be tested next. The process continues until a successful projection onto an element face, edge or vertex is found, allowing the displacement to be calculated.

One additional check to ensure that the appropriate projection is identified is to demand the unperturbed boundary normal at the centroid that is projected to be similar to the perturbed boundary normal at the identified projection point. This is important in thin regions where otherwise the projection could occur onto the wrong face of the model. In this work, the surface normals are considered to be similar when

$$\hat{n}_{unperturbed} \cdot \hat{n}_{perturbed} > \frac{\sqrt{3}}{2} \, . \tag{12}$$

The design velocity computation process is outlined in Fig. 4 and it is fully automated using a Python script.

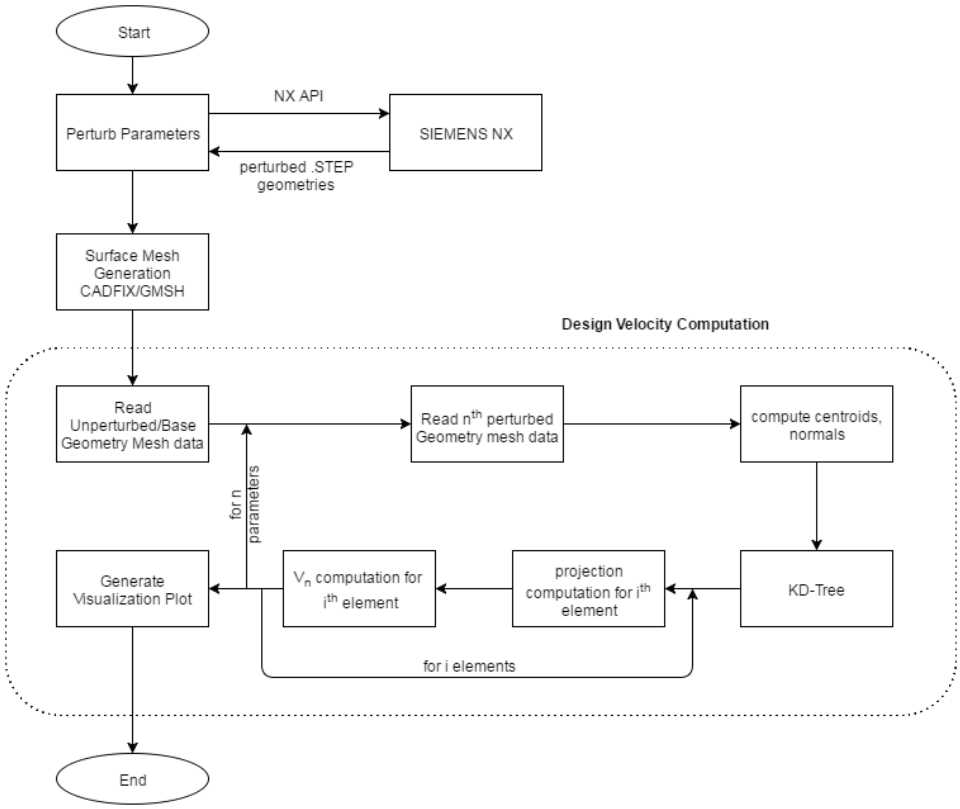


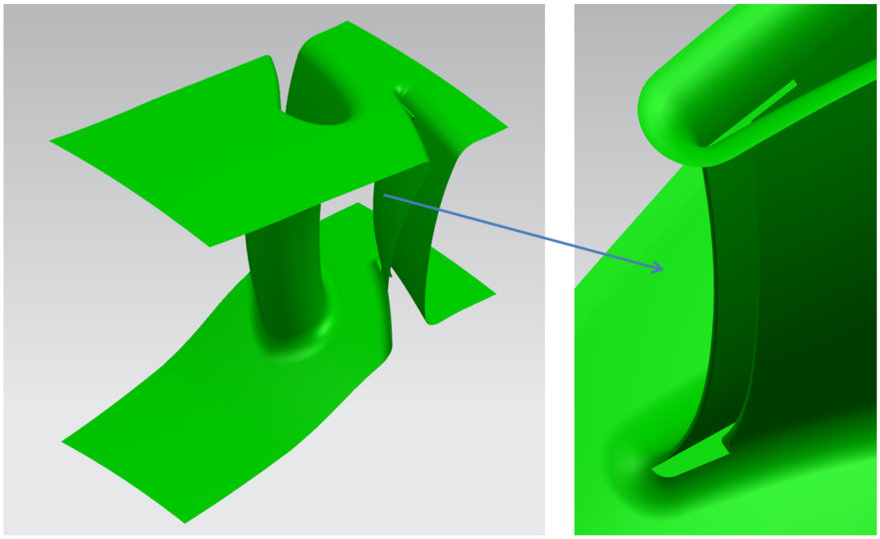
Figure 4: Design velocity computation process.

4 APPLICATIONS

4.1 Nozzle Guide Vane

In order to test the proposed approach, the first case used in this work is a state-of-the-art nozzle guide vane (NGV) of a high pressure turbine (HPT) provided by Rolls-Royce. The NGV geometry is shown in Fig. 5, where a 3D parametric CAD model is built in Siemens NX. The NGV has fillets at both ends and a cooling slot feature at the trailing edge (TE). This geometry is also investigated in [24], which illustrates different methods in order to quantify the impact of geometric variations on the NGV performance with focus on the capacity. Typically, the engine mass flow (and by association the capacity) is governed by the NGV design, which defines the narrowest cross section of the turbine. As a result, in this test case the capacity is considered as the objective function.

The CFD simulation requires only one periodic section of the engine's annulus. Consequently, the sector domain (Fig. 5) was used for both CFD and design velocities computation.



(a) (b)

Figure 5: 3D parametric CAD model in Siemens NX: a) NGV geometry and b) TE cooling slot.

The CAD model is parametrised using NX expressions and perturbed geometries are created using an iSIGHT workflow. Fig. 6 shows some of the 12 CAD parameters considered as exemplary design variables in this test case, together with the corresponding geometric variation that they cause.

For each perturbation of these parameters, a design velocity field was obtained using the QUB design velocity tool. In this case, the QUB tool was coupled with the open source grid generator Gmsh, which provided the mesh representations used for the projections. Design velocity fields for some of the parameters are shown in Fig. 7.

In order to be able to validate the results produced by linking these design velocity fields with the adjoint sensitivity map, a finite difference (FD) study has been conducted. Thus, 24 perturbed geometries (for + and - perturbations of the parameters) were generated using the iSIGHT workflow and an automated meshing procedure was executed using the BOXER

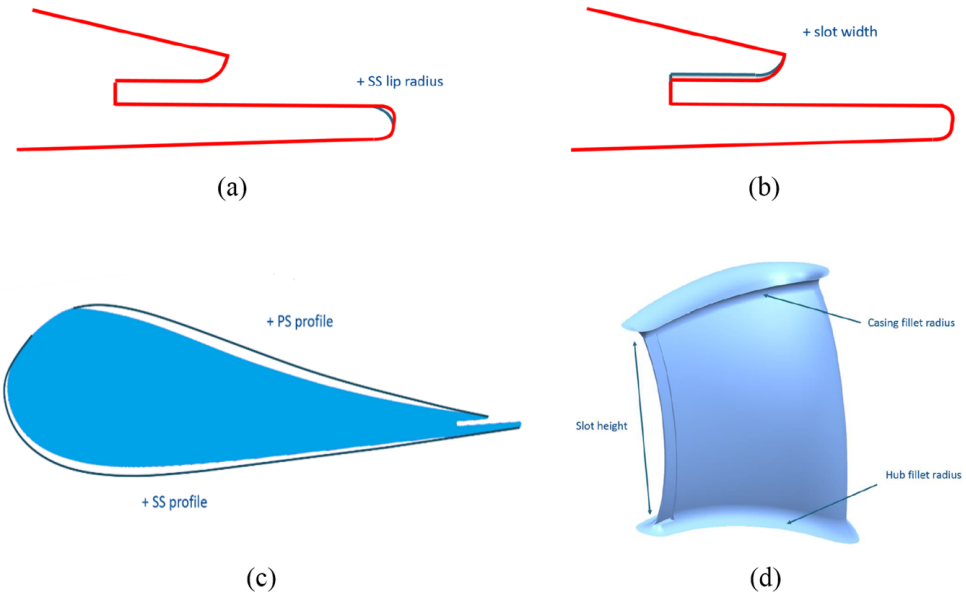


Figure 6: CAD feature parameters considered as design variables (not to scale): a) suction side lip radius, b) slot width, c) pressure side and suction side profile and d) slot height and fillet radius at hub and casing.

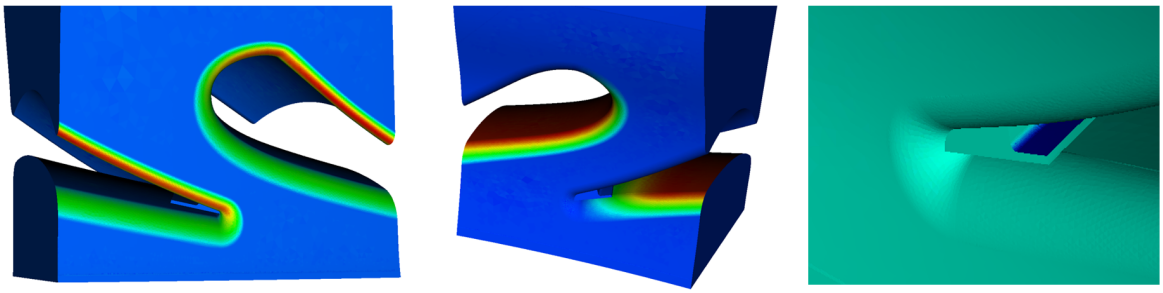


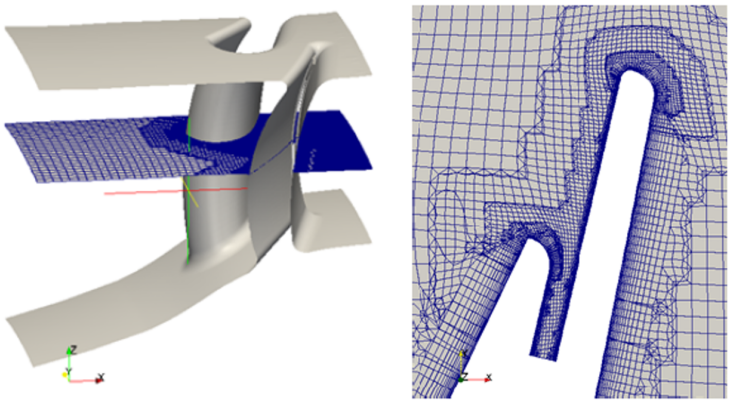
Figure 7: Design velocity fields due to negative parametric perturbations of: a) Casing fillet radius, b) SS profile and c) slot width.

meshing software. The produced meshes have approximately 9 million nodes and 13 million cells. Prior to conducting the CFD simulations, a mesh refinement study was performed to achieve a mesh-independent prediction of capacity. Fig. 8, illustrates the CFD domain, including a detailed view of mesh properties in the TE area.

The capacity results were obtained using the Rolls-Royce in-house CFD solver HYDRA, solving the steady RANS equations with Spalart-Allmaras turbulence model and wall functions. After convergence, the capacity q at the inlet was calculated as

$$q = \dot{m}\frac{\sqrt{T_t}}{p_t}\ , \tag{13}$$

where $\dot{m}$ denotes the inlet mass flow, T_t the total temperature and p_t the total pressure at the inlet using mass averaged values. The resulting capacity for the baseline geometry was found to match the expected capacity value very well and is taken as the reference value for all



(a) (b)

Figure 8: a) NGV CFD domain and b) CFD mesh around TE.

subsequent comparisons with perturbed geometries.

The CFD simulations have shown that a linear relationship between parameter variation and capacity change applies for all of the parameters considered. This means that the adjoint approach should be able to capture the relation and predict accurate results.

Apart from the primal CFD results, the adjoint sensitivity map for the baseline geometry (needed for the proposed approach) was obtained using the HYDRA adjoint solver. The result is illustrated in Fig. 1(a), according to which, the most sensitive area is the suction side.

Finally, the change in capacity caused by each parametric perturbation is predicted by taking the inner product of the sensitivity map with the corresponding design velocity field. The parametric sensitivities can then be computed by dividing with the perturbation step. The obtained derivatives are shown in Fig. 9, where they are compared with central FD values.

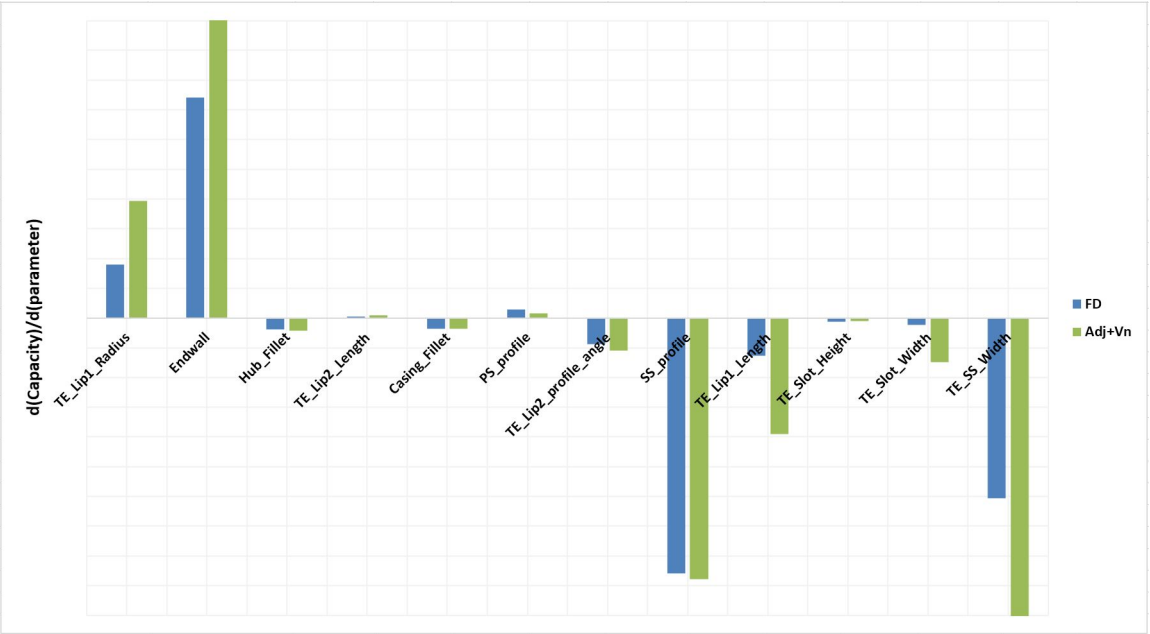


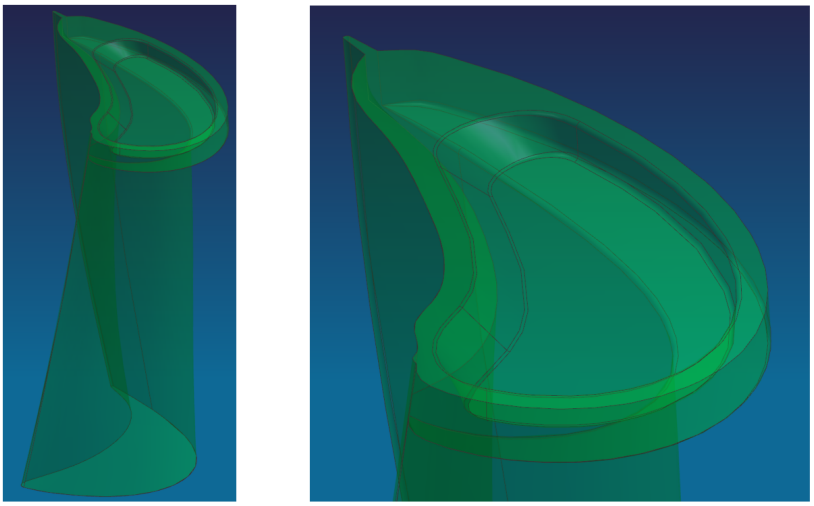
Figure 9: Capacity derivatives predicted by adjoint results (in green) and validation against FD (in blue).

As demonstrated in Fig. 9, for most parameters there is a good agreement between FD and adjoint derivatives. The CPU cost of the design velocity tool was around 2 hours (5 min per parametric perturbation), which is acceptable compared to the flow and adjoint solution. The main advantage of this approach is its efficiency, since it only requires 1 flow solution, 1 adjoint and 2 hours of post-processing, instead of 25 flow solutions (needed for the central FD).

4.2 Rotor blade with winglet

The second case tested in this work is a turbine rotor blade, also provided by Rolls-Royce. The rotor blade geometry is shown in Fig. 10(a), where a 3D parametric CAD model is built again in Siemens NX. The key feature of this blade is its winglet, a detailed view of which is included in Fig. 10(b). This feature is used to reduce the leakage mass flow, which leads to an increased rotor efficiency, by decreasing the driving pressure difference across the rotor's tip. The fluid stays inside the winglet and recirculations occur, as shown in Fig. 13(a), which may cause an increase in total pressure losses within the tip gap. A detailed study on winglets and squealers has been conducted in [25]. In what follows, the study is focused on the winglet geometry and total pressure loss between inlet and outlet is considered as the objective function.

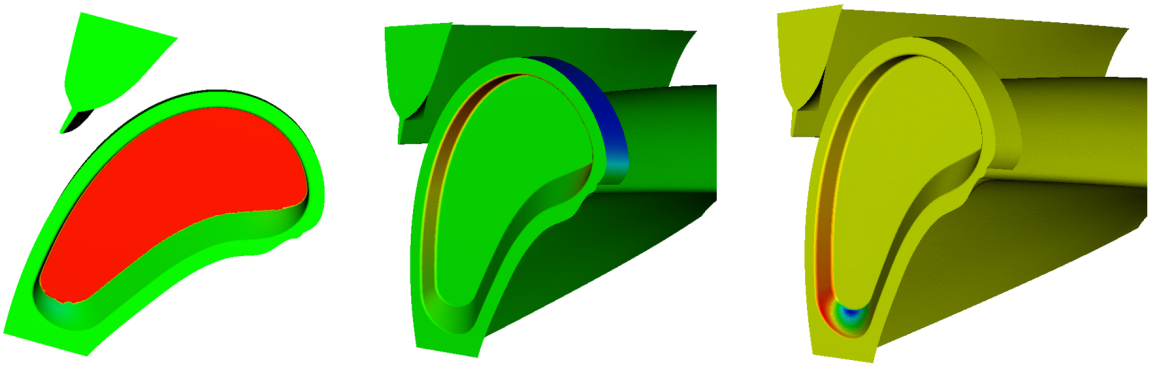
A similar case has been used in [26], where the same approach for calculating parametric sensitivities is implemented and promising results are obtained. However, in this work the further developed QUB tool is combined with the latest HYDRA adjoint version, leading to results of increased accuracy.



(a) (b)

Figure 10: 3D parametric CAD model in Siemens NX: a) Rotor blade geometry and b) winglet.

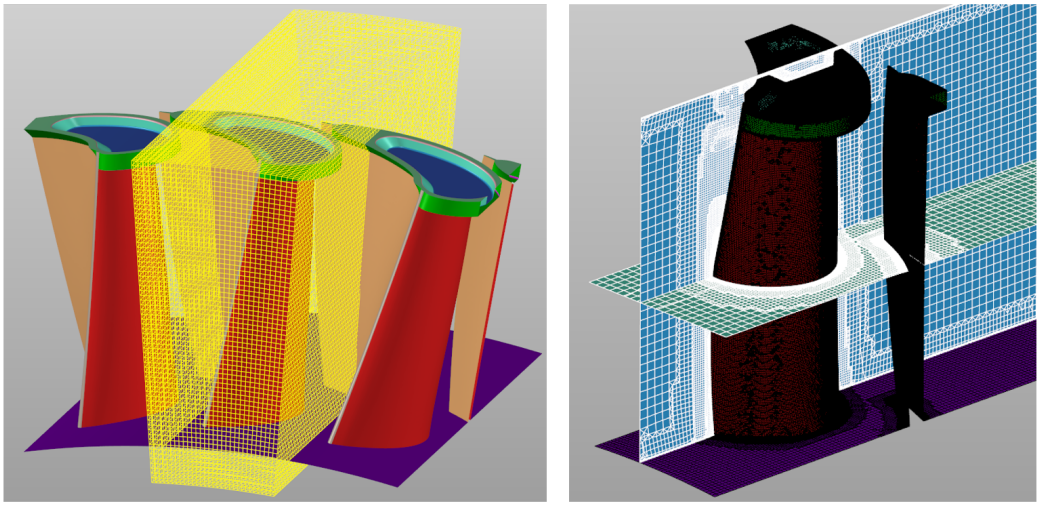
7 CAD parameters controlling the winglet geometry are considered as design variables. For each perturbation of these parameters, a design velocity field was calculated using the QUB tool, which was this time coupled with the commercial mesh generator CADfix, instead of Gmsh. Moreover, bigger perturbation steps were used in this case, in order to test the robustness of the approach. Some of the resulting design velocity fields are shown in Fig. 11.



(a) (b) (c)

Figure 11: Design velocity fields due to parametric perturbations: a) -CAVITY_HEIGHT, b) -SS_FRONT_OFFSET and c) -SS_REAR_ANGLE_PRIMARY.

The CFD mesh was generated again in BOXER, where a periodic sector was created from the initial geometry, as illustrated in Fig. 12(a). The produced CFD domain of the baseline geometry, shown in Fig. 12(b), has around 9 million nodes.

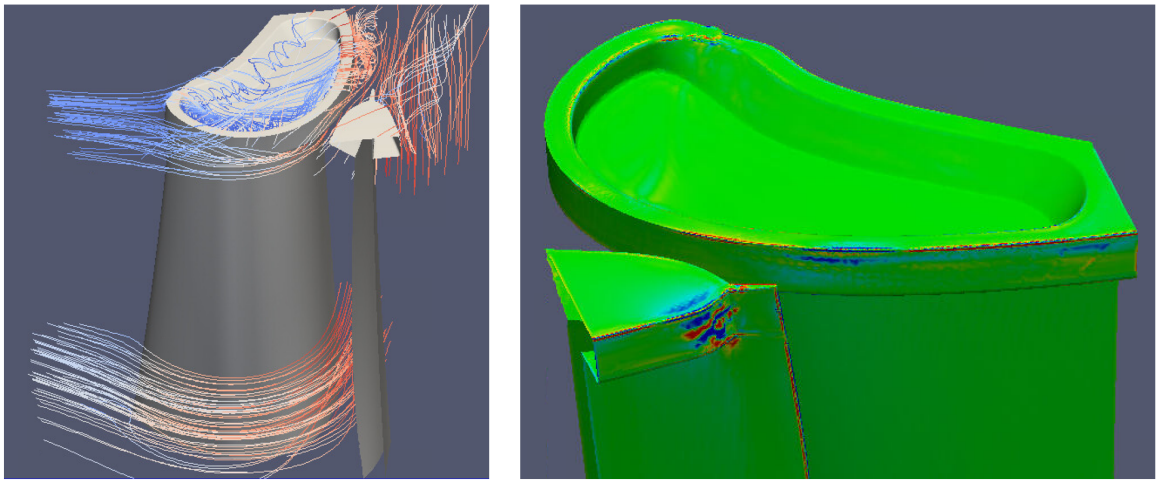


(a) (b)

Figure 12: BOXER mesh generation: a) Periodic section of initial geometry and b) Rotor CFD domain.

The primal and adjoint solutions were obtained from the HYDRA solvers, by using the steady RANS solver with a one equation SA turbulence model and wall functions. The converged flow for the baseline geometry is shown in Fig. 13(a) and the corresponding adjoint sensitivity map is illustrated in Fig. 13(b). The main sources of total pressure loss are the rotor tip vortices.

As in the previous test case, a finite difference (FD) study was also conducted. More specifically, 14 perturbed geometries were exported from Siemens NX and a Python script was used to execute an automated meshing (BOXER) and solving (HYDRA) procedure.



(a) (b)

Figure 13: Flow and adjoint solution: a) Streamlines coloured by relative velocity magnitude and b) Sensitivity map focused on winglet.

The derivatives of the pressure losses w.r.t. each parameter, predicted by linking the design velocities with the sensitivity map, are compared with the corresponding central FD values in Fig. 14. Since the results computed by the adjoint approach are again similar to the FD, further confidence is gained that the proposed method works as expected. The CPU cost of the design velocity tool was around 48 min, which is again negligible compared to the primal and adjoint solution.

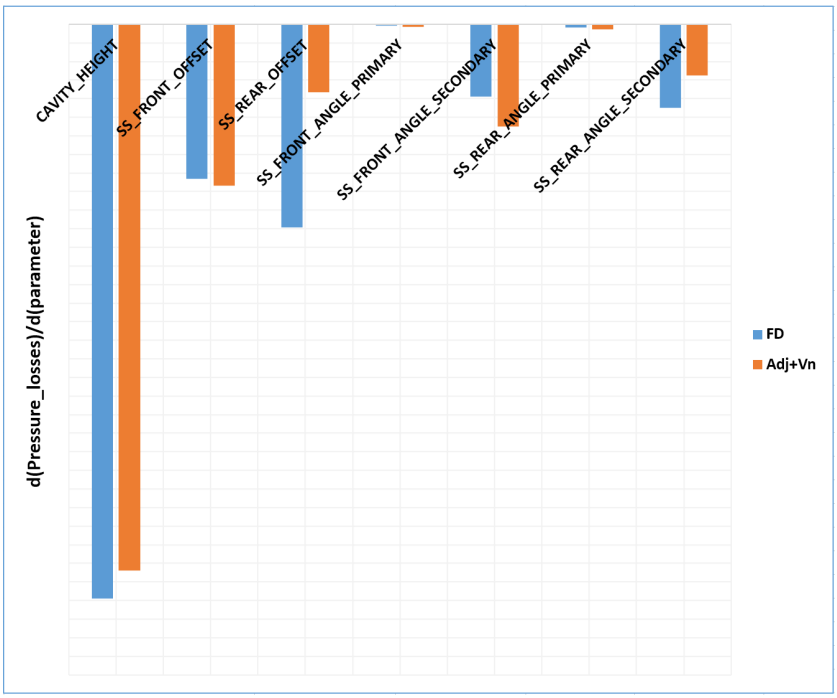


Figure 14: Pressure losses derivatives predicted by adjoint results (with orange) and validation against FD (with blue).

5 CONCLUSIONS

This paper has presented an efficient and robust method for calculating derivatives w.r.t. CAD parameters, based on link of adjoint surface sensitivities with geometric sensitivities, called design velocities, computed from a parametric CAD model. Design velocities are obtained from finite differences between the CAD model before and after a parameter perturbation. The method has been applied to two turbomachinery test cases and the gradient results have been shown to match those computed using finite differences.

Possible sources of error causing the remaining differences between the adjoint derivatives and the FD could be: a) the design velocities, b) the finite differences and c) the adjoint sensitivity map. The design velocity fields have been extensively checked for both cases and they were found to be smooth and without spurious design velocities caused by wrong projections. The FD cannot be considered to be exact, especially for the Rotor case where there was no refinement study carried out. The adjoint sensitivity map is also contributing to the differences, due to the noise that characterizes the surface sensitivities. However, since the sign of the derivatives is correctly calculated, the approach can already be used for steepest descent optimization.

In terms of future work, the possibility of improving the sensitivity map for BOXER meshes will be examined. Then, the work will be focused on passing the gradient information to an optimization algorithm, in order to create a fully automated CAD-based aerodynamic optimization system.

ACKNOWLEDGMENTS

This work has been conducted within the IODA project (http://ioda.sems.qmul.ac.uk), funded by the European Union HORIZON 2020 Framework Programme for Research and Innovation under Grant Agreement No. 642959. The authors would like to thank Rolls-Royce Deutschland for the permission to publish the work.

REFERENCES

[1] M.B. Giles, N.A. Pierce, *An introduction to the adjoint approach to design*. Flow, turbulence and combustion **65:** 393–415, 2000.

[2] A. Jameson, *Aerodynamic shape optimization using the adjoint method*. Lectures at the Von Karman Institute, Brussels, Belgium, 2003.

[3] M.B. Giles, M.C. Duta, J.D. Muller, N.A. Pierce, *Algorithm developments for discrete adjoint methods*. AIAA Journal **41:** 198–205, 2003.

[4] D.I. Papadimitriou, K.C. Giannakoglou, *Compressor blade optimization using a continuous adjoint formulation*. ASME Turbo Expo GT2006/90466, Barcelona, Spain, May 8-11, 2006.

[5] C. Othmer, *A continuous adjoint formulation for the computation of topological and surface sensitivities of ducted flows*. International Journal for Numerical Methods in Fluids **58:** 861–877, 2008.

[6] D.M. Fudge, D.W. Zingg, R. Haimes, *A CAD-free and a CAD-based geometry control system for aerodynamic shape optimization*. AIAA Paper 2005-0451, 2005.

[7] J.F. Dannenhoffer, R. Haimes, *Design sensitivity calculations directly on CAD-based geometry*. AIAA Paper 2015-1370, 2015.

[8] S. Xu, W. Jahn, J.D. Muller, *CAD-based shape optimisation with CFD using a discrete adjoint*. International Journal for Numerical Methods in Fluids **74:** 153–168, 2014.

[9] S. Xu, D. Radford, M. Meyer, J.D. Muller, *CAD-based adjoint shape optimisation of a one-stage turbine with geometric constraints*. ASME Turbo Expo GT2015-42237, Montreal, Canada, June 15-19, 2015.

[10] K.K. Choi, Kuang-Hua Chang, *A study of design velocity field computation for shape optimal design*. Finite Elements in Analysis and Design **15:** 317–341, 1994.

[11] E. Hardee, Kuang-Hua Chang, J. Tu, K.K. Choi, I. Grindeanu, X. Yu, *A CAD-based design parameterization for shape optimization of elastic solids*. Advances in Engineering Software **30:** 185–199, 1999.

[12] J. Kripac, *A mechanism for persistently naming topological entities in history-based parametric solid models*. Computer-Aided Design **29:** 113–122, 1997.

[13] J. Chen, M. Freytag, V. Shapiro, *Shape sensitivity of constructively represented geometric models*. Computer Aided Geometric Design **25:** 470–488, 2008.

[14] T.T. Robinson, C.G. Armstrong, H.S. Chua, C. Othmer, T. Grahs, *Optimizing parameterized CAD geometries using sensitivities based on adjoint functions*. Computer-Aided Design and Applications **9:** 253–268, 2012.

[15] P.M. Thompson, T.T. Robinson, C.G. Armstrong, *Efficient CAD-based aerodynamic design optimization with adjoint CFD data*. 21st AIAA Computational Fluid Dynamics Conference, San Diego, California, United States, 2013.

[16] A. Griewank, G.F. Corliss, *Automatic differentiation of algorithms: theory, implementation, and application*. Defense Technical Information Center, 1992.

[17] K.K. Choi, N.H. Kim, *Structural sensitivity analysis and optimization 1: Linear systems*. Springer, 2005.

[18] S. Shahpar, L. Lapworth, *Padram: parametric design and rapid meshing system for turbomachinery optimisation*. ASME Turbo Expo GT2003-38698, collocated with the 2003 International Joint Power Generation Conference, Atlanta, Georgia, USA, June 16-19, 2003.

[19] S. Shahpar, S. Caloni, *Aerodynamic optimization of high-pressure turbines for lean-burn combustion system*. Journal of Engineering for Gas Turbines and Power **135:** 055001, 2013.

[20] M.B. Giles, *On the iterative solution of adjoint equations*. Automatic Differentiation of Algorithms: 145–151, 2002.

[21] P. Moinier, J.D. Muller, M.B. Giles, *Edge-based multigrid and preconditioning for hybrid grids*. AIAA Journal **40:** 1954–1960, 2002.

[22] J.H. Friedman, J.L. Bentley, R.A. Finkel, *An algorithm for finding best matches in logarithmic expected time*. ACM Transactions on Mathematical Software **3:** 209–226, 1977.

[23] C. Ericson, *Real-time collision detection*. Morgan Kaufmann Publishers, 2005.

[24] L. Hogner, M. Meyer, A. Nasuf, P. Voigt, M. Voigt, K. Vogeler, C. Berridge, F. Goenaga, *Analysis of high pressure turbine nozzle guide vanes considering geometric variations*. In Proceedings of ASME Turbo Expo GT2016-57502, Seoul, South Korea, June 13-17, 2016.

[25] Z. Schabowski, H. Hodson, *The reduction of over tip leakage loss in unshrouded axial turbines using winglets and squealers*. ASME Turbo Expo GT2007-27623, Montreal, Canada, May 14-17, 2007.

[26] M.W. Finnegan, *Shape sensitivity of a turbine blade*. Master's thesis, School of Mechanical and Aerospace Engineering, Queen's University Belfast, 2012.

[27] OpenCASCADE: Open CASCADE Technology, 3D modeling & numerical simulation, 2012. [http://www.opencascade.com]

[28] C. Geuzaine, J.F. Remacle, *Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities*. International Journal for Numerical Methods in Engineering **79:** 1309–1331, 2009.

[29] CADfix: [http://www.iti-global.com/cadfix]

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A CODE-COUPLING APPROACH TO THE IMPLEMENTATION OF DISCRETE ADJOINT SOLVERS BASED ON AUTOMATIC DIFFERENTIATION

Jan Backhaus[1], **Anna Engels-Putzka**[1] **and Christian Frey**[1]

[1]German Aerospace Center (DLR)
Institute of Propulsion Technology
Linder Höhe, D-51147 Cologne, Germany
e-mail: jan.backhaus, anna.engels-putzka, christian.frey@dlr.de

Keywords: discrete adjoint; automatic differentiation; RANS; turbomachinery

Abstract. *We propose a method for selectively applying automatic differentiation (AD) by operator overloading to develop the discrete adjoint of a turbomachinery flow solver. A fully differentiated version of the solver is generated by operator overloading using the tapeless tangent mode of ADOL-C. The differentiated solver is coupled to an undifferentiated version of the same code using message passing. The automatic differentiation is used to calculate derivatives of the flux calculation routines. The flux derivatives depending on inner cell states are sparse, and this sparsity is exploited using analytical differentiation of the spatial discretization scheme. Subsequently the sparse matrix is communicated to the undifferentiated code for solution. Turbomachinery boundary conditions may have dense Jacobians and are therefore only evaluated during the solution process. The solution of the adjoint system of equations is achieved through a preconditioned GMRES, implemented inside the undifferentiated code. A modern three dimensional contra-rotating fan stage with engineering parameterization serves as application example in order to demonstrate the technique and to perform numerical validations. The validation of gradient results is performed by comparing against results from finite differences, and the tangent forward mode.*

1 INTRODUCTION

The adjoint method is a key technology to enable gradient based optimization methods with expensive fluid dynamics simulations. It also has interesting applications in mesh adaptation, robust design and the evaluation of shape variations either due to manufacturing tolerances or wear during operations. Since the adjoint method gained popularity much later than the application of flow simulations, it is often necessary to develop consistent adjoint solvers for long existing CFD solvers. One way to obtain an adjoint solver is to adjoin the underlying PDEs and discretize these; this is called the continuous approach. The discrete approach in contrast is to take the discretized PDEs which are implemented as the primal flow solver and adjoin these. We follow the discrete adjoint approach in this publication. There are two basic techniques for developing a discrete adjoint solver: the white box and the black box approach. For the white box approach, the developer reads a calculation routine, writes down the computation in analytical form, derives its adjoint and subsequently implements the derived calculation. The black box approach is to use a tool for automatic differentiation (AD) in reverse mode, which yields the exact adjoint calculation routine. Both methods have benefits and shortcomings. The white box approach has the potential for the most efficient result, but it requires a high level of insight into the code and is susceptible for inconsistencies between the primal and the adjoint code. While the black box approach solves these white box issues it can lead to inefficient results, especially when applied to iterative solution schemes for PDEs. In practice, a combination of both techniques is often used with success. Especially the selective application of AD inside a white box framework is described by a number of publications [1, 2, 3, 4]. Automatic differentiation can be achieved in two ways: source to source transformation and operator overloading. While the source to source technique is quite popular for solvers written in Fortran, the operator overloading approach is found more advantageous for solvers written in C. The following is a description of how selective application of AD by operator overloading may be achieved in the frame of a parallel solver, starting from the adjoint solver described in [5]. TRACE is a simulation suite for internal flows, with focus on turbomachinery applications; it is developed at DLR, in productive use at MTU AeroEngines and a research tool at several universities.

2 FLOW SOLVER

The most frequently used tool for industrial turbomachinery design optimizations are the compressible steady Reynolds-averaged Navier-Stokes (RANS) equations in a rotating frame of reference, discretized by the finite volume approach. These equations have the general form

$$\frac{\partial q}{\partial t} + \operatorname{div} F(q) - S(q) = 0 \tag{1}$$

with the conservative state in each cell denoted by $q = (\rho, \rho u, \rho v, \rho w, \rho E)$, the fluxes in three directions $F = (F^1, F^2, F^3)$ with $F^i : \mathbb{R}^5 \to \mathbb{R}^5$ and the source terms $S(q)$. For the discretization of convective fluxes we use Roe's TVD upwind scheme in combination with the MUSCL approach after van Leer, viscous fluxes are calculated by a central difference scheme. The steady solution is obtained by implicit pseudo-time marching.

Turbomachinery simulations are characterized by a set of specialized boundary conditions which influence how an adjoint solver can be constructed. Boundary conditions are applied by prescribing values on extra layers, called ghost cells, which extend the mesh beyond the boundary of the physical domain. We therefore divide the flow state into states in internal and

external cells

$$q = (q_{int}, q_{ext}) \tag{2}$$

where the external cells values are prescribed by a functional relationship

$$q_{ext} = \mathcal{F}(q_{int}, q_{ext}). \tag{3}$$

The class of non-local boundary conditions poses special challenges here. In the case of non-reflecting boundary conditions after Giles [6] the ghost-cell update is defined as the fixed point iteration

$$\Delta q_{ext}^{n+1} = \mathcal{F}(q_{int}^{n}, q_{ext}^{n}). \tag{4}$$

Whereas $\mathcal{F}$ involves a Fourier transformation in circumferential direction in order to suppress incoming waves at interfaces of the computational domain.

Blade rows with different rotational velocities are, in the steady case, coupled by Denton's mixing plane approach [7], which requires circumferentially mixed out states to become identical on each side of the mixing plane. Circumferential averaging of states is performed by band-wise integration over fluxes and then applying the inverse flux function to obtain the integral state. This reads in cylindrical coordinates

$$\bar{q}^F = F_c^{-1}\left(\frac{1}{\Delta\vartheta}\int_0^{\Delta\vartheta} F_c(q)d\vartheta\right), \tag{5}$$

where Δq_{ext}^{n+1} vanishes in the case of a converged flow solution.

The boundary conditions mentioned before depend on flow states in a circumferential direction. A span-wise dependency of states may additionally occur at exit surfaces, when the static pressure at the exit is prescribed by the radial equilibrium condition of pressure and centrifugal forces

$$dp = \rho V_\vartheta^2 \frac{dr}{r}, \tag{6}$$

whereas p denotes the static pressure, V_ϑ the circumferential component of the flow velocity and r the radius.

3 ADJOINT SOLVER

For the fields of application mentioned in the introduction it is desirable to calculate partial derivatives of cost functions $I(q)$, calculated from the flow solution q which depends on many design parameters α. An efficient way to evaluate $\frac{\partial I}{\partial \alpha_i}$, with α_i denoting the i-th design parameter, for a large number of parameters is the adjoint method. Since the adjoint approach is well documented in the literature, e.g. in [8], only the basic relations are repeated here. We want to evaluate

$$\frac{\partial I(q(\alpha))}{\partial \alpha_i}, \tag{7}$$

from the flow state q which is defined implicitly by fulfilling the discretized Navier-Stokes equations

$$R(\alpha, q(\alpha)) = 0. \tag{8}$$

One way this can be calculated is solving the adjoint system for the cost function I

$$\left(\frac{\partial R}{\partial q}\right)^t \psi = \left(\frac{\partial I}{\partial q}\right)^t \tag{9}$$

and afterwards evaluate the scalar product

$$\frac{\partial I}{\partial \alpha_i} = -\psi^t \frac{\partial R}{\partial \alpha_j} \tag{10}$$

for each design parameter. The computational costs of solving Equation (9) are about as high as those for solving the nonlinear system of Equations (1) while evaluating the matrix vector product in Equation (10) is comparatively cheap. The adjoint method is therefore efficient when the number of parameters is larger than the number of cost functions.

3.1 Flux linearization

A central point in the implementation of the adjoint solver is the exact calculation of $\frac{\partial R}{\partial q}$. The linearization of the discretized steady Eqn. (1) may be written as

$$\frac{\partial R_i}{\partial q_k} = \sum_j \frac{\partial F_j}{\partial q_k} + \delta_{ik} \frac{\partial S_i}{\partial q} = 0 \tag{11}$$

where i runs over all cells, j over the cell's faces and k over all states. Since the flux at a cell face depends only on a few flow states in neighboring internal cells q_{int}, the Jacobian matrix $\frac{\partial F}{\partial q_{int}}$ can be efficiently calculated by exploiting the sparsity pattern of the spatial discretization scheme.

While the nonlinear solver can be used with a variety of turbulence models, it is assumed here, that the eddy viscosity is not influenced by small geometric changes and may therefore be ignored for the adjoint system of equations. This is called the constant-eddy-viscosity (CEV) assumption. The validity of the CEV assumption for design optimizations is discussed by various authors, e.g. [9, 10].

3.2 Adjoint boundary conditions

While the flux Jacobian for inner cell states may be calculated with the above scheme, the fluxes on cell faces close to the boundary of the computational domain depend on external cell states and through Eqn. (3) on the boundary conditions. When differentiating Eqn. (11) with respect to internal and external flow states, one obtains the residual Jacobian matrix

$$\frac{\partial R}{\partial (q_{int}, q_{ext})} = (A_{int}\ A_{ext}). \tag{12}$$

Since q_{ext} depends on q_{int} through Eqn. (4) in the case of non-reflecting boundary conditions, one would have to differentiate this giving a fixed point iteration for the linear boundary condition

$$\Delta(\delta_{ext}^{n+1}) = \frac{\partial \mathcal{F}}{\partial (q_{int}, q_{ext})} \begin{pmatrix} \delta q_{int}^n \\ \delta q_{ext}^n \end{pmatrix} \tag{13}$$

However, the adjoint system of equations is, in contrast to the primal system, not solved by pseudo-time marching. Consequently, we cannot evolve Eqn. (13) parallel to the solution process. Instead we use

$$\frac{\partial \mathcal{F}}{\partial (q_{int}, q_{ext})} \begin{pmatrix} \delta q_{int}^n \\ T \delta q_{ext}^n \end{pmatrix} = 0, \tag{14}$$

with the adjoint boundary operator T. Since no routine in the primal solver exist which can be differentiated to obtain T, the operator is manually derived and implemented. The linear

operators T for the turbomachinery boundary conditions, employed here, are derived in [5, 11] and not repeated for brevity. Using the above equations, the linearized residual reads

$$\mathscr{L} = \begin{pmatrix} A_{int} & A_{ext} \end{pmatrix} \begin{pmatrix} Id \\ T \end{pmatrix}. \tag{15}$$

The adjoint is obtained by transposition

$$\mathscr{L}^* = \begin{pmatrix} Id & T^t \end{pmatrix} \begin{pmatrix} A_{int}^t \\ A_{ext}^t \end{pmatrix}. \tag{16}$$

3.3 Solution scheme

The adjoint system of equations (9) is solved by a preconditioned GMRES solver with restarts [12]. The available preconditioners are successive over-relaxation, (SSOR) and incomplete LU-decomposition with limited level-of-fill (ILU).

3.4 Selective use of AD

The introduction of AD into TRACE is described in [13]. The results of this work are used here as a starting point to describe how an AD based adjoint solver can be implemented in a grey box fashion.

More specifically we apply white box techniques to the spatial discretization stencil, objective functions and boundary conditions, while the flux calculation routines are differentiated in a black box fashion through AD in forward mode. From forward derivatives of the flux calculation routines we then build the adjoint system matrix. This grey box approach was chosen in order to obtain an efficient adjoint solver but at the same time avoid the cumbersome differentiation of the sophisticated flux computation routines. The reasons for treating boundary conditions differently from the rest of the flux computations are:

1. The primal boundary conditions are not implemented in simple function calls which could be treated by AD. They are constituted by a composition of functions controlled by the different solver modes and choices of models.

2. The fixed-point property of non-reflecting boundary conditions as discussed in section 3.2.

3. Non-local boundary operators, such as the non-reflecting boundary conditions, radial equilibrium boundary conditions as well as row coupling interfaces depend on a large number of cell states. The differentiation of these would add large dense contributions to the residual Jacobian matrix. In order to keep the memory consumption manageable they must therefore be evaluated on-the-fly during the adjoint solution process.

We employ the tangent forward mode of automatic differentiation which calculates directional derivatives by attaching to each variable in the calculation an additional derivative variable and propagating this derivative value alongside the computation:

$$y = f(x),\ \dot{y} = \frac{df}{dx}\dot{x}, \tag{17}$$

where $\frac{df}{dx}$ is analytically defined by the chain rule of differentiation for each elementary operation, i.e., all floating operations intrinsically provided by the C language. For a more comprehensive description of AD, the reader is referred to [14]. Since TRACE is written in the C

language, more specifically the C99-standard, we apply automatic differentiation by operator overloading, due to the easier implementation and to avoid the dependency on another compiler and the restrictions to supported language constructs. The operator overloading approach works by providing a class which re-defines all elementary operations to calculate y and $\dot{y}$ as in Eqn. (17). Here we use ADOL-C [15] in tapeless tangent mode. The general workflow for applying AD by operator overloading is to change the definition of the floating point datatype used for calculations from the standard floating point datatype to the AD class (called `adouble` in ADOL-C). During the execution one sets the differentiation seed $\dot{x}$ via a special member function (called `setADValue` in ADOL-C) of the independent variable x and runs the function to be differentiated. The directional derivative is then computed alongside the computation and its derivatives $\dot{y}$ can afterwards be obtained by calling another member function (called `getADValue` in ADOL-C) for the desired output variable. While this is the simplest way of applying AD, there are a few obstacles observed during the introduction of AD in TRACE:

- Operator overloading is not defined for C, only for C++. Even though C++ is sometimes called a superset of C, this is not exactly true. Not all programs conforming to the C99-standard are also valid C++ programs. See [16] for a more detailed view on this issue. In this context the most important issues are variable length arrays (VLA), which are not supported in C++, and unions containing active floating point variables[1]. After eliminating incompatibilities it must be communicated to all developers which constructs must be avoided and how these should be replaced. Automated testing of each new code revision is necessary to ensure that no incompatibilities are introduced.

- External libraries receiving and returning floating point numbers must be differentiated, by either automatic differentiation or, if possible, by high-level manual differentiation[2].

- All floating point variables passed to an external library for output, including C's standard libraries, for which no overloaded equivalent is provided by the AD tool, must be converted to standard datatypes. This requires wrapper-code around the actual library call. For often recurring calls this may be facilitated by using macros. Functions with variable number of arguments, e.g. `printf` type functions, require more elaborate solutions, e.g. variadic templates.

- Some computations are not differentiable, e.g. calculating the length of a vector which may become zero at some point during the calculation. For such results the AD value $\dot{y}$ takes the special values NaN or inf which propagate through the whole derivative computation. While most of the time it is simple to circumvent such calculations by improvements to the primal code (cf. [14] for a list of applicable techniques), finding and identifying such computations requires thorough testing.

3.5 Interfacing differentiated with non-differentiated code

When differentiating an existing code base using AD, one has to determine which variables are active, i.e., if variables are used inside the dependency path between the selected inputs and outputs. In the operator overloading approach a variable is marked as active by changing its datatype from the standard arithmetic type to the respective AD class. It is necessary to correctly

[1]The C++11 standard improves on some of these issue, but it is still necessary to define constructors for such unions

[2]The fast fourier transformation is a candidate for this

identify all variables that have to be active. Any missed out declaration either leads to compile time errors, or to wrong gradient results. Superfluously marking a passive variable as active is less problematic; this only increases the computational effort and memory requirements. Since correctness has to be reached before performance optimizations can be considered, one often starts by changing all floating point type declarations to the AD type and handle only the interface between this active types and the external libraries.

Selectively changing variables reduces the performance overhead to only those routines one wants to differentiate. However this requires more work from the developer to handle the interfaces between active and inactive variables. At these interfaces, all conversions from active to inactive variables have to be explicitly implemented by introducing conversion calls. Composite types complicate things further, since they may be used in contexts where their member variables are active in others where they are passive. Even combinations where only parts of the composite type have to be active may occur. A solution would be to create copies of the data type for each pattern of activation, copy and adapt the calling routines. Since the activation pattern may depend on usage scenarios, this would lead to a lot of copied routines with slightly varying activation patterns which is undesirable for code-maintainability reasons. This approach may be automatized by a script. However that script has to mimic parts of a source-to-source AD tool and therefore may become complicated, depending on the range of supported language constructs.

In order to circumvent the development of such a script another approach for the selective application was chosen here: Two executables are generated from the same code base using a compile time switch. In the unmodified executable, all variables are defined as standard-arithmetic types. In the differentiated executable, all variables are defined as the AD class. Both contain the same set of functions and may be tested independently from each other to ensure binary identical results from their primal calculations. In order to use exact derivatives inside the faster unmodified executable, the unmodified launches the differentiated executable and both perform the basic flow solver initialization, c.f. Fig. 1. The flux Jacobian matrix is generated inside the differentiated solver by iterating over each component of the flow state q^i in each internal cell, setting $\dot{q}^i$ to 1. After calling the flux-functions for the respective cell, the derivatives are obtained by reading all $\dot{F}^j$ as standard floating point values and storing these inside the flux Jacobian matrix in block compressed row storage (BCRS) format, where each submatrix is a 5x5 matrix. This matrix is subsequently communicated to the non-differentiated executable, where it constitutes the system matrix for the adjoint system of equations. Since the differentiated executable is terminated at this point, the additional memory for storing the differentiated flow field is freed. The adjoint system of equations is solved by inside the non-differentiated code as described in section 3.3.

4 VALIDATION

4.1 Procedure

For the validation of the sensitivities we now have three different validation procedures available. The first two are performed by running the complete code for a deformed mesh, and these are therefore primal procedures. One can run the undifferentiated executable calculating finite differences from the results, or one can run the differentiated executable and use the deformation at each mesh node as a seeding direction. These procedures are called FD-primal and AD-primal respectively. The other means of validation is to use finite differences instead of AD in the grey box adjoint code. This is called FD-adjoint in the following text. The solver to be

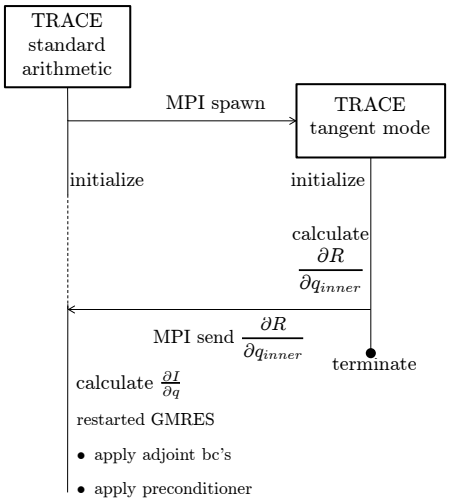


Figure 1: Sequence diagram for the code coupling strategy.

validated is the AD based adjoint, called AD-adjoint, consequently.

For the FD-primal evaluation first order accurate forward differences are obtained from the first terms of the Taylor series expansion

$$\frac{dI}{d\alpha_i} = \frac{I(\mathbf{x} + \mathbf{p}h) - I(\mathbf{x})}{h} + \mathcal{O}(h). \tag{18}$$

The grid coordinates $\mathbf{x}$ are perturbed by a deformation vector $\mathbf{p}$ which is obtained by modifying one design parameter α_i and generating the corresponding computational grid

$$\mathbf{p} = \mathbf{x}(\boldsymbol{\alpha} + \mathbf{e}_i\epsilon_i) - \mathbf{x}(\boldsymbol{\alpha}) \tag{19}$$

The choice of h is a difficult task, due to the linear dependence of the truncation error on the one hand, and cancellation errors on the other hand. Therefore different values of h are examined to find a range where the result is insensitive to the choice of h. Finite differences have the advantage of being very simple and they serve therefore as reference to rule out human error or flaws in the AD tool. However, as the associated approximation error may become very large finite differences cannot provide full confidence in the correctness of the results. The second simple approach is the AD-primal procedure, which serves as an alternative way to calculate derivatives in a forward manner. It has no associated step-width problem. This AD-primal calculates any cost function I available inside the flow solver and $\dot{I}$ with respect to one design parameter provided as deformed mesh in one sweep

$$I(x),\ \dot{I} = \frac{\partial I}{\partial \mathbf{x}}\mathbf{p}. \tag{20}$$

For $\mathbf{p}$ one can use the mesh deformations as in Eqn. (19) and apply these as seed direction for the AD variables on the grid points. In principle, this is the same as the finite difference evaluation described before with a step width approaching zero and no subtractive cancellation present. This is equivalent to the complex step derivative technique [17], which could be used alternatively. Since both primal procedures have a computational cost proportional to the number of design parameters, they are only used for validation purposes.

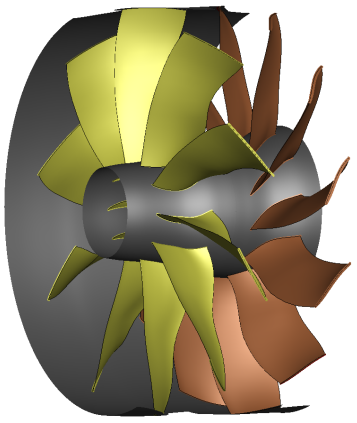


Figure 2: The contra-rotating integrated shrouded propfan CRISP 2.

The third solver is the FD-adjoint, where the flux Jacobian matrix is computed from finite differences. By means of this solver we are able to tell whether the Jacobian matrix calculated by the AD tool is valid and to assess the influence of errors in this matrix on the adjoint sensitivities. The finite difference step-width here is chosen to be proportional to the flow state and the factor of proportionality ε is varied to find a range of insensitivity of the finite differences.

4.2 CRISP 2

The second design of the contra-rotating shrouded propfan called CRISP 2 (Fig. 2) is a multidisciplinary optimization based design conducted at DLR [18] using the optimization framework AutoOpti [19]. The design targets are to develop a new engine concept for extremely high bypass-ratios (>20), in order to achieve a high fan efficiency and to reduce noise emissions for the important operating points. To this end, aerodynamic, acoustic and mechanic objectives or constraints were considered in the optimization [20]. The optimization has been carried out using a combination of evolutionary algorithms and Gaussian process meta models constructed from 3D simulations. The design used here has a mass flow rate of 158 kg/s, a total pressure ratio of 1.3 and an isentropic efficiency of 94 %. With a pre-shock Mach number of about 1.2 the flow is transonic. A validation of the CFD solver on modern contra-rotating fans against experimental results is described in [21]. See [22] for a description of how gradients from an adjoint solver may be included in this optimization. Different mesh resolutions can be generated from this process, here a coarse structured mesh consisting of 500 000 cells is used.

The CAD Model of the fan stage is parameterized by engineering CAD Parameters which are translated into B-Spline tensor product surfaces using the geometry tool Blade-Generator. Based on these surface definitions a structured grid is created using the mesh generator G3DHexa. Deformed meshes for the sensitivity calculation process are created by an elliptic mesh deformation tool [23].

From the 123 parameters of the original optimization we have selected 24 representative parameters. This was necessary to reduce the computational effort spent on the primal validation procedures. The selected parameters are the stagger angle, leading edge angle and trailing edge angle on four profiles of each of the two rotors.

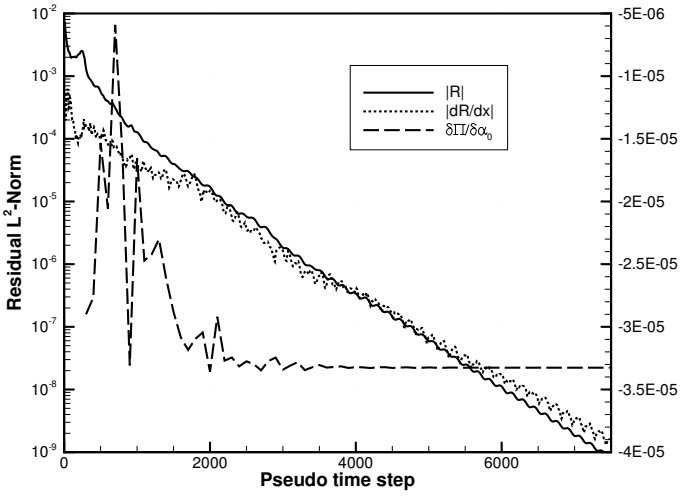


Figure 3: Residual, differentiated residual and convergence of sensitivity of the total pressure ratio for the AD-primal process

4.3 Validation using primal calculations

The convergence of the primal residual and the residual of the forward differentiation, AD-primal, are shown in Fig. 3, additionally for the total pressure ratio Π the evolution of the sensitivity $\frac{\partial \Pi}{\partial \alpha_0}$ is plotted exemplarily. All primal validation calculations are stopped when the residual falls below a threshold of 10^{-9}. While the residual still drops afterwards, the accuracy in all output quantities is sufficient, as can be seen for the value of the sensitivity in Fig. 3.

4.4 Validation using FD-adjoint

The adjoint solver based on finite differences is used to demonstrate the effect of approximations on the residual Jacobian matrix here. From Fig. 4 it can be seen that the convergence of the adjoint residual improves, when a larger finite difference step is used. Errors from the finite differences act as a regularization of the adjoint system matrix. The convergence behavior of the AD-adjoint solver is identical to the convergence with $\varepsilon = 10^{-11}$. However, the smallest error in the sensitivities for the first parameter, comparing AD-adjoint to FD-adjoint, can be observed for $\varepsilon = 10^{-4}$ in Fig. 5.

4.5 Comparison of Gradients

Figure 6 shows the comparison of the calculated gradients for the total pressure ratio from the four different approaches. One can see that these are in very good agreement. Minor deviations can be observed for some parameters, between the sensitivities from the primal and the adjoint approaches. Since the AD approaches show the same results as the FD approaches, an error in the AD tool or its application can be ruled out. The absolute errors of the gradients are calculated as the absolute value of the difference between AD-adjoint and AD-primal and plotted in Fig. 7.

The magnitude of errors is found to be sufficiently small for the optimization procedures described in [22].

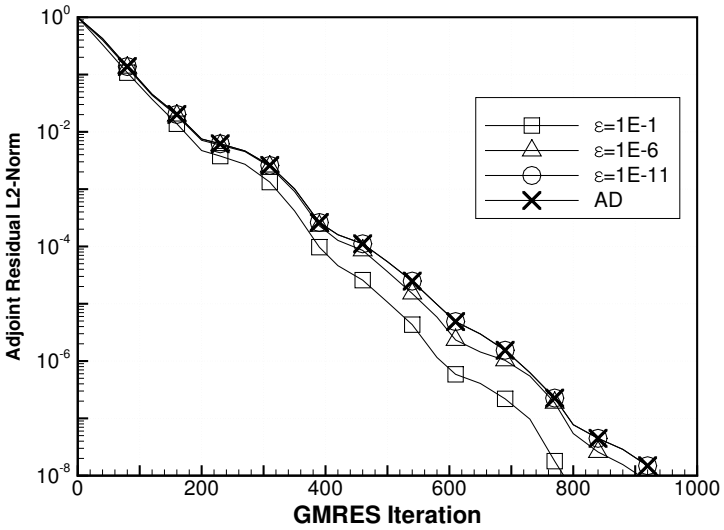


Figure 4: Residuals of the FD-adjoint process with varying step width ϵ.

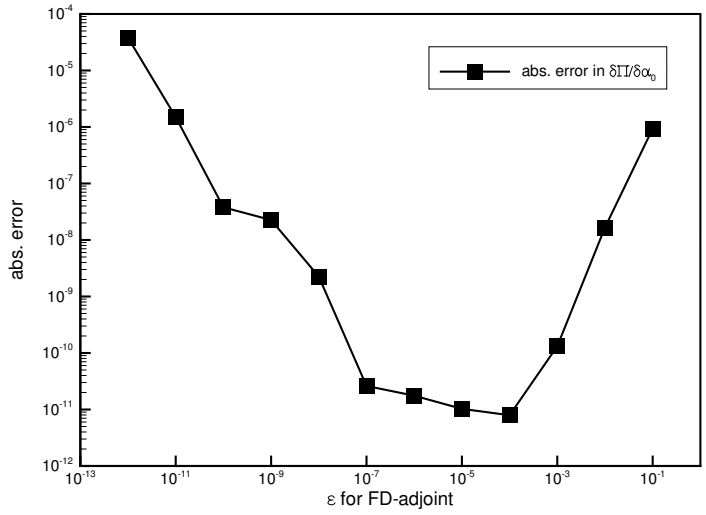


Figure 5: Errors in sensitivities for FD-Adjoint compared to AD-adjoint over step width ϵ.

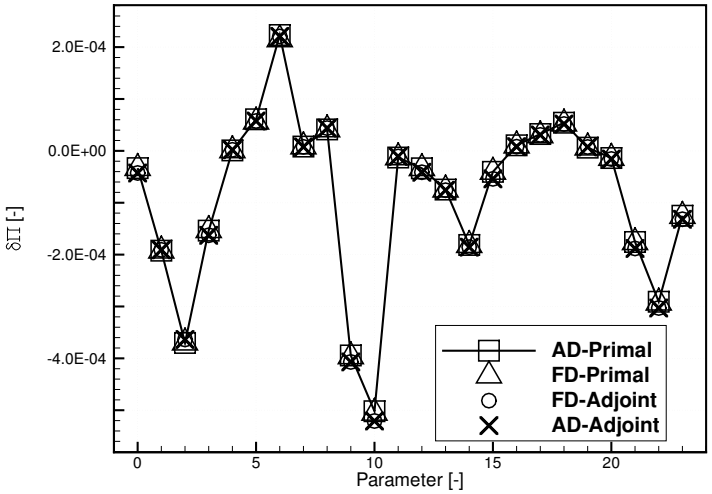


Figure 6: Validation of partial derivatives for the total pressure ratio of CRISP 2 w.r.t 24 CAD parameters.

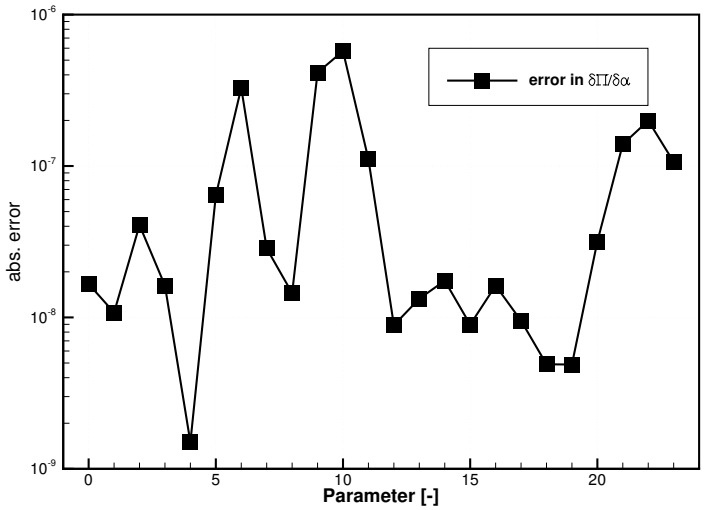


Figure 7: Error of the AD-adjoint sensitivities compared to AD-primal.

5 CONCLUSIONS AND OUTLOOK

Developing the adjoint for an existing CFD solver can become a laborious task which requires good knowledge of the solver and of the equations implemented therein. This amount of knowledge may be replaced, in parts, by the ability to work with AD tools. Correctly applied AD provides exact derivatives which stay consistent to the calculations, even when the solver is modified, given that certain coding guidelines are obeyed. Besides a lot of effort spent on the development of the AD tools, a complete black box approach is still not optimal for the iterative solution processes. The approach presented here shows how AD can be selectively applied for an industrial, parallelized turbomachinery solver written in the C language by using operator overloading. The approach taken here starts from a white box derivation and selectively includes black box techniques. A benefit of the approach is the ability to use the same AD-primal solver which is used to calculate the flux Jacobian for gradient validation purposes and therefore gain confidence in the consistency of the computed flow solution and the derivatives. An alternative approach is to derive an adjoint solver in a complete black box fashion through reverse mode AD and afterwards introduce white box improvements to lower the runtime and memory consumption. While this strategy has been pursued in [13], a thorough comparison of both approaches will be left for future work.

ACKNOWLEDGEMENT

This research was supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant number 20T1104B. We thank the Chair for Scientific Computing (SciComp) of TU Kaiserslautern, especially Max Sagebaum and Dr. Emre Özkaya for the fruitful cooperation in applying automatic differentiation.

REFERENCES

[1] F. Courty, A. Dervieux, B. Koobus, and L. Hascoet, "Reverse automatic differentiation for optimum design: from adjoint state assembly to gradient computation," *Optimization Methods and Software*, vol. 18, no. 5, pp. 615–627, 2003.

[2] M. Giles, D. Ghate, and M. Duta, "Using automatic differentiation for adjoint cfd code development," *Indo-French Workshop*, 2005.

[3] C. A. Mader, J. R. R. A. Martins, J. J. Alonso, and E. van der Weide, "ADJoint: An approach for the rapid development of discrete adjoint solvers," *AIAA Journal*, vol. 46, pp. 863–873, APR 2008. AIAA/ISSMO 11th Multidisciplinary Analysis and Optimization Conference, Portsmouth, VA, SEP 06-08, 2006.

[4] A. C. Marta, S. Shankaran, D. G. Holmes, and A. Stein, "Development of adjoint solvers for engineering gradient-based turbomachinery design applications," in *Proceedings of the ASME Turbo Expo 2009*, vol. Volume 7: Turbomachinery, Parts A and B, June 2009.

[5] C. Frey, D. Nürnberger, and H.-P. Kersken, "The discrete adjoint of a turbomachinery RANS solver," in *Proceedings of ASME-GT2009*, 2009.

[6] M. B. Giles, "Nonreflecting boundary conditions for Euler calculations," *AIAA Journal*, vol. 28, no. 12, pp. 2050–2058, 1990.

[7] J. Denton and U. Singh, "Time marching methods for turbomachinery flow calculations," VKI Lecture Series 1979-7, von Karman Institute., 1979.

[8] M. B. Giles and N. A. Pierce, "An introduction to the adjoint approach to design," *Flow, Turbulence and Combustion*, vol. 65, pp. 393–415, 2000.

[9] C. S. Kim, C. Kim, and O. H. Rho, "Feasibility study of constant eddy-viscosity assumption in gradient-based design optimization," *Journal of Aircraft*, vol. 40, no. 6, pp. 1168–1176, 2003.

[10] R. Dwight and J. Brezillon, "Effect of approximations of the discrete adjoint on gradient-based optimization," *AIAA Journal*, vol. 44, no. 12, pp. 3022–3071, 2006.

[11] C. Frey, A. Engels-Putzka, and E. Kügeler, "Adjoint boundary conditions for turbomachinery flows," in *ECCOMAS 2012 - European Congress on Computational Methods in Applied Sciences and Engineering, e-Book Full Papers*, 2012.

[12] Y. Saad, *Iterative methods for sparse linear systems. 2nd ed.* SIAM Society for Industrial and Applied Mathematics, Philadelphia., 2003.

[13] M. Sagebaum, E. Özkaya, and N. R. Gauger, "Challenges in the automatic differentiation of an industrial CFD solver," in *Evolutionary and Deterministic Methods for Design, Optimization and Control with Application to Industrial and Societal Problems (EUROGEN 2013)*, 2014.

[14] A. Griewank and A. Walther, *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Siam, 2008.

[15] A. Walther and A. Griewank, "Getting started with ADOL-C," *Combinatorial Scientific Computing*, pp. 181–202, 2012.

[16] B. Stroustrup, "C and C++: Case studies in compatibility," *C/C++ Users Journal*, vol. 20, no. 9, pp. 22–31, 2002.

[17] J. R. R. A. Martins and J. T. Hwang, "Review and unification of methods for computing derivatives of multidisciplinary computational models," *AIAA Journal*, vol. 51, pp. 2582–2599, November 2013.

[18] T. Lengyel-Kampmann, *Vergleichende aerodynamische Untersuchungen von gegenläufigen und konventionellen Fanstufen für Flugtriebwerke:*. PhD thesis, Ruhr-Universität Bochum, 2015.

[19] C. Voss, M. Aulich, B. Kaplan, and E. Nicke, "Automated multiobjective optimisation in axial compressor blade design," in *ASME Paper*, vol. 90420, 2006.

[20] D. Görke, A.-L. Le Denmat, T. Schmidt, F. Kocian, and E. Nicke, "Aerodynamic and mechanical optimization of CF/PEEK blades of a counter rotating fan," in *Proceedings of the ASME Turbo Expo 2012*, 2012.

[21] T. Lengyel-Kampmann, A. Bischoff, R. Meyer, and E. Nicke, "Design of an economical counter rotating fan: Comparison of the calculated and measured steady and unsteady results," in *ASME Turbo Expo 2012: Turbine Technical Conference and Exposition*, pp. 323–336, American Society of Mechanical Engineers, 2012.

[22] J. Backhaus, M. Aulich, C. Frey, T. Lengyel, and C. Voß, "Gradient enhanced surrogate models based on adjoint cfd methods for the design of a counter rotating turbofan," in *Proceedings of the ASME Turbo Expo 2012*, 2012.

[23] C. Voigt, C. Frey, and H.-P. Kersken, "Development of a generic surface mapping algorithm for fluid-structure-interaction simulations in turbomachinery," in *V European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010* (J. C. F. Pereira, A. Sequeira, and J. M. C. Pereira, eds.), June 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CAD KERNEL AND GRID GENERATION ALGORITHMIC DIFFERENTIATION FOR TURBOMACHINERY ADJOINT OPTIMIZATION

Ismael Sanchez Torreguitart[1], Tom Verstraete[2], and Lasse Mueller[1]

[1] Von Karman Institute for Fluid Dynamics (VKI)
Chausse de Waterloo 72, 1640 Rhode-Saint-Gense, Belgium
{ismael.sanchez.torreguitart, lasse.mueller}@vki.ac.be

[2] Queen Mary University of London
Mile End Road, E1 4NS, London, UK
t.verstraete@qmul.ac.uk

Keywords: CAD, grid, sensitivities, algorithmic differentiation, adjoint optimization

Abstract. *Although adjoint-based optimization methods have numerous advantages such as the efficient computation of the gradient virtually independent of the number of design variables, the methods are not yet picked up largely by industry. One major bottleneck herein is that adjoint methods mainly work on deforming the CFD grid and as such loose the connection to CAD, the industry adopted standard for the design of components. After the optimization a step is required which transforms the optimal shape, defined by grid points, back to a smooth CAD shape. This step is complicated, not easily automated and invariably impairs optimality of the shape as CAD systems will only approximate the optimal shape. The work presented herein is a first step to alleviate the problem by expressing the optimization problem through CAD parameters, rather than working directly on the CFD grid. This allows the optimal shape to remain defined within the CAD tool and through the differentiation of the CAD kernel and grid generation tools with algorithmic differentiation (AD) it is possible to obtain the grid sensitivities propagated to the CAD-based parameters. The in-house CAD and grid generation tools have been differentiated in forward mode as a first step. The VKI LS89 axial turbine vane is selected as a demonstrator to test the methodology. The profile is parameterized by 24 design parameters and the geometrical as well as grid sensitivities are obtained for each parameter, showing the influence of each design parameter on the profile and on the CFD grid coordinates. Finally, the sensitivities obtained by AD are compared against second order accurate central finite difference approximations through changing the perturbation step.*

1 INTRODUCTION

Aerodynamic shape optimization using low-cost, efficient and accurate computational methods is key for the aerospace industry, as it provides the opportunity to make significant design improvements at the early stage of the design chain.

For gradient-based optimization strategies it is essential to compute the total gradient of the design chain, which is the derivative or gradient of the cost function with respect to a vector of design variables. This is usually done by calculating the derivatives separately for the various disciplines used in the design chain (e.g. grid generation, Computational Fluid Mechanics, Computational Solid Mechanics, etc.) and then combining them to obtain the total gradient. The derivatives can be obtained using different methods, either analytically, by finite differences or by differentiating the code(s). Adjoint-based optimization methods [1, 2, 3] have been introduced for more than two decades and have witnessed a tremendous improvement over time. They have become increasingly popular amongst the gradient-based optimization methods for its efficiency to compute the gradients at a cost that is independent of the number of design variables [4]. However, one major bottleneck herein is that adjoint methods mainly work on altering the shape to be optimized by deforming the CFD grid [5] and as such loose the connection to computer aided design (CAD), the industry adopted standard for the design of components. After the optimization, a step is required which transforms the optimal shape, defined by grid points, back to a smooth CAD shape. Many different procedures have been developed to approximate a given set of grid points with NURBS curves or surfaces [6, 7]. Despite the fact that these methods can be successfully employed to obtain a smooth CAD shape, the fitting error may invariably impair optimality of the shape as CAD systems will only approximate the optimal shape.

The emphasis of this work is to present a method in which:

1. the optimal shape remains defined within the CAD tool. The optimization problem herein is expressed by CAD parameters that are directly used in defining the CAD geometry by means of Bézier and B-spline curves.

2. the in-house CAD and grid generation tools are automatically differentiated in forward mode to obtain the exact derivatives of the grid coordinates with respect to the CAD-based design parameters. This allows to accurately predict the sensitivities and circumvent the errors introduced by finite differences.

3. the differentiation of the CAD kernel and grid generation tool in forward mode will then serve to validate the sensitivities in reverse mode. The reverse capability would allow enriching the design space by introducing more design parameters without penalising the computational cost.

The work is carried out using a computer aided design and optimization tool for turbomachinery applications (CADO) [8], which has been used so far mainly in gradient free optimization methods, and hence the work presented herein represents the first steps to differentiate the entire CADO code to make it work in gradient based optimisation. The LS89 [9, 10, 11, 12] axial high pressure turbine nozzle guide vane (HPT NGV) is selected as a demonstrator to test the methodology. This profile was designed and optimized at the von Karman Institute for Fluid

Dynamics (VKI) for a subsonic outlet flow by the inverse method [9]. It was extensively tested at the VKI laboratories for a large range of Reynolds and Mach numbers and this experimental data was then used to improve the in-house CFD prediction capabilities in the late 80s.
In the following sections the approach to construct the geometry, fluid computational grid and to compute the geometrical and grid sensitivities is presented. Also discussed are the sensitivities obtained by AD and compared against FD second order central derivatives by tuning the step size.

2 METHODOLOGY

The design chain process shown in Fig. 1 starts typically from a set of input design parameters that allow the designer to build a CAD model, from which it is necessary to generate a computational fluid grid to perform CFD analysis. After obtaining a flow converged solution by the primal solver, the cost function is evaluated as a post-processing step. The differentiation of the design chain allows to compute the performance sensitivities which are necessary for gradient based optimization algorithms. The following subsections present a description of the methods employed in this work for the CAD and grid generation steps and the calculation of the geometrical and grid sensitivities.

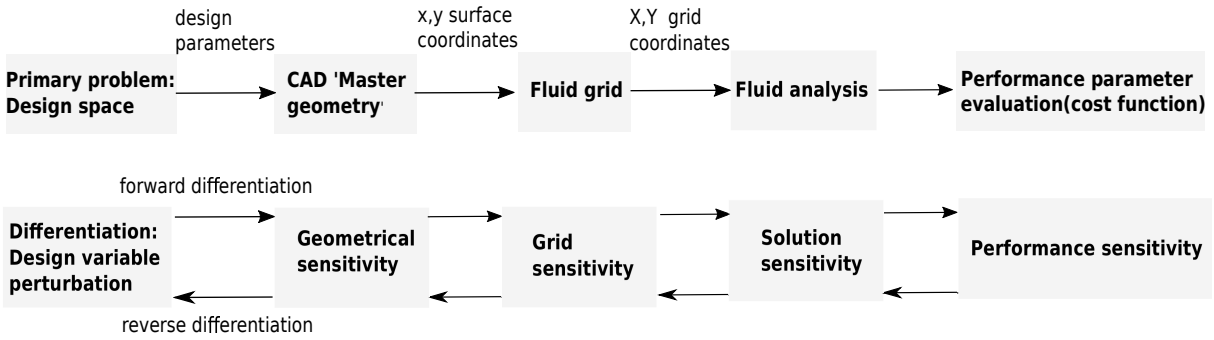


Figure 1: Chain to compute the performance sensitivity

2.1 Using CAD parameters as design variables

In gradient based optimization, it is vital to know the performance sensitivities with respect to design parameter changes. Typically, the grid point coordinates have been used as design variables in order to optimize a design [5]. This means that an additional step is required to convert the optimal shape defined by the grid back to a CAD shape. However, it is not guaranteed that the final CAD shape will meet all the manufacturing design requirements and constrains. Furthermore, the generated CAD shape will only approximate the optimal shape given by the CFD grid point cloud [6]. Hence there is an arising need to keep the CAD geometry in the optimization loop.

2.1.1 Constructing a turbine profile with Bézier and B-spline curves

The construction of the turbine profile using Bézier curves is based on the parametrization described in [13]. The turbine profile, which is shown in Fig. 2, is defined herein with three independent curves, two B-spline curves for the suction side (SS) and pressure side (PS) and a

circular arc at the TE to close the profile. In order to construct the profile, first a camber line is constructed, which is used to define the position of the control points of the SS and PS B-spline curves relative to the camber line. The camber line is defined by an inlet blade metal angle, an exit blade metal angle, an axial chord length and a stagger angle. The position of the leading edge and trailing edge is defined by the axial chord length and the stagger angle. By defining an inlet and outlet lines respecting the inlet and outlet blade metal angles, one can intersect the lines and define the point P_{mid} shown in Fig. 2.

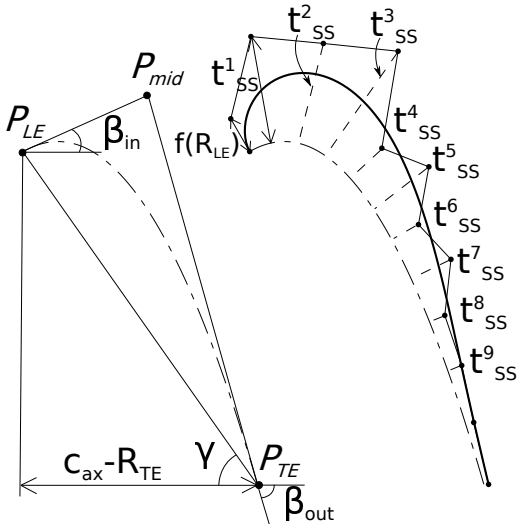


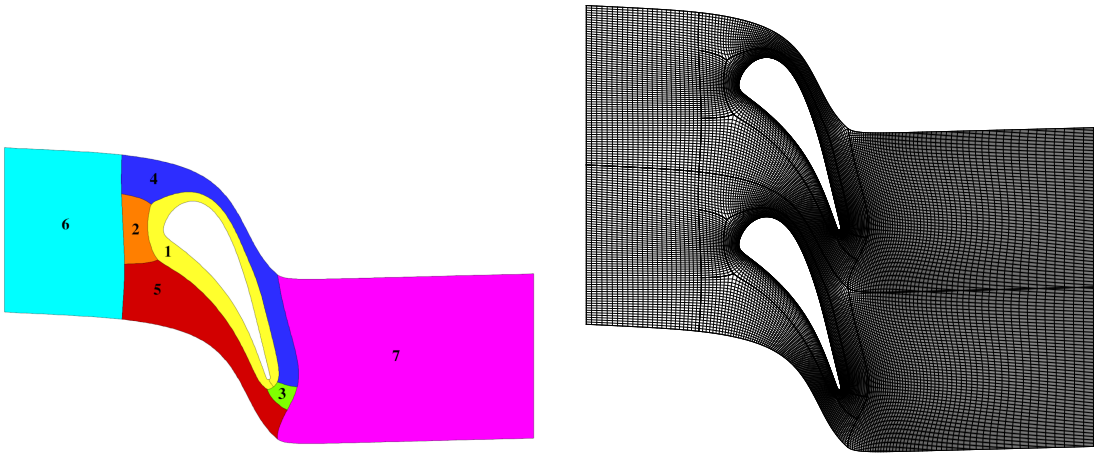
Figure 2: Construction of the suction side by a B-spline curve

The points P_{LE},P_{mid}, P_{TE} define the control points of the 2nd order Bézier curve describing the camber line. By specifying a stretch factor and a number of points, the camber line is divided into a number of intervals. For each point obtained on the camber line (except the first and last two points) a normal distance is specified by the designer to obtain the corresponding control point of the SS and PS B-spline curves. The normal distance of the first control point relative to the camber line is a function of the LE radius, in order to guarantee G2 geometric continuity (i.e. equal curvature) between the SS and PS B-splines at the stagnation point of the Leading Edge. The normal distance of the last control point relative to the camber line is equal to the TE radius. For the second last point, the normal distance is computed by specifying a wedge angle. Usually more control points are placed on the SS than on the PS because the highly curved shape of the SS is deemed to play a bigger role than the PS during the aerodynamic shape optimization. Finally, the shape of the cascade is fixed after specifying the pitch, which can be computed for a given solidity. The solidity plays an important role in turbomachinery design practice and this justifies the use of this parameter as an independent variable.

2.2 Generation of Block-structured smoothed grids

In the present work a structured grid was preferred as opposed to an unstructured grid to have better accuracy in the flow results. The generation of a boundary conformal curvilinear grid for the turbomachinery flow geometry of this study was achieved by subdividing the domain into multiple structured grid blocks which are independent from each other but share a common interface. A single block approach for a relatively highly staggered profile, such as the LS89

turbine profile, would result in a poor quality grid due to the high skewness of the cells. Therefore, a multi-block approach was preferred because it allows to have more detailed control of the element size and shape. Figure 3(a) shows the selected multi-block structured topology for the LS89 CFD domain. One O-grid block is placed around the profile and there are six additional H-grid blocks, four of them distributed around the profile and the remaining two being used as the inlet and outlet blocks. The first step in the grid generation process consist of generating



(a) LS89 multi-block structured grid topology (b) LS89 smoothed multi-block structured grid

Figure 3: LS89 mesh topology and final smoothed grid

an initial grid on the edges of the different blocks. The internal grid is subsequently initialized using Transfinite Interpolation (TFI) equations from the boundary mesh points, which are described by the Eqns. 1 and 2 [14]. Equations 1 and 2 allow to compute the internal grid given the coordinates of the boundaries for $\eta = 0, 1$ and $\xi = 0, 1$ (see Fig. 4).

$$\begin{aligned} X(\xi,\eta) = {} & (1-u)x(0,\eta) + ux(1,\eta) + (1-v)x(\xi,0) + vx(\xi,1) \\ & - (1-u)(1-v)x(0,0) - u(1-s)x(1,0) - (1-u)vx(0,1) - uvx(1,1) \end{aligned} \tag{1}$$

$$\begin{aligned} Y(\xi,\eta) = {} & (1-u)y(0,\eta) + uy(1,\eta) + (1-v)y(\xi,0) + vy(\xi,1) \\ & - (1-u)(1-v)y(0,0) - u(1-s)y(1,0) - (1-u)vy(0,1) - uvy(1,1) \end{aligned} \tag{2}$$

The parameters *u* and *v* from Eqns. 3 and 4 are blending formulas proposed by Soni [15] for the normalised arc-length control functions $s_1(\xi)$, $s_2(\xi)$, $t_1(\eta)$, $t_2(\eta)$ along the boundary edges. The normalised arc-length control functions intuitively express the position of each grid point along the edge as a percentage of the total length of the edge. In the present work, $s_1(\xi)$ and $s_2(\xi)$ are defined along the boundary edges spanning between $t_1(\eta = 0)$ and $t_2(\eta = 1)$ respectively, whereas $t_1(\eta)$ and $t_2(\eta)$ are defined along the boundary edges spanning between $s_1(\xi = 0)$ and $s_2(\xi = 1)$.

$$u(\xi,\eta) = \frac{(1 - t_1(\eta))s_1(\xi) + t_1(\eta)s_2(\xi)}{1 - (s_2(\xi) - s_1(\xi))(t_2(\eta) - t_1(\eta))} \tag{3}$$

$$v(\xi,\eta) = \frac{(1 - s_1(\xi))t_1(\eta) + s_1(\xi)t_2(\eta)}{1 - (t_2(\eta) - t_1(\eta))(s_2(\xi) - s_1(\xi))} \tag{4}$$

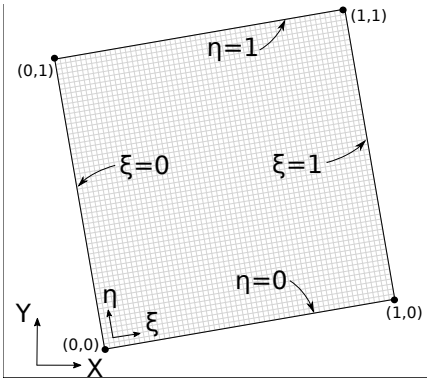


Figure 4: Two dimensional block

The generation of an algebraic grid for each block is efficient and easy to implement but there is no guarantee that the grid lines will not intersect, fold, or that the corners from the side will not be propagated inside the domain. Given the difficulty to generate successful smooth grids using TFI equations, it is convenient to smooth them afterwards by solving a partial differential equation (PDE) of some type (elliptic, hyperbolic, parabolic). The elliptic PDE is commonly used for grid generation for ducted flows. The Elliptic grid generation equations shown in Eqns. 5-6 were pioneered by Thompson et. al. [16] and are employed in this work.

$$\xi_{xx} + \xi_{yy} = P(\xi, \eta) \tag{5}$$

$$\eta_{xx} + \eta_{yy} = Q(\xi, \eta) \tag{6}$$

These equations, which are also referred to as the Poisson equations, are formulated in the body-fitted non-orthogonal physical space XY and need to be transformed to the uniform orthogonal computational space $\xi\eta$ in order to solve them efficiently. The inverse transformed system of equations then becomes [14]:

$$\alpha x_{\xi\xi} - 2\beta x_{\xi\eta} + \gamma x_{\eta\eta} \quad = -I^2(P x_\xi + Q x_\eta) \tag{7}$$

$$\alpha y_{\xi\xi} - 2\beta y_{\xi\eta} + \gamma y_{\eta\eta} \quad = -I^2(P y_\xi + Q y_\eta) \tag{8}$$

In Eqns. 7-8, the greek symbols α, β, γ represent the scale metric factors of the coordinate transformation and they are defined by $\alpha = x_\eta^2 + y_\eta^2$, $\beta = x_\xi x_\eta + y_\xi y_\eta$, $\gamma = x_\xi^2 + y_\xi^2$. The inverse of the Jacobian is given by $I = \frac{1}{J} = x_\xi y_\eta - x_\eta y_\xi$. The system of equations in the transformed space can be approximated using Finite Differences and then solved by means of a standard GS-(S)SOR (Gauss-Seidel Symmetric Successive Over-relaxation) approach.
The P and Q source terms are used to control the geometrical aspects of the cells. Grid behaviour control is achieved through the introduction of forcing function terms in the manner of Steger and Sorenson [17] in the O-grid block. The Steger and Sorenson method is an iterative forcing function approach that adjusts the boundary forcing functions to meet a prescribed angle and spacing constraint at the boundary. The forcing function is then interpolated into the domain based on the boundary values of the source terms. In all the H-grid blocks, the source terms have been set to zero and hence the elliptical equations being solved are the so called Laplace equations.

2.3 Algorithmic Differentiation

Algorithmic Differentiation (AD) is a technique for computing analytic derivatives of programs [18]. The idea of AD is to differentiate analytically each elementary mathematical operation performed by the code and accumulating the derivatives by applying the chain rule in an automatic fashion. In this way, it is possible to build up the cost function or derivative of the output with respect to the input variable, which can then be used for an optimization algorithm.

One distinguishes between the forward and reverse mode of AD. In forward mode one is interested to know the derivative of the outputs depending on one setting of the input sensitivities, whereas in reverse mode one is interested in the derivative of all inputs for one single output value. Both modes allows to accurately compute the sensitivities with the best possible machine accuracy. To compute the derivatives for all input parameters in the forward mode it is necessary to run as many simulations as the number of input parameters. Conversely, in the reverse mode you need as many runs as the number of output variables. AD methods can be implemented by either source to source transformation or operator overloading. The former one consists in transferring the primal evaluation trace into a differentiated evaluation trace. The second is based on operator and function overloading. For a more detailed description of the different techniques for evaluating derivatives using AD the reader is referred to [18].

In this work, all the sensitivities have been calculated using ADOL-C [19], an AD tool developed at the Department of Mathematics at the University of Paderborn that is written in C++ and uses overloaded operators and functions. The ADOL-C derivative evaluation routines can be called from C, C++, Fortran and any other language that can be linked with C [19]. Both the geometrical and grid sensitivities have been calculated in tapeless forward mode, which means that the operation count of the added lines necessary to evaluate the sensitivities is a small multiple of that for the underlying code to evaluate the output variable.

3 APPLICATION

The LS89 axial turbine aerofoil designed at the Von Karman Institute for Fluid Dynamics is selected to test the method presented in this paper as a first demonstrator. The LS89 is

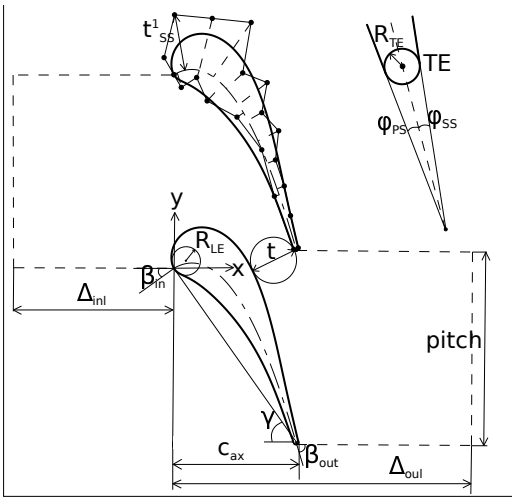


Figure 5: LS89 profile geometry

a relatively modern high pressure turbine nozzle guide vane (HPT NGV) profile specifically designed for subsonic flows downstream of an aero-engine combustion chamber.

Design parameters	Acronyms	Value
Solidity	σ	1.118
Axial Chord Length	c_{ax}	36.955 mm
Stagger Angle	γ	54.9°
LE Radius	R_{LE}	4.126 mm
TE Radius	R_{TE}	0.710 mm
TE Wedge Angle SS	φ_{SS}	4.0°
TE Wedge Angle PS	φ_{PS}	2.5°
Inlet Angle	β_{in}	0.0°
Outlet Angle	β_{out}	74.0°
Stretching Factors PS	k_{PS}	1.175
PS Thickness (x4)	t^1_{PS}, ..., t^4_{PS}	see Tab. 2
Stretching Factors SS	k_{SS}	1.1
SS Thickness (x9)	t^1_{SS}, ..., t^9_{SS}	see Tab. 2

Table 1: Design parameters used to define the LS89 blade profile

Bézier Control Point Index j	t^j_{SS} [mm]	t^j_{PS} [mm]
1	16.750	2.250
2	15.900	6.650
3	19.690	2.300
4	6.750	0.040
5	10.750	
6	4.750	
7	6.850	
8	2.565	
9	2.295	

Table 2: Bézier control point normal distances relative to the camber line

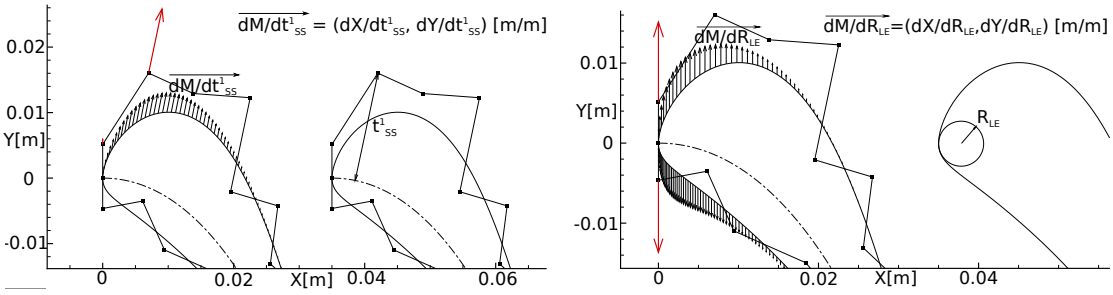
The profile was designed and optimized in the two-dimensional space at the VKI for a 0.9 downstream isentropic Mach number by an inverse method described in [9]. It was tested at the VKI laboratories with the aim to use the experimental data as a benchmark to validate the VKI CFD prediction capabilities in the late 80s. The complete experimental results were published during the 1990 International Gas Turbine Conference held in Brussels [10]. The final report is the VKI Technical Note TN174 [11]. The LS89 has been tested for different Reynolds and Mach numbers in the VKI-CT2 test facility. A description of different set ups for optimization studies can be found in [12].
The LS89 profile geometry shown in Fig. 5 comes first from point data [11] that has been fitted with the parametrization method described before. The geometry is defined using 24 geometrical parameters shown in Tab. 1 and 2. The position of the Bézier control points of the suction side and pressure side is shown in Fig. 5.

4 RESULTS AND DISCUSSION

The geometrical and the grid sensitivities are obtained for all the design parameters, showing the influence that each design parameter has on the aerofoil surface X-Y coordinates and the CFD grid X-Y coordinates respectively. The discussion herein is limited to the parameters t^1_{SS}, R_{LE}, c_{ax} (see Tab. 1) for the sake of simplicity. In the following subsections, the AD geometrical sensitivities, which are defined as the change in the X-Y turbine profile coordinates with respect to a design parameter change, will be discussed first. The grid sensitivities, which are defined as the change in the X-Y grid coordinates with respect to a design parameter change will be discussed next. Finally, the sensitivities obtained by AD will be compared against FD.

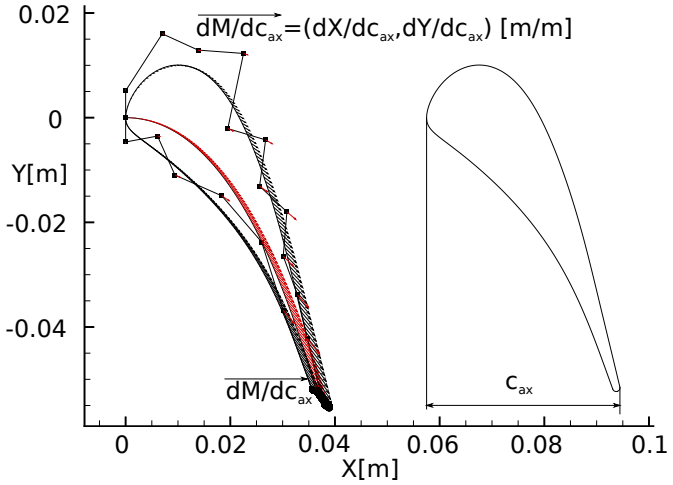
4.1 AD geometrical sensitivities

Figures 6(a), 6(b) and 6(c) show the magnitude and direction of the geometrical sensitivities obtained for the t^1_{SS}, R_{LE}, c_{ax} design parameters, respectively. Different scale factors are used to display the length of the vectors. Figure 6(a) clearly shows that an infinitesimally small



(a) t^1_{SS}, Vector Scale Factor = 0.01

(b) R_{LE}, Vector Scale Factor = 0.0016



(c) c_{ax}, Vector Scale Factor = 0.0025

Figure 6: AD geometrical sensitivities

perturbation of t^1_{SS} causes the movement of the points along the suction side shoulder in the normal direction in which the suction side first Bézier control point position is specified relative to the camber line. A similar perturbation in the R_{LE} forces the first control point of the suction and pressure side to move away from the camber line because their position is a function of the LE radius, making the points along the LE region to move in the y direction accordingly

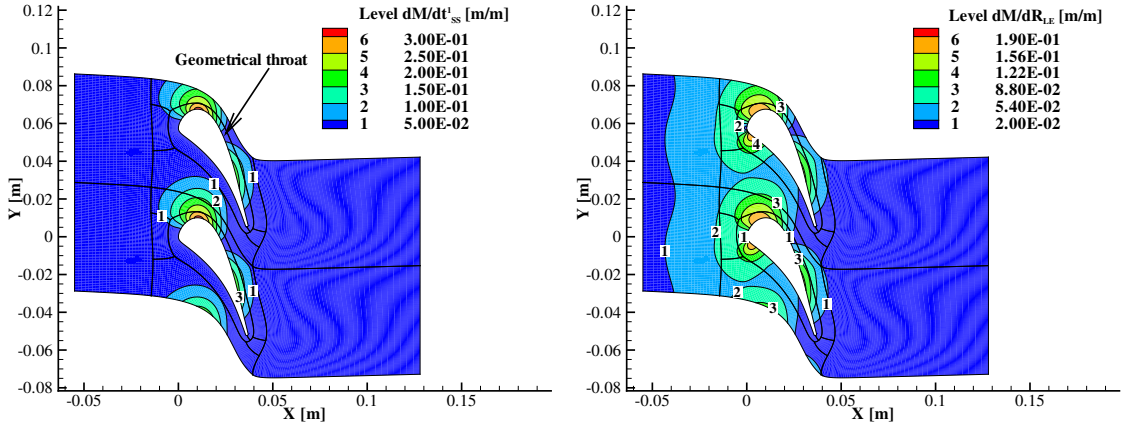
as displayed in Fig. 6(b). Figure 6(c) shows that a perturbation in the c_{ax} causes the points along the camber line, SS and PS curves to move away from the stagnation point at the LE. Although the scale factor does not allow to show the magnitude of all the sensitivities, all the control points (except the first of the SS and PS curves) sensitivity vectors point away from the stagnation point because the TE point (P_{TE} from Fig. 2) also moves away in both x and y directions. The TE point movement is not in the x direction only because an increase of axial chord by virtue of a constant stagger angle requires the TE point (and the camber line) to move in the (x,y) space.

4.2 AD grid sensitivities

Figures 7(a), 7(b) and 7(c) show the magnitude of the grid sensitivities obtained for the t^1_{SS}, R_{LE}, c_{ax} design parameters, respectively. In general, a perturbation in the design parameter can cause two type of perturbations in the CFD grid point coordinates. First, the grid point distribution along the edge of the curve itself can change due to the fact that the normalised arc-length functions used to distribute points in the edge undergo some changes. This type of perturbation that moves the surface grid points tangentially to the surface can also propagate into the interior domain of the grid but it has a tendency to decay rapidly. Secondly, the grid points coordinates in the fluid domain change as well according to the overall grid point field movement sensed by them. The source of this perturbation is stronger than for the first type and originates in the region where the geometrical sensitivities are not zero. Figure 7(a) clearly shows that a small perturbation in the t^1_{SS} causes the grid points in the fluid domain above the SS shoulder to move and this movement also propagates into the PS fluid domain through the periodic interface. The rear suction side grid points are also affected due to the fact that the arc-length distribution changes. The geometrical throat is used herein to split the SS curve into two edges to have more control of the grid point distribution along the SS curve. This means that the geometrical throat on the SS (see Fig. 7(a)) is the end of the SS pre-throat and the beginning of the SS post-throat edges, which are using different arc-length distributions. Therefore, it is expected that a perturbation in SSThck1 causes a grid point position variation along the SS pre-throat and post-throat edges because the length of these edges changes. However, as the geometrical throat X-Y coordinates remain fixed, the grid points in the vicinity of the SS geometrical throat do not sense any perturbation.

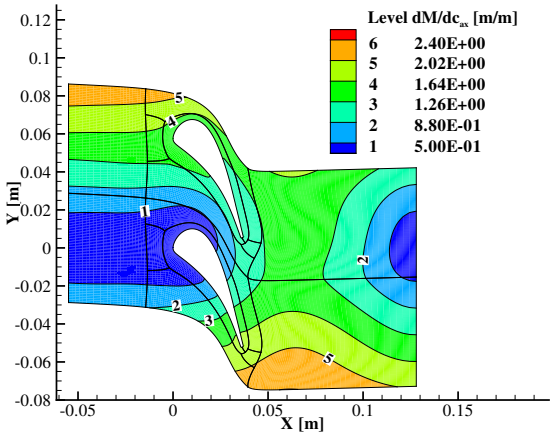
Figure 7(b) shows that the biggest influence of R_{LE} is in the vicinity of the largest geometrical sensitivities (see Fig. 6(b)). The perturbation on the SS also propagates into the PS interior domain through the periodic interface. Also, the changes in the normalised arc-length functions for the PS and SS edges result into grid point movements in both PS and SS curves. As the SS is more curved than the PS, it is expected that a perturbation in the R_{LE} changes the total length of the SS curve more than for the PS. This is the reason why the R_{LE} grid sensitivities at the rear SS are bigger than the rear PS. The grid point displacements along the SS and PS also propagate into the interior of the fluid domain but decay more rapidly than in those that propagate from the LE region. The minimum sensitivity areas around the profile correspond to those grid points that are fixed, like the stagnation point, the geometrical throat position, and the SS and PS end points towards the TE. In these points, the arc-length functions do not change because they take the values zero or one, depending the orientation in which the grid point distribution has been applied to the edges.

Figure 7(c) shows a non-periodic character, for the displacement field is not repeated passage after passage. This is because the solidity, as an independent parameter, remains constant under a change of the axial chord length, resulting is a proportional change in pitch. Identical points



(a) t^1_{SS} AD grid sensitivities

(b) R_{LE} grid sensitivities



(c) c_{ax} grid sensitivities

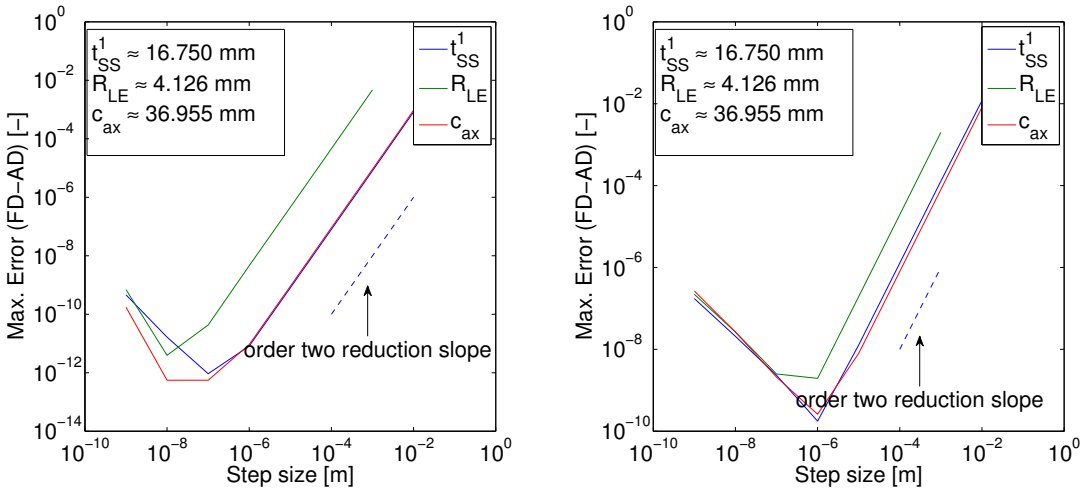
Figure 7: AD grid sensitivities

on the opposite side from the periodic boundary will experience a jump in the y-component of the grid sensitivity, proportional to the change in pitch. In the figure, the jump value has been added to the top CFD domain such that a continuous field is visualized. The choice to use solidity as an independent parameter is inspired by it's importance in classical turbomachinery design practice.

4.3 AD vs FD comparison

The geometrical and grid sensitivities obtained by AD are compared against FD second order central derivatives by tuning the step size. Figures 8(a) and 8(b) show the maximum error in magnitude between the FD approximated and the exact AD sensitivities for the t^1_{SS}, R_{LE} and c_{ax} design parameters. The error is very sensitive to the chosen step size and the step size needs to be relatively small to have a small error. Figures 8(a) and 8(b) show that the error in both cases reduces with order two as the step size is reduced up to a critical step size, as to be expected for a central second order FD scheme. If the step size is chosen too small below this critical step size, the numerical representation of the FD approximation becomes unstable and more error-prone due to the limited machine accuracy. The error increases with order one approximately if the step size is reduced below the critical step size. Figures 9(a), 9(b) and 9(c) show the grid sensitivity error for the t^1_{SS}, R_{LE} and c_{ax} design parameters if the step size

is 1E-03 [m]. It has to be noted that the maximum grid sensitivity error shown in Fig. 9(b) is relatively large, as it is only two orders of magnitude lower than the maximum grid sensitivities shown in Fig. 7(b), and will eventually result in wrong gradient information.



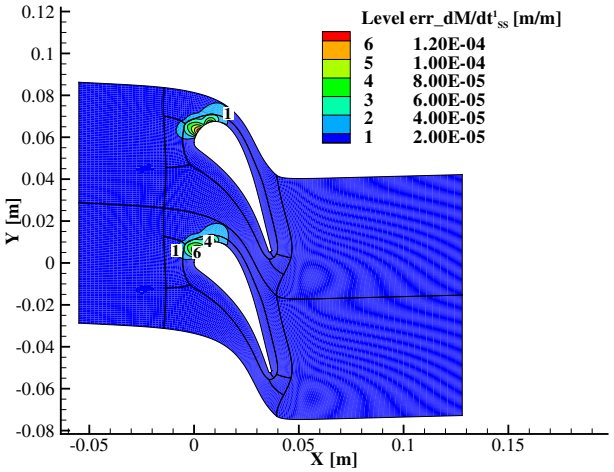
(a) Geometrical sensitivity error

(b) Grid sensitivity error

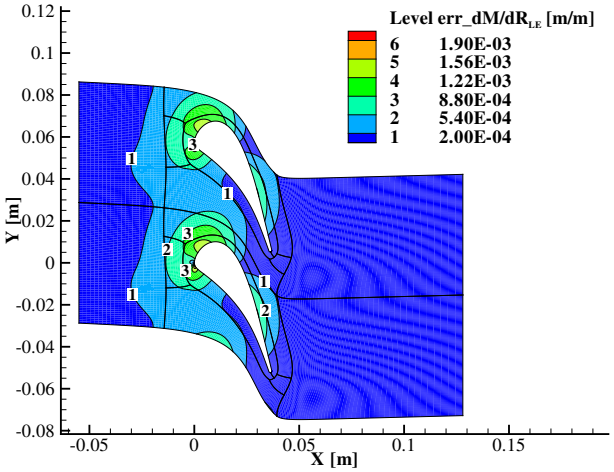
Figure 8: Max Error between the FD and AD

5 CONCLUSIONS

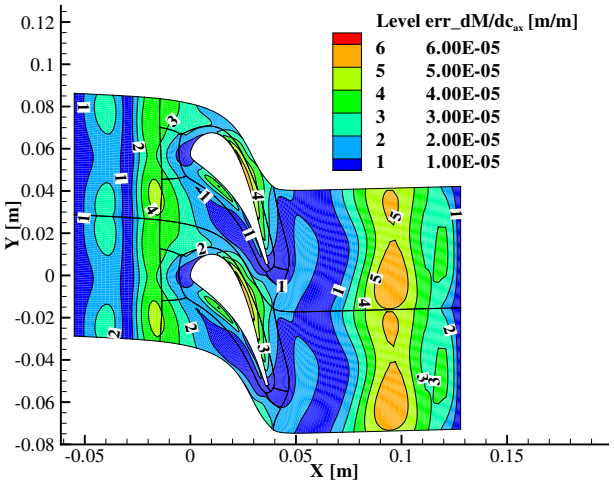
The method presented herein allows to integrate the CAD kernel and grid generation tools within the CFD adjoint-based optimization algorithm by expressing the optimization problem through CAD parameters, rather than working directly on the CFD grid. This allows the optimal shape to remain defined within the CAD tool and to differentiate the CAD kernel and grid generation tool with algorithmic differentiation (AD) to obtain the grid sensitivities propagated to the CAD-based parameters. The in-house CAD and grid generation tools have been differentiated in forward mode and tested in the VKI LS89 axial high pressure turbine nozzle guide vane. The two dimensional profile, constructed by means of Bézier and B-spline curves, is parametrized by 24 design parameters and the geometrical as well as grid sensitivities are obtained for each, showing the influence that each design parameter has on the profile and on the CFD grid coordinates. Finally, the sensitivities obtained by AD are compared against finite differences second order central derivatives by tuning the step size. The error introduced by FD is very sensitive to the chosen step size and its magnitude can be relatively close to the sensitivity value itself if the step size is not chosen appropriately. The algorithmic differentiation of the cad kernel and grid generation tool is a promising approach to obtain accurate sensitivities with the best machine accuracy. Although in the present approach the AD sensitivities are obtained in forward mode mainly, algorithmic differentiation also offers the reverse capability. Further work will focus in differentiating the CAD kernel and grid generation tool in reverse mode. The sensitivities computed in forward mode will serve to validate those obtained by the reverse mode. The differentiated code will be integrated afterwards with the CFD solver and the gradient-based optimization algorithm. Finally, the method will be extended to more complicated 3D turbomachinery test cases.



(a) t^1_{SS} grid sensitivity error, step size = 1E-03



(b) R_{LE} grid sensitivity error, step size = 1E-03



(c) AxCh grid sensitivity error, step size = 1E-03

Figure 9: Grid sensitivity error for step size = 1E-03

REFERENCES

[1] Jameson, A., "Aerodynamic design via control theory," *Journal of scientific computing*, Vol. 3, No. 3, 1988, pp. 233–260.

[2] Pironneau, O., "On optimum design in fluid mechanics," *Journal of Fluid Mechanics*, Vol. 64, No. 01, 1974, pp. 97–110.

[3] Giles, M. B. and Pierce, N. A., "An introduction to the adjoint approach to design," *Flow, turbulence and combustion*, Vol. 65, No. 3-4, 2000, pp. 393–415.

[4] Peter, J. E. and Dwight, R. P., "Numerical sensitivity analysis for aerodynamic optimization: A survey of approaches," *Computers & Fluids*, Vol. 39, No. 3, 2010, pp. 373–391.

[5] Jameson, A., "Efficient aerodynamic shape optimization," *AIAA paper*, Vol. 4369, 2004, pp. 2004.

[6] Becker, G., Schäfer, M., and Jameson, A., "An advanced NURBS fitting procedure for post-processing of grid-based shape optimizations," *49th AIAA Aerospace Sciences Meeting, Orlando, FL, USA*, 2011.

[7] Leal, N., Leal, E., and Branch, J. W., "Simple method for constructing nurbs surfaces from unorganized points," *Proceedings of the 19th international meshing roundtable*, Springer, 2010, pp. 161–175.

[8] Verstraete, T., "Cado: A computer aided design and optimization tool for turbomachinery applications," *2nd Int. Conf. on Engineering Optimization, Lisbon, Portugal, September*, 2010, pp. 6–9.

[9] Van den Braembussche, R., Léonard, O., and Nekmouche, L., "Subsonic and transonic blade design by means of analysis codes," *Computational Methods for Aerodynamic Design (Inverse) and Optimization, AGARD CP*, Vol. 463, 1990.

[10] Arts, T. and De Rouvroit, M. L., "Aero-Thermal Performance of a Two Dimensional Highly Loaded Transonic Turbine Nozzle Guide Vane: A Test Case for Inviscid and Viscous Flow Computations," *ASME 1990 International Gas Turbine and Aeroengine Congress and Exposition*, American Society of Mechanical Engineers, 1990, pp. V001T01A106–V001T01A106.

[11] Arts, T., Lambertderouvroit, M., and Rutherford, A., "Aero-thermal investigation of a highly loaded transonic linear turbine guide vane cascade. A test case for inviscid and viscous flow computations," *NASA STI/Recon Technical Report N*, Vol. 91, 1990, pp. 23437.

[12] Mueller, L. and Verstraete, T., *The LS89 Optimization Test Case*, 2016, Available at `http://aboutflow.sems.qmul.ac.uk/content/events/munich2016/benchmark/testcase2/files/LS89TestCase.pdf`, url date 2016-02-03.

[13] Pierret, S., *Designing turbomachinery blades by means of the function approximation concept based on artificial neural network, genetic algorithm, and the navier-stokes equations*, Ph.D. thesis, Von Karman Institute for Fluid Dynamics, 12 1999.

[14] Thompson, J. F., Soni, B. K., and Weatherill, N. P., *Handbook of grid generation*, CRC press, 1998.

[15] SONI, B., "Two-and three-dimensional grid generation for internal flow applications of computational fluid dynamics," *7th Computational Physics Conference*, 1985, p. 1526.

[16] Thompson, J. F., Thames, F. C., and Mastin, C. W., "Automatic numerical generation of body-fitted curvilinear coordinate system for field containing any number of arbitrary two-dimensional bodies," *Journal of computational physics*, Vol. 15, No. 3, 1974, pp. 299–319.

[17] Steger, J. and Sorenson, R., "Automatic mesh-point clustering near a boundary in grid generation with elliptic partial differential equations," *Journal of computational physics*, Vol. 33, No. 3, 1979, pp. 405–410.

[18] Griewank, A. and Walther, A., *Evaluating derivatives: principles and techniques of algorithmic differentiation*, Siam, 2008.

[19] Walther, A. and Griewank, A., "A Package for the Automatic Differentiation of Algorithms Written in C/C+," *URL https://projects. coin-or. org/ADOL-C*, 2014.

[20] Piegl, L. and Tiller, W., *The NURBS book*, Springer Science & Business Media, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ALTERNATIVE SOLUTION ALGORITHMS FOR PRIMAL AND ADJOINT INCOMPRESSIBLE NAVIER-STOKES

Mattia Oriani[1,2], **Guillaume Pierrot**[1]

[1]ESI Group
99 Rue des Solets, Rungis 94513, France
{mattia.oriani,guillaume.pierrot}@esi-group.com

[2] Queen Mary, University of London
Mile End Road, London E1 4NS, UK
m.oriani@qmul.ac.uk

Keywords: Oseen Preconditioners, Adjoint Navier-Stokes, Block-Coupled, SIMPLEC, V-Coupled, Augmented Lagrangian.

Abstract. *Regardless of the specific discretisation framework, the discrete incompressible Navier-Stokes equations present themselves in the form of a non-linear, saddle-point Oseen-type system. Traditional CFD codes typically solve the system via the well-known SIMPLE-like algorithms, which are essentially block preconditioners based on Schur complement theory. Due to their "segregated" nature, which reduces to iteratively solving a sequence of linear systems smaller than the full Oseen and better conditioned, traditional SIMPLE-like algorithms have long been considered as the only viable strategy.*

However, recent progress in computational power and linear solver capabilities has led researchers to develop, for Oseen-type systems (and discrete Navier-Stokes in particular), a number of alternative preconditioners and solution schemes, found to be more efficient than SIMPLE-like strategies but previously deemed practically unfeasible in industrial contexts.

The improved efficiency of novel preconditioners entails a) faster, more stable convergence and b) the possibility of driving residuals below more strict tolerances, which is sometimes difficult with SIMPLE due to stagnating behaviour. The second aspect in particular is extremely relevant in the context of adjoint-based optimisation, as evidence suggests that an adjoint system may be affected by convergence issues when the primal flow solution is not well converged.

In this work, we present some solution schemes (both traditional and novel) implemented for the Mixed Hybrid Finite Volumes Navier-Stokes solver we introduced in our previous work. Performance, in terms of robustness and convergence properties, is assessed on a series of benchmark test cases. We also turn our attention to the discrete adjoint Navier-Stokes problem itself, which in essence requires solving a linear system similar to the original Oseen and therefore may benefit from the same preconditioning techniques. We show how the primal algorithms are adapted to the adjoint system, and we run a series of adjoint test cases to compare performance of various solution schemes.

1 INTRODUCTION

In numerical optimisation, gradient-based methods are defined as a way of actively searching for a local minimum of a given cost function J via minimisation algorithms that make use of its sensitivity, i.e. its gradient with respect to a chosen set of design parameters α. Typical examples related to fluid mechanics include: reducing the drag force exerted on an aircraft wing by seeking the optimal airfoil profile; minimising the total pressure loss of a flow through a duct by modifying the shape of the duct's cross section.

Performing numerical optimisation in an industrial context can be prohibitively costly for two reasons. Firstly, each iteration of the optimisation loop requires an evaluation of J, and therefore a CFD solve. Secondly, extra computations are required in order to get the sensitivity $\frac{dJ}{d\alpha}$. The second issue in particular is a blocking factor in an industrial context: gradient computation via traditional methods - finite differencing (FD) or forward-mode Algorithmic Differentiation (AD) [11] - demands as many extra CFD solves as the number of design variables n_α, which may be in the order of millions. This is notably the case for shape optimisation processes where we want to compute the *surface sensitivity*, i.e. the gradient of J w.r.t. the coordinates of each mesh node lying on the surface to be optimised.

The *adjoint method* [12, 13, 22] provides a workaround to the problem, as it allows to compute the sensitivity of J at a cost that is essentially independent of n_α. Our work is based on the *discrete adjoint* approach, which consists in applying adjoint theory directly to the set of discrete equations describing the original (or *primal*) problem. In short, a discrete adjoint boils down to solving a linear system in the form:

$$\left(\frac{\partial \mathbf{r}}{\partial \mathbf{u}}\right)^T \mathbf{u}^* = -\frac{\partial J}{\partial \mathbf{u}} \tag{1}$$

where $\mathbf{r}$ is the primal residual vector and $\mathbf{u}^*$ the discrete *adjoint field*.

The adjoint method, powerful as it is, suffers from stability and convergence issues. Solving linear system (1) is not always a trivial task, and it was shown [20] that, if the primal solution is not properly converged, then the adjoint might not converge at all. Evidence suggests that such lack of robustness is to be blamed in part on the quality of the CFD solver itself, which in turns depends on two key factors:

- the level of consistency and robustness of the discrete operators used in the primal;
- the efficiency of the algorithm used to converge the primal flow field.

In other words, even though a conventional CFD solver is able to produce results that are "good enough" for mere engineering purposes, its discrete adjoint counterpart may be unstable or too inaccurate to be of any use, since its quality is highly sensitive to how the primal governing equations were discretised, as well as the tolerance down to which they were solved.

In our previous work [17, 18, 19] we addressed the task of finding robust discretisation schemes for the Navier-Stokes equations, which we summarise in the next section. In this paper we will turn our attention to developing improved solution algorithms for both primal and adjoint systems.

1.1 Previous work: discretisation schemes

Traditional CFD solvers often rely on classical Finite Volumes (FV) schemes to discretise the incompressible Navier-Stokes equations, which seem to produce discrete adjoints that some-

what lack robustness. In our recent work we attempted to tackle such issue, focusing our research on non-standard discretisation schemes. We looked in particular into a class of methods formerly known as Mimetic Finite Differences (MFD), now renamed Virtual Elements Methods (VE) [5], originally developed for pure anisotropic diffusion problems [3, 4] and subsequently extended to convection-diffusion operators [7, 26] and 1st-order accurate Navier-Stokes [8]. Compared to FV, VE operators sport a series of attractive features, most notably:

- they "mimic" at a discrete level certain key properties of their continuous counterparts (e.g. they satisfy the discrete Gauss-Green formula);
- they are perfectly consistent up to a set order of accuracy, which can be increased to an arbitrary level by attaching extra degrees of freedom to each mesh element;
- they are fully implicit and free of numerical artefacts (e.g. the Non-Orthogonal Correctors typically found in FV schemes);
- their requirements in terms of mesh quality are minimal, i.e. they can deal with highly skewed, non-orthogonal or non-convex elements.

We refer the reader to our previous publications for details on our specific implementation of a VE diffusion operator [17], the addition of a convective term [18], the extension to stabilised 2nd-order accuracy and our implementation of a Mixed Hybrid Finite Volumes (MHFV) Navier-Stokes solver [19]. Here we limit ourselves to providing a very simplified outline of our framework.

The main degrees of freedom for a generic mesh cell (Figure 1) are:

- face-averaged velocities $\vec{U}_F$, with components u_F, v_F, w_F;
- cell-averaged pressure p_C.

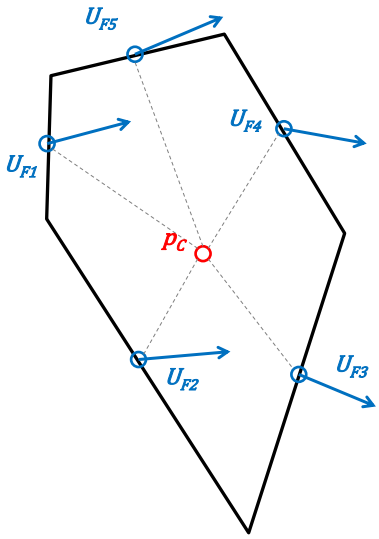


Figure 1: Placement of main MHFV flow variables on a generic 2D cell.

Considering now the incompressible, steady-state Navier-Stokes equations

$$\begin{cases} \vec{U} \cdot \nabla \vec{U} - \nabla \cdot \nu \nabla \vec{U} = -\nabla p + \vec{g} \\ \nabla \cdot \vec{U} = 0 \end{cases}, \tag{2}$$

when discretised via our MHFV operators they yield an Oseen-type saddle-point system

$$\begin{bmatrix} \mathcal{F}_{\Phi,x} & 0 & 0 & \mathcal{G}_x \\ 0 & \mathcal{F}_{\Phi,y} & 0 & \mathcal{G}_y \\ 0 & 0 & \mathcal{F}_{\Phi,z} & \mathcal{G}_z \\ \mathcal{D}_x & \mathcal{D}_y & \mathcal{D}_z & 0 \end{bmatrix} \begin{pmatrix} \mathbf{u}_F \\ \mathbf{v}_F \\ \mathbf{w}_F \\ \mathbf{p}_C \end{pmatrix} = \begin{pmatrix} \mathbf{g}_{F,u} \\ \mathbf{g}_{F,v} \\ \mathbf{g}_{F,w} \\ \mathbf{0} \end{pmatrix} \tag{3}$$

or, in a more compact form:

$$\begin{bmatrix} \mathcal{F}_\Phi & \mathcal{G} \\ \mathcal{D} & 0 \end{bmatrix} \begin{pmatrix} \mathbf{U}_F \\ \mathbf{p}_C \end{pmatrix} = \begin{pmatrix} \mathbf{g}_F \\ \mathbf{0} \end{pmatrix}. \tag{4}$$

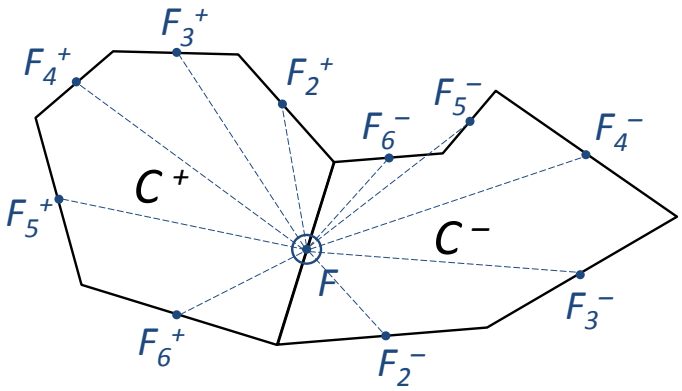


Figure 2: Face-to-face stencil on a generic 2D mesh.

- $\mathcal{F}_{\Phi,i}$ stands for any version of the MHFV hybrid convection-diffusion operator described in [19], acting on the ith velocity component. It is identical in all 3 spatial directions (boundary conditions aside), and it is of course non-linear, since it depends on the convective flux Φ_F which is a function of $\vec{U}_F$. However, when linearised w.r.t. Φ_F, $\mathcal{F}_\Phi$ is evidently block-diagonal, i.e. it only acts on the velocity component in its own direction.

  The operator is derived by defining a *face flux* Ψ_{FC}, through face F outward w.r.t. cell C, catering for both convective and diffusive fluxes and acting on its respective component of $\vec{U}_{F'}$ defined on all faces F' belonging to C. For instance, for u_F:

$$\Psi^u_{FC} = \sum_{F' \in C} n^C_{FF'} u_{F'}. \tag{5}$$

  Derivation of the $n^C_{FF'}$ coefficients, omitted here, is the core of the VE/MHFV scheme. We then impose flux conservation across each face, i.e.

$$\Psi^u_{FC^+} + \Psi^u_{FC^-} = 0 \tag{6}$$

where C^+ and C^- denote the two cells connected by face F, which yields the above mentioned hybrid convection-diffusion operator. $\mathcal{F}_\Phi$ therefore operates on a face-to-face stencil, as the one shown in Figure 2;

- the divergence operator $\mathcal{D}$ is simply derived from the Gauss divergence theorem applied over each cell, i.e.

$$\mathcal{D}_C\left(\vec{U}_F\right) = \sum_{F\in\partial C} \vec{U}_F \cdot \vec{F} \tag{7}$$

where $\vec{F} = \{F_x, F_y, F_z\}$ is the area vector of face F, outward w.r.t. cell C;

- the gradient operator $\mathcal{G}$ acting on the pressure is, for a 1st-order scheme, the transpose of the divergence operator, i.e.

$$\mathcal{G}_F\left(p_C\right) = \vec{F}\left(p_{C^-} - p_{C^+}\right). \tag{8}$$

For a 2nd-order scheme, the operator also includes least-squares approximations of pressure face values:

$$\mathcal{G}_F\left(p_C\right) = \vec{F}\left\{\left[p_{C^-} + \nabla^{LSQ}_{C^-} p \cdot (\vec{x}_F - \vec{x}_{C^-})\right] - \left[p_{C^+} + \nabla^{LSQ}_{C^+} p \cdot (\vec{x}_F - \vec{x}_{C^+})\right]\right\}. \tag{9}$$

1.2 MHFV discrete adjoint

Let us now turn our attention to the discrete adjoint Navier-Stokes problem. In order to assemble the adjoint system (1) we make use of the Equational Differentiation (ED) methodology which we describe in [21]. In a nutshell, ED stands as a clear formulation of the fact that we differentiate the primal (4) itself, and not the solution algorithm used to solve it (which is what AD does). In other words, ED requires assembling (1) explicitly and independently of the primal solving algorithm. Assembly of the Jacobian matrix $\frac{\partial \mathbf{r}}{\partial \mathbf{u}}$ and the adjoint right-hand side $-\frac{\partial J}{\partial \mathbf{u}}$ may be done via any viable method, such as hand-differentiation, FD or AD tools. In our case, we make use of FD techniques combined with colouring algorithms in order to make the assembly cost independent of the problem size. One of the main advantages of such approach is that it allows to assemble the system in an automatic, non-intrusive and black-box-like fashion. However, since one of our goals here is to devise preconditioning techniques for the adjoint system, it is paramount that we analyse its structure.

Let us consider the Jacobian $\frac{\partial \mathbf{r}}{\partial \mathbf{u}}$ of the MHFV Navier-Stokes problem, i.e. the matrix of partial derivatives of the residual vector of system (4) with respect to each flow variable. The gradient and divergence operators, $\mathcal{G}$ and $\mathcal{D}$, are linear with respect to all variables, and therefore their corresponding entries in the Jacobian correspond to the operators themselves. The convection-diffusion operator $\mathcal{F}_\Phi$, however, is non-linear with respect to $\mathbf{U}_F$: this is mainly due to its dependency on the convective flux Φ_F, although further non-linearities may be also be present if the operator includes certain solution-dependent discretisation schemes (flux limiters, weighting coefficients...), which is notably the case for our Upwind Least Squares (ULSQR) stabilisation technique, often used in our tests. The full Jacobian $\mathcal{A}$ takes the form:

$$\mathcal{A} = \begin{bmatrix} \widetilde{\mathcal{F}_{xx}} & \widetilde{\mathcal{F}_{xy}} & \widetilde{\mathcal{F}_{xz}} & \mathcal{G}_x \\ \widetilde{\mathcal{F}_{yx}} & \widetilde{\mathcal{F}_{yy}} & \widetilde{\mathcal{F}_{yz}} & \mathcal{G}_y \\ \widetilde{\mathcal{F}_{zx}} & \widetilde{\mathcal{F}_{zy}} & \widetilde{\mathcal{F}_{zz}} & \mathcal{G}_z \\ \mathcal{D}_x & \mathcal{D}_y & \mathcal{D}_z & 0 \end{bmatrix} \tag{10}$$

or, in compact notation:

$$\mathcal{A} = \begin{bmatrix} \widetilde{\mathcal{F}} & \mathcal{G} \\ \mathcal{D} & 0 \end{bmatrix}. \tag{11}$$

Unlike its primal counterpart, the adjoint momentum block $\widetilde{\mathcal{F}}$ is not block-diagonal, due to the fact that convective fluxes Φ_F depend on all components of $\vec{U}_F$, and therefore so does the momentum residual, regardless of its specific direction. Each entry $\widetilde{\mathcal{F}_{ij}}$ in (10) corresponds to the partial derivative of the momentum residual in the ith direction with respect to the jth velocity component. Defining now $\mathbf{g}_F^* = -\frac{\partial J}{\partial \mathbf{U}_F}$ and $\mathbf{g}_C^* = -\frac{\partial J}{\partial \mathbf{p}_C}$, J being the cost function, we can assemble the full adjoint system:

$$\begin{bmatrix} \widetilde{\mathcal{F}}^T & \mathcal{D}^T \\ \mathcal{G}^T & 0 \end{bmatrix} \begin{pmatrix} \mathbf{U}_F^* \\ \mathbf{p}_C^* \end{pmatrix} = \begin{pmatrix} \mathbf{g}_F^* \\ \mathbf{g}_C^* \end{pmatrix}, \tag{12}$$

where $\mathbf{U}_F^*$ and $\mathbf{p}_C^*$ are the adjoint velocity and pressure fields, respectively. The system is evidently an Oseen-type saddle-point problem, much like the primal it is derived from.

2 PRECONDITIONING OF OSEEN-TYPE PROBLEMS

In the past few decades, constant increases in computing power have driven the advances in algorithm development, specifically towards block preconditioners for Oseen-type systems more efficient than the traditional SIMPLE-like strategies. Research has successfully produced a number of alternative algorithms, although mostly restricted so far to the FE community: variants of classical ILU preconditioners [28]; those based on the so-called *approximate commutators* [9] - in particular the *Pressure Convection-Diffusion* (PCD) commutator [15]; the *Augmented Lagrangian* approach [1]. Several interesting comparisons amongst various Navier-Stokes preconditioners have also been published [23, 24].

We showed above how our discrete Navier-Stokes problem, both primal and adjoint, indeed takes the form of a saddle-point Oseen-type problem. The main difference is that the primal is non-linear due to the contributions of convective terms to $\mathcal{F}_\Phi$, which depend on $\mathbf{U}_F$, and therefore requires an outer iterative procedure (*Picard iteration*) on top of the inner one that solves the linearised Oseen problem itself. However, since there is no interest in obtaining the exact Oseen solution at each Picard iteration, the two are typically performed at the same time in a one-shot fashion: this gives rise to the well-known CFD solution algorithms, which are in fact the mere combination of preconditioners for linearised Oseen-type systems with outer non-linear iterations. The adjoint is simply a linear saddle-point problem; it is therefore perfectly legitimate to devise generic algorithms that can be adapted to both primal and adjoint, provided that a few subtle differences are taken care of. For this reason, in the following sections we will first describe preconditioners for a generic Oseen system written in the following notation:

$$\begin{bmatrix} \mathbb{F} & \mathbb{G} \\ \mathbb{D} & 0 \end{bmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_u \\ \mathbf{g}_p \end{pmatrix}, \tag{13}$$

then adapt them to the MHFV framework and highlight the differences between primal and adjoint.

Notice that we set the lower right matrix block in (13) to zero, which is the case for our Navier-Stokes scheme (4) and for its adjoint (12). This is typically true for FE schemes satisfying the so-called inf-sup stability condition, but false in general - most notably for classical FV using Rhie-Chow interpolation, as well as for many common stabilised FE schemes.

In the following sections we cover some of the algorithms we implemented so far, which are mostly MHFV adaptations of preconditioners developed for traditional FV or FE schemes. It is worth mentioning that, although our main goals are a) to achieve a better converged primal in order to produce a more robust discrete adjoint and b) to devise efficient ways of solving the adjoint itself, the potential benefits of our investigation go beyond the scope of adjoint-based optimisation, since the ability to solve the discrete Navier-Stokes to higher accuracy and in as few iterations as possible is highly desirable in the industry, regardless of adjoint computation.

2.1 SIMPLEC

Classical FV often make use of the ever-popular SIMPLE-like solution algorithms (see e.g. [25]). The efficiency of SIMPLE-like preconditioners is debatable at best: they are somewhat stable but they exhibit a rather poor convergence rate, they are prone to stagnation and their performance is affected by mesh refinement. They mostly owe their popularity to legacy reasons, being amongst the first devised working methods, and to their segregated nature, as they require solving linear systems that are relatively small and better conditioned in comparison to the full Oseen.

Despite traditionally being presented as *segregated algorithms*, highlighting the fact that they solve separately for velocity and pressure, SIMPLE-like strategies can in fact be seen as a way of preconditioning the discrete Oseen problem [24]. Notice that the generic Oseen matrix in (13) can be factorized as

$$\begin{bmatrix} \mathbb{F} & \mathbb{G} \\ \mathbb{D} & 0 \end{bmatrix} = \begin{bmatrix} \mathbb{F} & 0 \\ \mathbb{D} & -\mathbb{S} \end{bmatrix} \begin{bmatrix} \mathbb{I} & \mathbb{F}^{-1}\mathbb{G} \\ 0 & \mathbb{I} \end{bmatrix}, \tag{14}$$

where $\mathbb{I}$ is the identity matrix and $\mathbb{S}$ is known as *Schur complement*:

$$\mathbb{S} = \mathbb{D}\mathbb{F}^{-1}\mathbb{G}. \tag{15}$$

This suggests a potentially very efficient way of preconditioning the Oseen system; in fact, an exact Schur complement would provide an exact preconditioner, i.e. it would allow us to solve the linearised (13) in one iteration only. However, this would require inverting operator $\mathbb{F}$, which would be computationally extremely expensive in real-life engineering applications. Hence, practicality dictates that we compute an approximate Schur complement instead:

$$\widehat{\mathbb{S}} = \mathbb{D}\widehat{\mathbb{F}^{-1}}\mathbb{G} \tag{16}$$

and solve iteratively (relaxing if necessary) as follows:

1. solve $\mathbb{F}\mathbf{U}^{n+1/2} = \mathbf{g}_u - \mathbb{G}\mathbf{p}^n$ (*predictor step* for velocity);
2. solve $\widehat{\mathbb{S}}\delta\mathbf{p} = \mathbb{D}\mathbf{U}^{n+1/2} - \mathbf{g}_p$ (*pseudo-Laplacian* for pressure correction);
3. update pressure: $\mathbf{p}^{n+1} = \mathbf{p}^n + \delta\mathbf{p}$;
4. update velocity: $\mathbf{U}^{n+1} = \mathbf{U}^{n+1/2} - \widehat{\mathbb{F}^{-1}}\mathbb{G}\delta\mathbf{p}$ (*corrector step*);
5. if non-linear, update operator $\mathbb{F}$ (Picard step).

If $\mathbb{F}$ happens to be block-diagonal, as is the case for the primal Navier-Stokes system, then step 1 can be split into three separate linear solves, each corresponding to the momentum equation in its respective spatial dimension.

In its most basic implementation, SIMPLE approximates the inverse of $\mathbb{F}$ with the inverse of its main diagonal:

$$\mathbb{F}^{-1} \approx \widehat{\mathbb{F}^{-1}} = \left(\text{DIAG}\left(\mathbb{F}\right)\right)^{-1}. \tag{17}$$

In this paper, we focus on the variant of SIMPLE known as SIMPLEC. It operates by adding to the momentum equations in (13) some form of implicit relaxation by factor α:

$$\mathbb{F}_{\alpha} = \mathbb{F} + \alpha \text{DIAG}\left(\mathbb{F}\right) \tag{18}$$

and subsequently approximating the inverse of $\mathbb{F}$ as:

$$\mathbb{F}^{-1} \approx \widehat{\mathbb{F}^{-1}} = \frac{1}{\alpha}\left(\text{DIAG}\left(\mathbb{F}\right)\right)^{-1}. \tag{19}$$

It has been observed [25] that SIMPLEC, in the steady-state case, yields a pseudo-Laplacian pressure equation aimed at correcting the velocity increment $\left(\mathbf{U}^{n+1/2} - \mathbf{U}^{n}\right)$ rather than the velocity itself, and since it acts on relaxed velocity increments, it does not require relaxing the pressure correction step.

We previously adapted SIMPLEC to the MHFV framework. The main difference with respect to classical FV is that, when relaxing our operator $\mathcal{F}_{\Phi}$, instead of using its diagonal coefficients as in (18), we apply *inertial relaxation* in the form:

$$\mathcal{F}_{\Phi,\alpha} = \mathcal{F}_{\Phi} + \alpha\left[\text{diag}\left(\beta_F\right)\right] \tag{20}$$

where β_F is a suitable scaling factor defined for each face, related to the hybridisation procedure (i.e. the elimination of cell-averaged velocity components in the convection-diffusion operator) and the flux conservation (6) we impose at each face. Our inertial relaxation is proportional to the local Reynolds number Re_F, meaning that stronger relaxation is applied in areas where convection-to-diffusion ratio is higher. The explicit definition of β_F, omitted here, can be found in [19].

For the primal, the overall iterative algorithm is analogous to the generic one outlined above, the only difference being that, since we are dealing with incompressible flow, there is no source term for the continuity equation.

Primal SIMPLEC:

1. solve relaxed momentum:

$$\mathcal{F}_{\Phi^n,\alpha}\mathbf{U}_F^{n+1/2} = \mathbf{g}_F - \mathcal{G}\mathbf{p}_C^n + \alpha\left[\text{diag}\left(\beta_F\right)\right]\mathbf{U}_F^n; \tag{21}$$

2. solve pseudo-Laplacian:

$$\mathcal{D}\left[\text{diag}\left(\frac{1}{\alpha\beta_F}\right)\right]\mathcal{D}^T\delta\mathbf{p}_C = \mathcal{D}\mathbf{U}_F^{n+1/2}; \tag{22}$$

3. update pressure:

$$\mathbf{p}_C^{n+1} = \mathbf{p}_C^n + \delta\mathbf{p}_C; \tag{23}$$

4. update velocity:

$$\mathbf{U}_F^{n+1} = \mathbf{U}_F^{n+1/2} - \left[\text{diag}\left(\frac{1}{\alpha\beta_F}\right)\right]\mathcal{D}^T\delta\mathbf{p}_C; \tag{24}$$

5. update convective fluxes and assemble new operator $\mathcal{F}_{\Phi^{n+1},\alpha}$.

Notice that here, in the Schur complement appearing in step 2, we use the transpose of the divergence operator $\mathcal{D}^T$ rather than the gradient operator $\mathcal{G}$. As mentioned in Section 1.1, the two are identical - barring boundary conditions - when the pressure gradient term is discretised via the 1st-order scheme (8). For the 2nd-order scheme (9) that is no longer the case, meaning that by using $\mathcal{D}^T$ we introduce a further level of approximation in the Schur complement. We do so in order not to degrade the sparsity pattern of the pseudo-Laplacian, which is already rather challenging for standard linear solvers even with 1st-order connectivity. Since this approximation only affects the pressure correction step, the overall algorithm still converges.

SIMPLEC can be easily adapted to the adjoint system. The main difference is that the adjoint momentum operator $\widetilde{\mathcal{F}}^T$ is not block-diagonal; since the ability to solve separately for each velocity component is arguably one of the most attractive features of SIMPLE-like algorithms, we aim to to maintain such de-coupled nature in the adjoint version. We therefore need to find a suitable block-diagonal approximation to $\widetilde{\mathcal{F}}^T$ to be used as system matrix for the velocity prediction step, whilst treating all extra-diagonal blocks explicitly. Intuitively, we choose to use the transpose of operator $\mathcal{F}_\Phi$ itself, assembled using convective flux values Φ_F taken at the last primal iteration and therefore, presumably, converged.

As we did for the primal, we apply inertial relaxation in the form (20) to the adjoint momentum equation, which is necessary for steady-state SIMPLEC. We choose for simplicity to recycle the same scaling factor β_F from the primal, but we reserve the possibility to set a relaxation factor α independently. As for the adjoint pressure correction step, again we replace $\mathcal{G}^T$ with $\mathcal{D}$, regardless of the order of accuracy of $\mathcal{G}$, in order not to deteriorate the sparsity pattern of the Schur complement.

Adjoint SIMPLEC:

1. solve relaxed adjoint momentum:

$$\mathcal{F}_{\Phi,\alpha}^T\mathbf{U}_F^{*\,n+1/2} = \mathbf{g}_F^* - \mathcal{D}^T\mathbf{p}_C^{*\,n} - \left(\widetilde{\mathcal{F}}^T - \mathcal{F}_\Phi^T\right)\mathbf{U}_F^{*\,n} + \alpha\left[\text{diag}\left(\beta_F\right)\right]\mathbf{U}_\mathbf{F}^{*\,n}. \tag{25}$$

2. solve pseudo-Laplacian:

$$\mathcal{D}\left[\text{diag}\left(\frac{1}{\alpha\beta_F}\right)\right]\mathcal{D}^T\delta\mathbf{p}_C^* = \mathcal{G}^T\mathbf{U}_F^{*\,n+1/2} - \mathbf{g}_C^*. \tag{26}$$

3. update adjoint pressure:

$$\mathbf{p}_C^{*\,n+1} = \mathbf{p}_C^{*\,n} + \delta\mathbf{p}_C^*; \tag{27}$$

4. update adjoint velocity:

$$\mathbf{U}_F^{*\,n+1} = \mathbf{U}_F^{*\,n+1/2} - \left[\text{diag}\left(\frac{1}{\alpha\beta_F}\right)\right]\mathcal{D}^T\delta\mathbf{p}_C^*. \tag{28}$$

The term $\left(\widetilde{\mathcal{F}}^T - \mathcal{F}_\Phi^T\right) \mathbf{U}_F^{*\,n}$ on the right-hand side of (25) is the contribution due to the coupling of adjoint velocity components that we treat explicitly, and in adjoint jargon it is known as *Adjoint Transposed Convection* (ATC), a name borrowed from continuous adjoint theory.

It is also worth noticing that our adjoint SIMPLEC algorithm is very similar to what we would obtain if we were to apply AD to our solver via the so-called *Christianson's method*. The method, described in [6], operates by 1) applying reverse-mode AD to the last CFD iteration only, and 2) iterating until the adjoint states are converged. This can be interpreted as transposing the primal algorithm by using $\mathcal{F}_\Phi^T$ with a converged Φ_F for the momentum linear solves, with all off-diagonal blocks of the Jacobian - the ATC - relegated to the right-hand side.

2.2 Block-Coupled

The Block-Coupled (BCPL) solution strategy is arguably the most intuitive and straightforward way of solving a system like (13). In our framework, for the primal, it simply requires solving (4) after linearisation, i.e. with a frozen convective flux Φ_F, then update the value of Φ_F with the newly computed $\mathbf{U}_F$, re-assemble operator $\mathcal{F}_\Phi$, and iterate. In other words, BCPL proceeds from one Picard iteration to the next, without any inner Oseen iterations; it does not in general require any relaxation.

As for the adjoint system, since it is linear, a BCPL approach reduces to one linear solve with no iterations at all except for those that may be performed internally by the linear solver in use (which may or may not be iterative).

The BCPL approach entails two major drawbacks:

- it has been observed [27] that the saddle-point nature of system (4) poses a challenge for standard linear solvers, as it is difficult to precondition and, for certain types of solvers, even difficult to factorise/solve via direct methods due to the so-called *zero-block* on the main diagonal in the discrete continuity equation;

- while more traditional algorithms (such as SIMPLEC) require at each iteration solving multiple smaller linear systems, the BCPL approach requires solving the full linearised Oseen system (4) which can indeed be extremely large for industrial cases - even more so in our MHFV framework, since the number of velocity unknowns scales with the number of faces in the mesh, and not cells like in FV.

2.3 V-Coupled

We mentioned in Section 2.1 how the adjoint version of SIMPLEC solves for $\mathbf{u}_F^*$, $\mathbf{v}_F^*$ and $\mathbf{w}_F^*$ in a segregated fashion, moving the ATC to the right-hand side. Although we did not record any issues in our experiments, the ATC is known to be a troublesome term especially when treated explicitly, causing severe instabilities to the point where some researchers, such as [14, 20], resort to arbitrarily damping it, or even eliminating it completely in sensitive areas, in the hope that the final sensitivities won't be excessively affected qualitatively.

The issue is particularly evident in continuous adjoints, where the ATC appears explicitly at a PDE level, thus posing the challenge of finding a suitable discretisation as well as a stable way of treating it in the solution algorithm. However, it is reasonable to expect robustness issues in a discrete adjoint as well, especially at high Re, since high values on the right-hand side of (25) might cause instabilities, leading to divergence unless heavy relaxation is applied.

The obvious solution is to treat the ATC implicitly; we do so in our *V-Coupled* (VCPL) algorithm, which is simply a version of the adjoint SIMPLEC from Section 2.1 in which, in the

predictor step, we couple all velocity components by keeping them on the left-hand side:

$$\widetilde{\mathcal{F}}_{\alpha}^{T}\mathbf{U}_{F}^{*\,n+1/2} = \mathbf{g}_{F}^{*} - \mathcal{D}^{T}\mathbf{p}_{C}^{*\,n} + \alpha\left[\mathrm{diag}\left(\beta_{F}\right)\right]\mathbf{U}_{\mathbf{F}}^{*\,n}; \tag{29}$$

the pressure correction step (26), on the other hand, remains unchanged.

Besides tackling the above mentioned stability issues, by treating the ATC implicitly we can also expect an overall reduction on the iteration count. On the downside, the VCPL approach (29) comes with the obvious drawback of having to store larger matrices and solve larger linear systems in comparison with SIMPLEC.

Of course the VCPL approach is only relevant to the adjoint system; applying VCPL to the primal would be exactly equivalent to applying SIMPLEC, the only difference being that we would be solving at each iteration a single, larger linear system rather than breaking it down into three smaller ones.

2.4 Augmented Lagrangian

The Augmented Lagrangian (AL) preconditioning scheme for Oseen-type problems was first presented by [1], and further developed with several variants by e.g. [2, 16]. AL-based preconditioners have been so far investigated mostly within FE frameworks, and have been proven to be theoretically almost optimal [2] in terms of mesh and Re-dependency. Despite its drawbacks, we therefore deem it worth investigating the AL methodology and attempting to adapt it to our MHFV scheme.

The AL core idea is to re-write system (13) as

$$\begin{bmatrix} \mathbb{F}_{\gamma} & \mathbb{G} \\ \mathbb{D} & 0 \end{bmatrix}\begin{pmatrix} \mathbf{U} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_{u,\gamma} \\ \mathbf{g}_{p} \end{pmatrix}, \tag{30}$$

where

$$\mathbb{F}_{\gamma} = \mathbb{F} + \gamma\mathbb{G}\mathbb{W}^{-1}\mathbb{D} \tag{31}$$

and

$$\mathbf{g}_{u,\gamma} = \mathbf{g}_{u} + \gamma\mathbb{G}\mathbb{W}^{-1}\mathbf{g}_{p}. \tag{32}$$

Systems (30) and (13) are clearly equivalent, since (31) and (32) simply add to the velocity blocks a term proportional to residual of the continuity equation. In other words, the AL method corresponds to adding to the momentum equations a penalisation term which is driven to zero for a converged solution. In (31) and (32), γ is a positive augmentation factor, and $\mathbb{W}$ an arbitrary symmetric positive definite (SPD) matrix.

As observed by [2], an advantage of using an AL formulation (besides theoretical near-optimality of the preconditioner itself) is that the issue of finding a good approximation for the Schur complement (15) is circumvented. If the augmentation factor is large enough, then the penalisation term will prevail on the operator $\mathbb{F}$ itself, thus justifying the approximation

$$\mathbb{F}_{\gamma}^{-1} \approx \widehat{\mathbb{F}_{\gamma}^{-1}} = \left(\gamma\mathbb{G}\mathbb{W}^{-1}\mathbb{D}\right)^{-1} \tag{33}$$

which yields an approximate Schur complement in the form:

$$\widehat{\mathbb{S}} = \mathbb{D}\left(\gamma\mathbb{G}\mathbb{W}^{-1}\mathbb{D}\right)^{-1}\mathbb{G} = \frac{1}{\gamma}\mathbb{W}. \tag{34}$$

The generic AL iterative procedure is outlined as follows:

1. solve $\mathbb{F}_\gamma \mathbf{U}^{n+1} = \mathbf{g}_{u,\gamma} - \mathbb{G}\mathbf{p}^n$ (penalised momentum equation);
2. solve $\frac{1}{\gamma}\mathbb{W}\delta\mathbf{p} = \mathbb{D}\mathbf{U}^{n+1} - \mathbf{g}_p$ (pressure correction);
3. update pressure: $\mathbf{p}^{n+1} = \mathbf{p}^n + \delta\mathbf{p}$;
4. if non-linear, assemble new augmented momentum operator $\mathbf{F}_\gamma$.

In FE, matrix $\mathbb{W}$ is often chosen to be the so-called pressure mass matrix or, for practical reasons, a diagonal approximation of it (usually a lumped mass matrix, or else simply its main diagonal); in that case, the pressure correction step 2 involving the approximate Schur complement (34) simply requires inverting a diagonal matrix, i.e. it doesn't actually involve a linear solve. The main drawbacks of AL-based preconditioners are:

- the augmented momentum operator $\mathbb{F}_\gamma$ is no longer block-diagonal, because each of the velocity components contributes to the penalisation term of the momentum equations in all spatial dimensions. Therefore, step 1 in the procedure above entails a single coupled linear solve for u, v and w at the same time and, unlike with SIMPLEC, it cannot be de-coupled into segregated smaller systems, at least for the basic AL formulation;
- as shown by [1], too large values of γ cause the penalised block $\mathbb{F}_\gamma$ to become increasingly ill-conditioned since $\mathbb{G}\mathbb{W}^{-1}\mathbb{D}$ is a singular matrix, and therefore increasingly challenging for linear solvers. On the other hand, the approximated Schur complement (34) is only close to the exact one if γ is large enough, hence if γ is chosen too small, the overall algorithm may underperform or even diverge. A trade-off between these two extremes is thus required, possibly combined with inexact solves of the augmented momentum equations.

Developing an AL preconditioner for our primal MHFV solver is fairly straight-forward. We already described in Section 1.1 the divergence operator $\mathcal{D}$ (7); we proceed by adding a penalisation term in the form (31) to our MHFV momentum operator as follows:

$$\mathcal{F}_{\Phi,\gamma} = \mathcal{F}_\Phi - \gamma\mu\mathcal{D}^T\left[\text{diag}\left(\frac{1}{|C|}\right)\right]\mathcal{D}, \tag{35}$$

whereas the right-hand side remains unchanged since there is no source term for the continuity equation. A few observations are in order:

- as we did for the SIMPLEC Schur complement, here we use $\mathcal{D}^T$ rather than $\mathcal{G}$ in the AL augmentation term. We already highlighted how the two are identical in the case of 1st-order pressure gradient scheme (8), and since a 2nd-order accurate $\mathcal{G}$ as defined in (9) would further increase the complexity of the already challenging augmented operator (35), we choose to stick with the 1st-order operator $\mathcal{D}^T$. It is important to stress that such choice does not in any way affect the order of accuracy of the solution itself: the AL algorithm ultimately drives $\mathcal{D}\mathbf{U}_F$, and therefore the penalisation term, to zero, and the final flow field satisfies the original non-augmented Navier-Stokes problem (4);
- $|C|$ is the volume of cell C, and the diagonal matrix diag $(|C|)$ plays in (35) the role of $\mathbb{W}$ in the generic formulation (31). In a FV-like framework, such as ours, this is indeed the equivalent of the FE pressure mass matrix, and since it is diagonal it can be inverted with no further approximation;

- besides the augmentation parameter γ, we also multiply the penalisation term by a scaling factor μ, in order to maintain γ within a range of values that work reasonably well regardless of the specific mesh size and problem physics. Following suggestions from [1], we set μ to scale with the velocity: $\mu = \max|\vec{U}_C|$, where $|\vec{U}_C|$ the cell-averaged velocity magnitude.

We can now outline our primal AL iterative procedure.

Primal Augmented Lagrangian:

1. solve augmented momentum:

$$\mathcal{F}_{\Phi^n,\gamma}\mathbf{U}_F^{n+1} = \mathbf{g}_F - \mathcal{G}\mathbf{p}_C^n; \tag{36}$$

2. compute pressure correction:

$$\delta\mathbf{p}_C = -\gamma\mu\left[\text{diag}\left(\frac{1}{|C|}\right)\right]\mathcal{D}\mathbf{U}_F^{n+1}; \tag{37}$$

3. update pressure:

$$\mathbf{p}_C^{n+1} = \mathbf{p}_C^n + \delta\mathbf{p}_C; \tag{38}$$

4. update convective fluxes and assemble new augmented operator $\mathcal{F}_{\Phi^{n+1},\gamma}$.

Lastly, we propose here an adaptation of the AL preconditioner to the adjoint Navier-Stokes system. Following the generic methodology above, the penalised adjoint momentum operator becomes

$$\widetilde{\mathcal{F}}_\gamma^T = \widetilde{\mathcal{F}}^T - \gamma\mu\mathcal{D}^T\left[\text{diag}\left(\frac{1}{|C|}\right)\right]\mathcal{D}, \tag{39}$$

and its right-hand side:

$$\mathbf{g}_{F,\gamma}^* = \mathbf{g}_F^* - \gamma\mu\mathcal{D}^T\left[\text{diag}\left(\frac{1}{|C|}\right)\right]\mathbf{g}_C^*. \tag{40}$$

The scaling factor μ is taken from the primal, while the penalisation coefficient γ is defined independently.

Notice that, as we did for the primal, in (39) we replaced $\mathcal{G}^T$ with $\mathcal{D}$ in the adjoint augmentation term in order to avoid excessively complex connectivities. For the adjoint, however, we need to take a few extra measures: the adjoint continuity equation dictates that we penalise by a quantity proportional to $\left(\mathcal{G}^T\mathbf{U}_F^* - \mathbf{g}_C^*\right)$ exactly; therefore, if $\mathcal{G}^T \neq \mathcal{D}$, penalising for $(\mathcal{D}\mathbf{U}_F^* - \mathbf{g}_C^*)$ as we do in (39) and (40) may never converge, as it would be the equivalent to solving the adjoint of a 1st-order accurate primal, which would be inconsistent with the actual Jacobian. A way around the issue is to treat all the (presumably small) 2nd-order contributions explicitly, as shown in the algorithm below.

Adjoint Augmented Lagrangian:

1. solve adjoint augmented momentum:

$$\widetilde{\mathcal{F}}_\gamma^T\mathbf{U}_F^{*\,n+1} = \mathbf{g}_{F,\gamma}^* - \mathcal{D}^T\mathbf{p}_C^{*\,n} - \gamma\mu\mathcal{D}^T\left[\text{diag}\left(\frac{1}{|C|}\right)\right]\left(\mathcal{G}^T - \mathcal{D}\right)\mathbf{U}_F^{*\,n}; \tag{41}$$

2. compute adjoint pressure correction:

$$\delta \mathbf{p}_C^* = -\gamma\mu \left[\text{diag}\left(\frac{1}{|C|} \right) \right] \left(\mathcal{G}^T \mathbf{U}_{\mathbf{F}}^{*\,n+1} - \mathbf{g}_C^* \right); \tag{42}$$

3. update adjoint pressure:

$$\mathbf{p}_C^{*\,n+1} = \mathbf{p}_C^{*\,n} + \delta \mathbf{p}_C^*. \tag{43}$$

3 NUMERICAL RESULTS

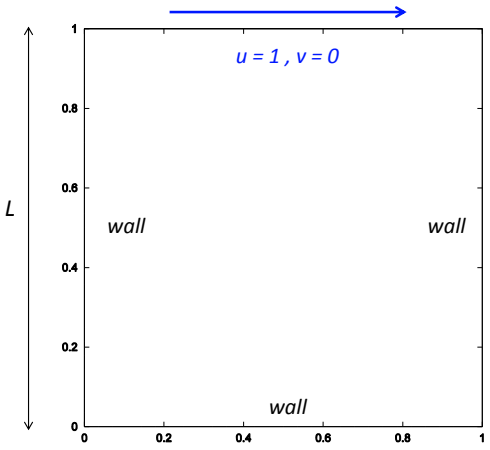


Figure 3: Lid-driven cavity test case.

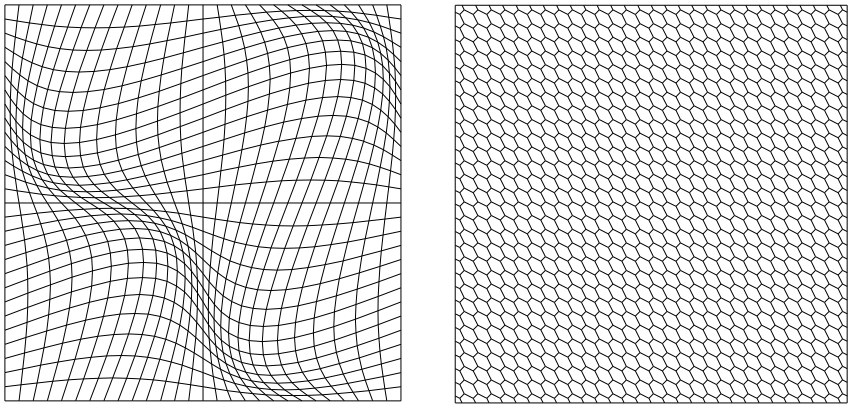


Figure 4: Mesh types used for the lid-driven cavity test case.

For a first evaluation of the performance of our (primal) algorithms, we test them on the well-known 2D lid-driven cavity benchmark test case, set-up as shown in Figure 3. We run it at two different Reynolds numbers, $Re = 10^2$ and $Re = 10^3$, on the two different mesh types shown

in Figure 4, with Type A being a quadrilateral, highly distorted/non-orthogonal mesh, and Type B a more regular, polygonal honeycomb-like mesh. We test on a series of progressively refined meshes of both types in order to assess h-dependency.

As for the discretisation schemes, we use 2nd-order MHFV operators for both velocity and pressure, the first stabilised via our ULSQR technique described in [19], the second as in (9). All simulations are run down to a tolerance of 10^{-4} on scaled residuals.

3.1 Block-Coupled

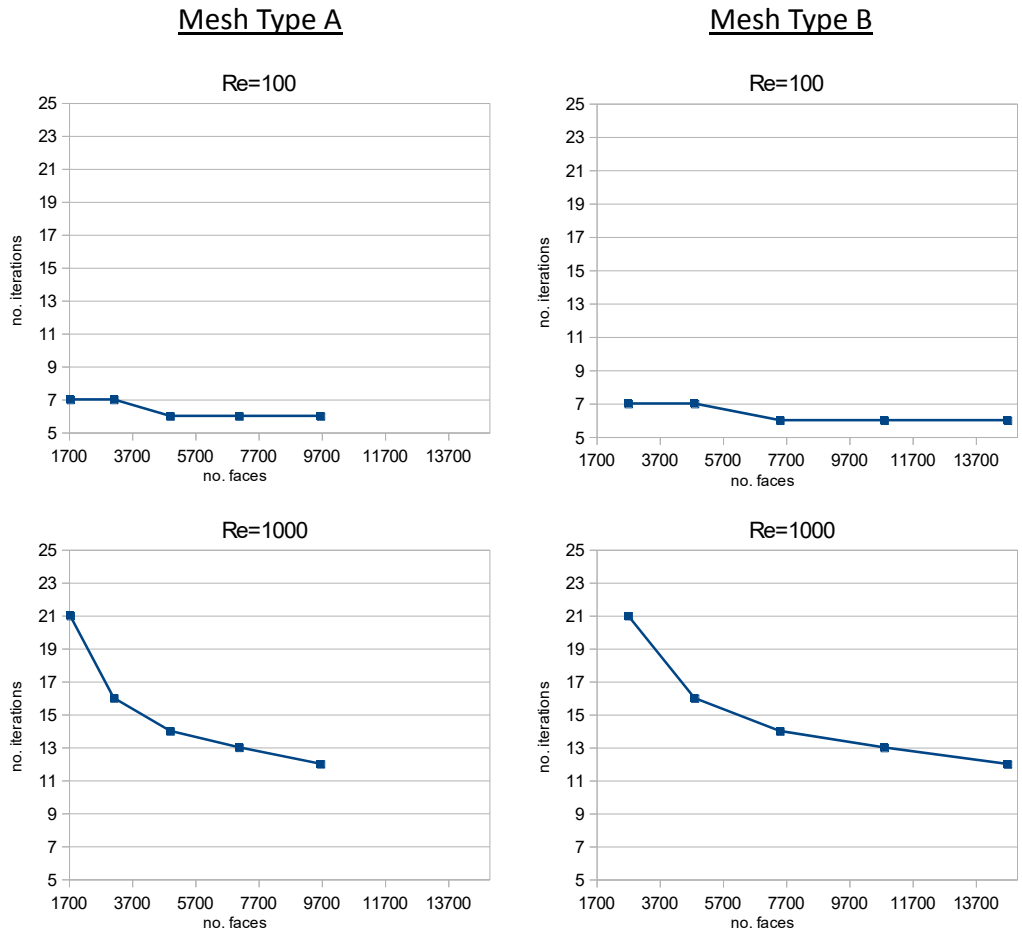


Figure 5: MHFV Block-Coupled. Lid-driven cavity test case. Iteration count in function of mesh type, mesh size and Re.

Results for BCPL are shown in Figure 5 where we plot, for each mesh type and each Re, the iteration count n_{iter} against the number of mesh faces, i.e. against mesh refinement.

The figures are rather positive: at $Re = 10^2$ the iteration count is consistently small (below 10), and does not appear to be affected by mesh size or quality; at $Re = 10^3$ we record an asymptotic behaviour of n_{iter} with respect to mesh size, namely the iteration count is higher on coarser meshes and tends to settle around a value in the order of 10 as we refine the mesh; such behaviour is identical over both mesh types. Evidence thus suggests that our BCPL algorithm is independent of mesh quality and mesh size (barring coarse mesh cases, where it underperforms

slightly), and only marginally affected by an increase of Re.

3.2 SIMPLEC

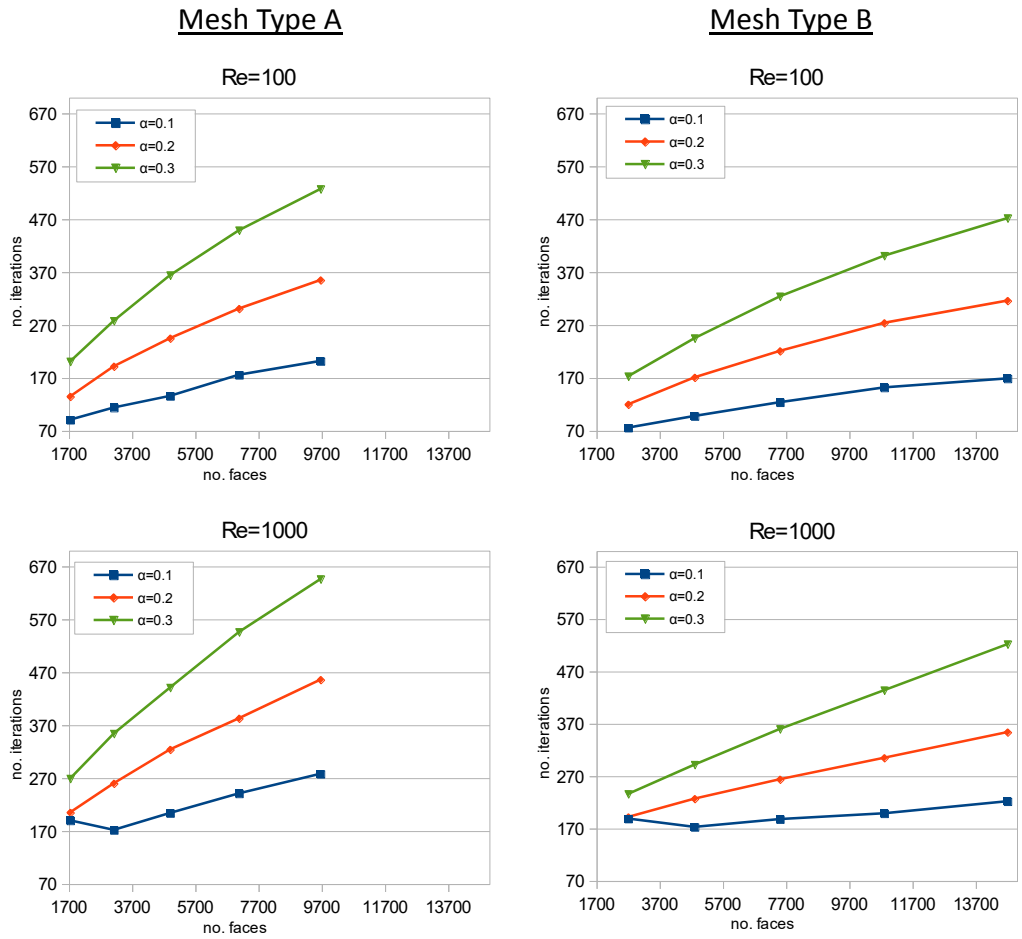


Figure 6: MHFV SIMPLEC. Lid-driven cavity test case. Iteration count in function of mesh type, mesh size, Re and relaxation factor α.

We perform on our SIMPLEC scheme the same tests presented above for BCPL. For SIMPLEC we also try out different values of relaxation factor α, in order to verify whether the optimal value depends on the other parameters.

The resulting graphs are reported in Figure 6. It appears that, for both mesh types and both Re, setting $\alpha = 0.1$ gives the best performance in terms of both iteration count and algorithm scaling against problem size. Results are encouraging in that sense, as that they suggest that the optimal relaxation factor does not depend on mesh quality, and it is only slightly influenced by grid coarseness and problem physics.

On the other hand, these results also highlight two severe limitations of SIMPLEC itself:

- the iteration count scales with the problem size, i.e. refining the mesh causes n_{iter} to grow. The iteration count itself is in general very high, roughly an order of magnitude greater than with the BCPL approach;

- the algorithm heavily depends on numerical parameters, and in particular on mesh type. Notice how n_{iter} is in general much higher, and grows more rapidly with the problem size, for mesh Type A. This can be attributed to the irregularity of Type A, which features several strongly distorted cells as well as a wide range of cell volumes/face areas - whereas Type B is only slightly non-orthogonal, and perfectly regular in terms of element size. This might be causing SIMPLEC to underperform on Type A compared to Type B, despite the considerably smaller problem size.

3.3 Augmented Lagrangian

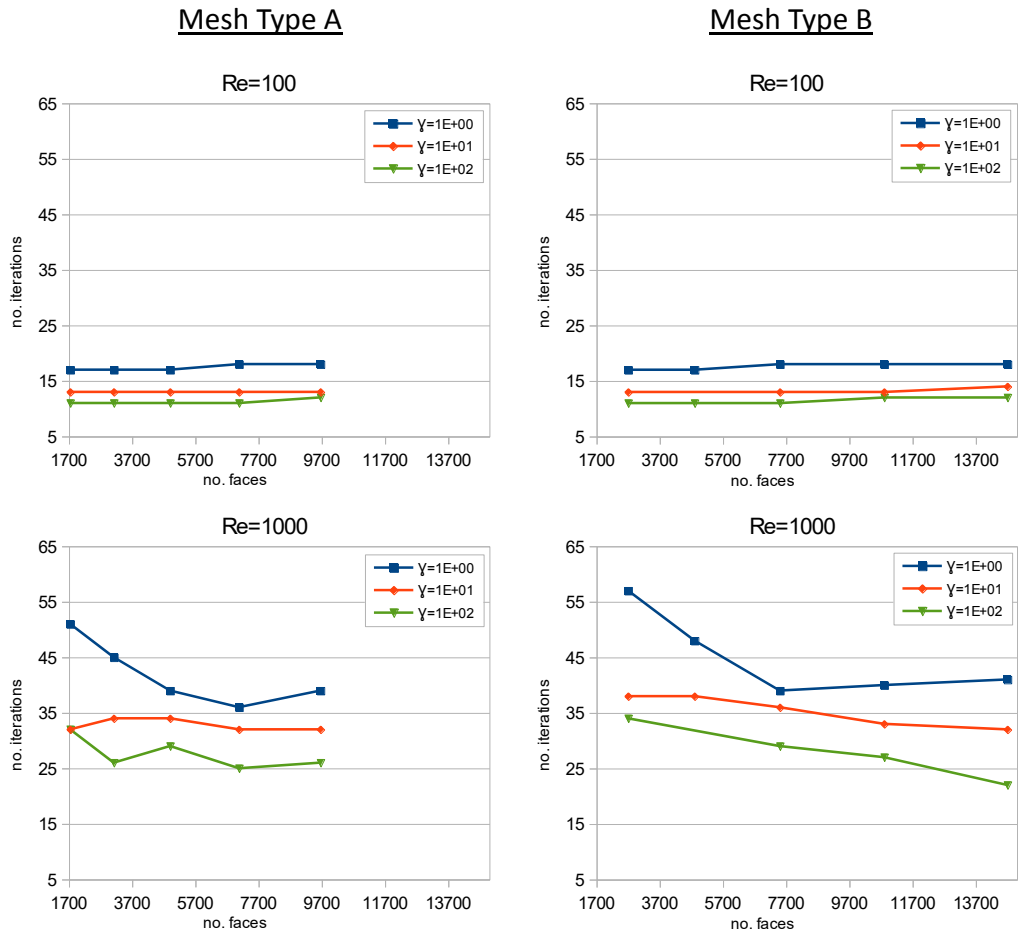


Figure 7: MHFV Augmented Lagrangian. Lid-driven cavity test case. Iteration count in function of mesh type, mesh size, Re and augmentation factor γ.

Once again we run the lid-driven test case to validate the performance of AL in function of Re, mesh quality and mesh size; we also test for different values of penalisation factor γ. Results are reported in Figure 7. The curves show that there is no definite correlation between n_{iter} and grid size nor grid quality, hence our AL implementation is completely mesh independent. Similarly to the other preconditioners, we observe however a slight increase in iteration count for higher Re values. The value of n_{iter} itself remains in the order of 10, hence much lower than that of SIMPLEC and fairly close to BCPL results, thus confirming the near-optimality of

the AL algorithm.

Results also highlight how higher values of γ consistently reduce the total number of iterations; this is expected because, when γ is high, the convection-diffusion operator $\mathcal{F}_\Phi$ in (35) becomes negligible compared to the penalisation term, and therefore the diagonal approximate Schur complement used in (37) is close to the exact one. However, as observed in Section 2.4, too high values of γ will cause the augmented operator $\mathcal{F}_{\Phi,\gamma}$ to be nearly singular, thus hindering the linear solve for velocity prediction. Keeping γ in the order of 1 appears to be a resonable choice; in order to run tests at higher γ values, we circumvent the linear solver issue by using direct solvers, which is however not a viable option for real industrial cases.

3.4 Comparison of primal Navier-Stokes solution algorithms

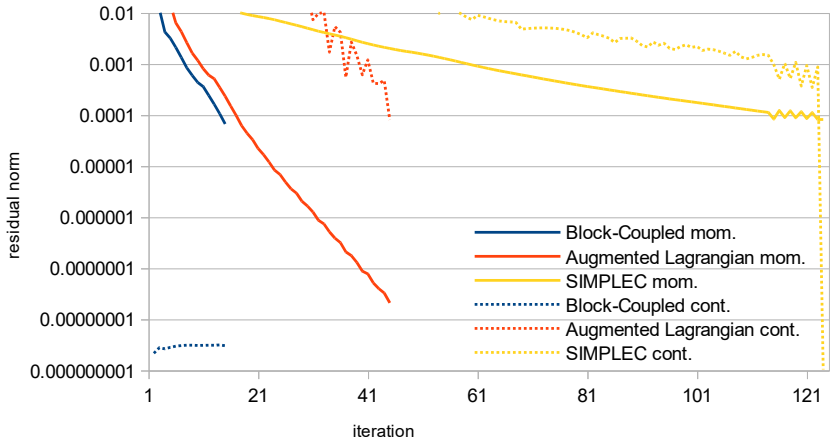
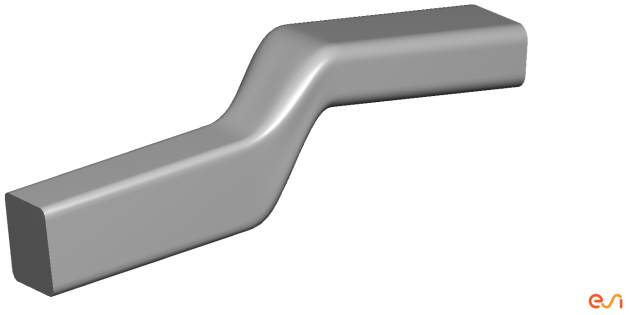


Figure 8: S-bend test case. Convergence history of the x-momentum and continuity scaled residual norms for different solution algorithms.

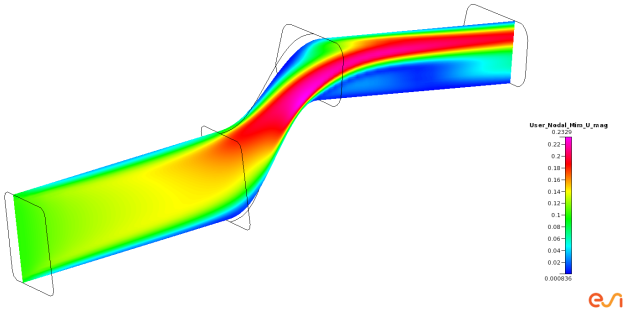
We run a further comparison of the primal BCPL, SIMPLEC and AL algorithms on a 3D flow solve: the S-bend, a popular benchmark test case involving an internal flow simulation through an S-shaped air duct. We run the S-bend at a fairly low $Re \approx 400$ and over a relatively coarse, hexahedral mesh (approximately 41k elements and 126k faces). We show in Figure 9 the S-bend geometry and some solution contours.

For the current test, we set the following algorithm parameters: no relaxation for BCPL; $\alpha = 0.1$ for SIMPLEC (experimentally determined to be close to optimal); $\gamma = 5$ for AL (which we found to produce good overall convergence properties whilst maintaining the augmented system relatively well-conditioned). Tolerance is set to 10^{-4} for both momentum and continuity equations. As for the numerical scheme, we set our MHFV solver to 2nd-order accuracy for both pressure and velocity, the latter stabilised via the previously mentioned ULSQR strategy.

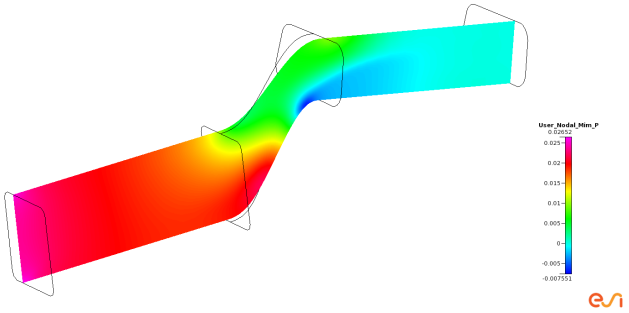
A comparison is shown in Figure 8, where we report the history of the x-momentum and continuity scaled residuals for each preconditioner. The graph demonstrates once again the definite superiority of both BCPL and AL over SIMPLEC in terms of convergence rate. The reader may also notice a stagnation-like behaviour of SIMPLEC at the last few iterations, also noticeable in other graphs throughout this paper. This is due to the fact that our CFD solver only performs a linear solve if the initial residual norm for the specific equation being solved is above the global tolerance; not all segregated equations reach convergence at the exact same



(a)



(b)



(c)

Figure 9: S-bend test case. Domain geometry (a), velocity magnitude (b), pressure field (c).

iteration, hence a synchronisation phase happens at the end of the process, essentially causing the equations to be solved alternately until all residual norms align just below tolerance.

It is remarkable how BCPL and AL exhibit very similar convergence properties. The difference is that AL runs several extra iterations, due to the fact that the continuity equation does not converge at the same rate for the two: BCPL solves it exactly at each step, while AL iteratively reduces the continuity residual via the penalisation mechanism. The extra AL iterations evidently result in a better converged momentum equation, at the expense of continuity being solved only down the set global tolerance, whilst BCPL solves it exactly - or almost, depending on the linear solver in use.

3.5 Comparison of adjoint Navier-Stokes solution algorithms

To test and compare performances of our adjoint preconditioners, we first run a simulation on the simple 2D square box test case represented in Figure 10, a popular benchmark case

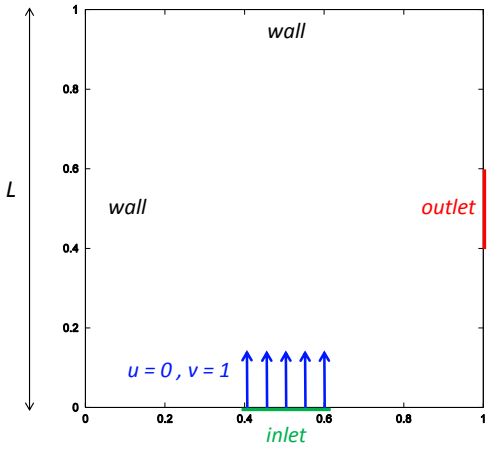


Figure 10: Square box test case.

often used to illustrate adjoint-based optimisation techniques; the cost function J, necessary to compute the adjoint right-hand side, is defined as the total pressure drop across the box's boundary $\partial\Omega$:

$$J = -\int_{\partial\Omega} \left(p + \frac{1}{2}|\vec{U}|^2 \right) \tag{44}$$

We run the primal at $Re \approx 10^3$ on a regular polygonal mesh (Type B in Figure 4); MHFV operators are all 2nd-order accurate. We converge the primal down to a normalised resudual of 10^{-6} in order to secure a robust discrete adjoint; for the adjoint itself, we set a tolerance of 10^{-4}. We relax by $\alpha = 0.3$ both SIMPLEC and VCPL, while for AL we use $\gamma = 10$. We report

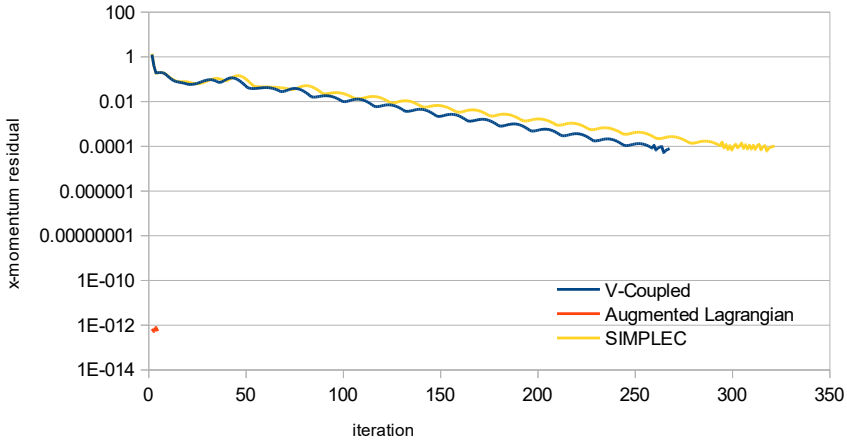


Figure 11: Square box test case. Convergence history of the adjoint x-momentum scaled residual norm for different solution algorithms.

in Figure 11 the convergence history for the adjoint x-momentum. SIMPLEC and VCPL both exhibit a somewhat oscillatory behaviour, and both converge roughly at the same rate. The gain in iteration count obtained by opting for VCPL over SIMPLEC is moderate at best; a series of other test cases, not reported here, also confirm that the reduction on n_{iter}, when present,

is limited to a maximum of about 10%, at the considerable cost of having to solve a velocity coupled system at each iteration. It should also be mentioned, however, that we mostly test on fairly low-Re; we therefore maintain the argument that, for higher Re simulations, treating the ATC implicitly may bring about significant advantages in terms of solver stability.

AL, on the other hand, outperforms by far the SIMPLE-like preconditioners. For this particular test case the adjoint momentum residual falls well below tolerance - in fact, down to machine precision - from the very first iteration, whilst the adjoint continuity equation, not shown here, converges to the given tolerance in 4 iterations only. Therefore, as for the primal, AL displays a behaviour very close to the optimal BCPL approach - the adjoint system is linear, hence a fully implicit adjoint BCPL approach does not need any outer iterations.

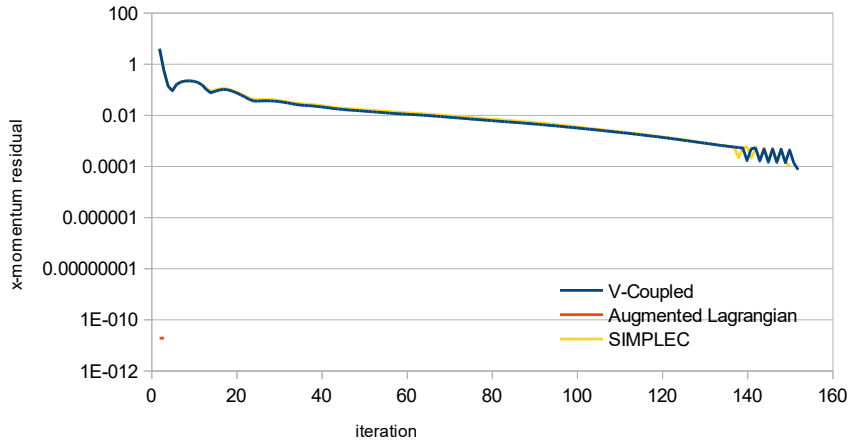
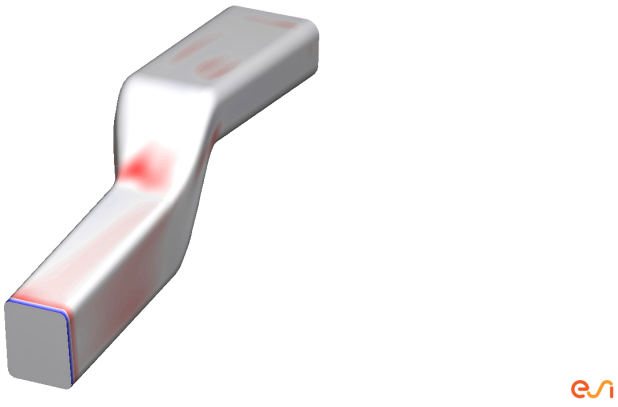


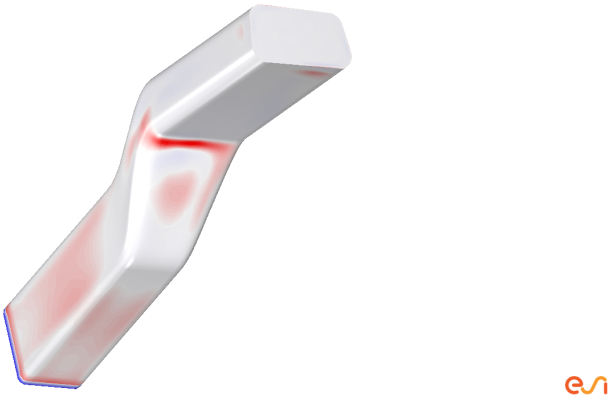
Figure 12: S-bend test case. Convergence history of the adjoint x-momentum scaled residual norm for different solution algorithms.

To further validate, we also assemble and solve the discrete adjoint of the 3D S-bend test case whose primal we described in section 3.4; again we define as cost function the total pressure drop (44). We use an optimal relaxation factor, found to be $\alpha = 0.28$, for both SIMPLEC and VCPL, and an augmentation factor $\gamma = 10$ for AL. Results are reported in Figure 12; we also show, for the sake of generating reader interest, the computed surface sensitivity field in Figure 13.

Results on the S-bend confirm the observations previously made on the 2D square box: SIMPLEC and VCPL perform comparably (in fact, to be precise, for the S-bend VCPL even takes a few more iterations than SIMPLEC due to the final residual synchronisation process explained in Section 3.4), whilst AL drops the adjoint momentum residual down to near-machine precision from the very start, and only takes 3 iterations to converge the adjoint continuity below tolerance. It should be mentioned, however, that all of the drawbacks highlighted for the primal AL also affect the adjoint AL, namely: on one hand, the difficulties in solving the augmented momentum other than via direct methods for too large values of γ; on the other hand, the failure of the overall AL algorithm for a γ too small.



(a)



(b)

Figure 13: S-bend test case. Surface sensitivity field: red areas indicate zones where pulling the surface outwards results in total pressure loss reduction.

4 CONCLUSIONS

Despite our MHFV discretisation scheme being an alternative, non-standard form of the discrete Navier-Stokes equations, numerical results shown in this paper reveal how the adaptation of preconditioning techniques developed for classical Finite Volumes or Finite Elements schemes leads to algorithms exhibiting behaviours similar to those described in traditional literature. SIMPLE-type schemes suffer from poor convergence rates and excessive dependency on mesh quality and size; alternative schemes, such as BCPL and AL, although fully mesh-independent and theoretically near-optimal, are affected by several practical issues, notably caused by the size and/or complexity of the linear solves involved. The same applies when such algorithms are run for the adjoint Navier-Stokes system.

Convergence properties and robustness aside, there is one other aspect, paramount in industrial contexts, which we haven't yet delved into: the cost in terms of CPU time. Our tests show that SIMPLEC benefits from a lower CPU time per iteration, as it involves solving smaller, better conditioned systems. On the other hand, the CPU time per iteration is found to scale similarly for all algorithms, and since the BCPL/AL iteration count is an order of magnitude lower than for SIMPLEC, the former two remain by far the best choice.

Interestingly enough, an AL iteration appears to be roughly twice as expensive as a BCPL one, despite AL solving for smaller matrices; this is easily explained by the fact that AL matri-

ces, although smaller, are also less sparse: the face-to-face connectivity (Figure 2) repeated on the off-diagonal blocks of the augmented momentum operator easily gives rise to AL matrices with more non-zeroes than the full Oseen matrix itself. Of course the issue - which is scheme-dependent and may be less aggravating for traditional FV methods - could be circumvented via a more efficient preconditioning of the AL linear solves but, as we mentioned, this can prove to be a very challenging task and beyond the scope of our current research.

In the near future we therefore plan to implement a MHFV version of the *Modified Augmented Lagrangian* (MAL) described in [2], which is indeed designed to alleviate some of the issues discussed above whilst still benefiting from most advantages brought about by AL. We also plan to conduct further research related to Navier-Stokes preconditioning in general, namely by investigating approximate commutators such as the *Pressure Convection-Diffusion* (PCD) [15] and the *Least Squares approximate commutator* [10].

ACKNOWLEDGEMENTS

This work was funded by the EU through the FP7-PEOPLE-2012-ITN "AboutFlow" Grant agreement number 317006.

REFERENCES

[1] M. Benzi, M. A. Olshanskii, An augmented Lagrangian-based approach to the Oseen problem. *SIAM Journal on Scientific Computing*, **28**, 2095–2113, 2006.

[2] M. Benzi, M. A. Olshanskii, Z. Wang, Modified augmented Lagrangian preconditioners for the incompressible Navier-Stokes equations. *International Journal for Numerical Methods in Fluids*, **66**(4), 486–508, 2011.

[3] F. Brezzi, K. Lipnikov, M. Shashkov, A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Mathematical Models and Methods in Applied Sciences*, **15**(10), 1533–1551, 2005.

[4] F. Brezzi, K. Lipnikov, M. Shashkov, Convergence of the Mimetic Finite Difference method for diffusion problems on polyhedral meshes. *SIAM Journal on Numerical Analysis*, **43**(5), 1872–1896, 2005.

[5] F. Brezzi, R.S. Falk, L.D. Marini, Basic principles of mixed virtual element method. *Mathematical Modeling and Numerical Analysis*, **48**(4), 1227–1240, 2014.

[6] B. Christianson, Reverse accumulation and attractive fixed points. *Optimization Methods and Software*, **3**, 311–326, 1994.

[7] J. Droniou, Remarks on discretizations of convection terms in hybrid mimetic mixed methods. *Networks and Heterogeneous Media*, **5**(3), 545–563.

[8] J. Droniou, R. Eymard, Study of the mixed finite volume method for Stokes and Navier-Stokes equations. *Numerical Methods for Partial Differential Equations*, **25**(1)., 137–171, 2008.

[9] H. Elman, V. E. Howle, J. Shadid, R. Shuttleworth, R. Tuminaro, Block preconditioners based on approximate commutators. *SIAM Journal of Scientific Computing*, **27**(5), 1651–1668, 2006.

[10] H. Elman, V. E. Howle, J. Shadid, D. Silvester, R. Tuminaro, Least squares preconditioners for stabilized discretizations of the NavierStokes equations. *SIAM Journal of Scientific Computing*, **30**(1), 290–311, 2007.

[11] A. Griewank, A. Walther, Evaluating derivatives: principles and techniques of Algorithmic Differentiation. *SIAM, second edition*, 2008.

[12] A. Jameson, Aerodynamic design via control theory. *Journal of Scientific Computing*, **3**, 233–260, 1988.

[13] A. Jameson, Optimum aerodynamic design using CFD and control theory. *AIAA Paper* **95-1729**, 1995.

[14] G. Karpouzas, E.M. Papoutis-Kiachagias, T. Schumacher, E. de Villiers, K.C. Giannakoglou, C. Othmer, Adjoint optimization for vehicle external aerodynamics. *International Journal of Automotive Engineering*, **7**, 2016.

[15] D. Kay, D. Loghin, A. Wathen, A Preconditioner for the Steady-State Navier-Stokes Equations. *SIAM Journal of Scientific Computing*, **24**(1), 237–256, 2002.

[16] M. A. Olshanskii, M. Benzi, An augmented Lagrangian approach to linearized problems in hydrodynamic stability. *SIAM Journal on Scientific Computing*, **30**, 1459–1473, 2008.

[17] M. Oriani, G. Pierrot, Alleviating adjoint solver robustness issues via mimetic CFD discretization scheme. *OPT-I: International Conference on Engineering and Applied Sciences Optimization*, Kos (Greece), 4-6 June, 2014.

[18] M. Oriani, G. Pierrot, Towards converged adjoint state for large industrial cases by improving the discretization schemes. *WCCM XIECCM VECFD VIIACM-ECCOMAS*, Barcelona (Spain), July 20-25, 2014.

[19] M. Oriani, G. Pierrot, A Mixed Hybrid Finite Volume scheme for incompressible Navier-Stokes. *NAFEMS World Congress*, San Diego (California), 21-24 June, 2015.

[20] C. Othmer, Adjoint methods for car aerodynamics. *Journal of Mathematics in Industry*, **4**(6), 2014.

[21] G. Pierrot, Equational Differentiation of incompressible flow solvers as a middle ground between continuous and discrete adjoint methodologies. *WCCM XIECCM VECFD VIIACM-ECCOMAS*, Barcelona (Spain), July 20-25, 2014.

[22] O. Pironneau, On optimum design in fluid mechanics. *Journal of Fluid Mechanics*, **64**, 97–110, 1974.

[23] M. Rehman, C. Vuik, G. Segal, A comparison of preconditioners for incompressible Navier-Stokes solvers. *International Journal for Numerical Methods in Fluids*, **57**(12), 1731–1751, 2008.

[24] A. Segal, M. Rehman, C. Vuik, Preconditioners for incompressible Navier-Stokes solvers. *Numerical Mathematics: Theory, Methods and Applications*, **3**(3), 245–275, 2010.

[25] H. Veersteg, W. Malalasekera, An introduction to Computational Fluid Dynamics: the Finite Volume method. *Pearson Education, Harlow*, 2007.

[26] L.B. Da Veiga, J. Droniou, G. Manzini, A unified approach to handle convection terms in Finite Volumes and Mimetic Discretization Methods for elliptic problems. *IMA Journal of Numerical Analysis*, **31**(4), 1357–1401, 2011.

[27] C. Vuik, A. Segal, Solution of the coupled Navier-Stokes equations. *TU Delft internal report* **95-28**, 1995.

[28] S. Wille, O. Staff, A.F.D. Loula, Efficient a priori pivoting schemes for a sparse direct Gaussian equation solver for the mixed finite element formulation of the Navier-Stokes equations. *Applied Mathematical Modelling*, **28**(7), 607–616, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DEFROSTER NOZZLE SHAPE OPTIMIZATION USING THE CONTINUOUS ADJOINT METHOD

L.A. Germanou[2,1], E.M. Papoutsis-Kiachagias[2], A. Delacroix[1] and K.C. Giannakoglou[2]

[1]Toyota Motor Europe NV/SA,
e-mail: antoine.delacroix@toyota-europe.com

[2] Parallel CFD & Optimization Unit, School of Mechanical Engineering, National Technical University of Athens, Greece,
e-mail: {lefkatza,vaggelisp}@gmail.com, kgianna@central.ntua.gr

Keywords: Continuous Adjoint Method, Shape Optimization, HVAC Defroster Nozzle, Defrosting–demisting performance.

Abstract. *This paper presents the use of the continuous adjoint method, developed by the Parallel CFD & Optimization Unit of National Technical University of Athens (NTUA) in the OpenFOAM environment, for the shape optimization of a passenger car defroster nozzle, including experimental validation performed at Toyota Motor Europe (TME). The defroster nozzle plays a major role in the demisting-defogging of the windshield, by blowing high velocity hot air jets supplied by the HVAC (Heating, Ventilation and Air Conditioning) unit of the vehicle. For a well–designed defroster nozzle, the time required for dispelling condensation or frost on the windshield must be reasonable; the nozzle must also have the capability to perform uniform defrosting from the bottom of the windshield to its top, without patches of condensation. In view of the above, an appropriate objective function, to be minimized, is the integral of the difference of the air velocity from a target (desirable) one over a thin control volume defined close to the windshield, inside the car cabin. To set up the optimization problem, the shape of a reference defroster nozzle is allowed to vary according to the morphing capabilities of a volumetric NURBS tool developed by NTUA; the latter is also used for deforming the computational mesh at each optimization cycle, by adapting it to the changed defroster shape. The CFD analysis is based on RANS, using the k-ϵ turbulence model. The optimization loop uses the gradient of the objective function with respect to the coordinates of the control points of a volumetric B–splines lattice, which is computed using the continuous adjoint method. Experimental tests performed to measure the actual velocity pattern on the windshield include velocity measurements with a hot-wire anemometer. A convincing comparison between CFD analysis and measurements is presented. The improved demisting performance of the geometry resulted from the adjoint optimization was also experimentally validated, using rapid prototyping to manufacture the designed defroster nozzle.*

1 INTRODUCTION

The safety and thermal comfort of automotive passengers are the most important factors in the development of the automotive HVAC system, [6]. HVAC is responsible for the demisting and defrosting of the vehicle's windows and for creating/maintaining a pleasant climate inside the cabin by controlling air humidity and temperature. The defroster nozzle, as part of the HVAC system of vehicles, plays a major role in the demisting-defrosting of the windscreen. Demisting refers to any function intended to remove a film of condensate from the internal face of the surface of the windscreen. Defrosting refers to any function intended to eliminate frost or ice from the external surface of the windscreen. The HVAC unit provides hot air to the nozzle which is, blowing high velocity air jets to the windscreen.

Among other, windshield defrosting performance constitutes a compulsory test according to national and international legislation since it has a significant impact on driving safety. The formation of frost on the windshield and front door glasses during cold season can be proved dangerous as it is veiling the drivers view and disturbing driving. Therefore, defroster performance is seriously taken into consideration during the design of HVAC system. On the other hand, the HVAC system has to meet the following performance requirements, [1]: (a) the time required for dispelling condensation or frost on the windshield must be reasonable and (b) uniform defrosting ideally on the whole surface of the windshield must be ensured, for the latter to become clear without being spotty or with condensation patches.

According to previous research [1, 4], the demist pattern is related to the air velocity distribution on the windshield, reflecting the defrosting performance to some extent. The defroster nozzle must, therefore, be designed to provide optimal air velocity distribution. This requirement is expressed in the form of an (integral) objective function to be minimized using a gradient–based optimization method.

CFD–based optimization methods based on either evolutionary algorithms or gradient–based optimization can be used for the design of the defroster nozzle. In the latter case, the adjoint method computes the gradient of the objective function with respect to (w.r.t.) the design variables, with a cost which is independent of the number of design variables. This paper presents the use of the continuous adjoint method, developed by NTUA [8, 9, 10, 11] and implemented within OpenFOAM, for the shape optimization of a passenger car defroster nozzle, including experimental validation performed at Toyota Motor Europe (TME).

To set–up the optimization problem, the shape of a reference defroster nozzle is morphed using a volumetric B–splines–based morphing tool; the same tool undertakes the deformation of the CFD mesh at each optimization cycle. The CFD analysis is based on RANS, using the k-ϵ turbulence model. The optimization loop uses the gradient of the objective function w.r.t. the coordinates of the volumetric B–splines lattice, computed using the adjoint method.

Experimental tests to measure the actual velocity pattern on the windshield including velocity measurements with a hot–wire anemometer, were performed at the TME premises, in Belgium. A convincing comparison between CFD analysis and measurements is presented at first. Then, the improved demisting performance of the nozzle geometry resulted from the adjoint optimization is confirmed on the basis of experiments; rapid prototyping techniques were used to manufacture the optimized nozzle.

2 CFD & EXPERIMENTAL ANALYSIS

The flow problem is governed by the RANS equations for steady-state, incompressible flow, coupled with a turbulence model. The mean flow equations are:

$$R^p = -\frac{\partial v_j}{\partial x_j} = 0 \tag{1}$$

$$R^v_i = v_j \frac{\partial v_i}{\partial x_j} - \frac{\partial}{\partial x_j}\Big[(\nu + \nu_t)\Big(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\Big)\Big] + \frac{\partial p}{\partial x_i} = 0 \qquad i = 1, 2, 3 \tag{2}$$

where v_i are the velocity components, p stands for static pressured divided by the constant density ρ, ν is the constant bulk viscosity and ν_t is the turbulent viscosity. Turbulent viscosity results from the solution of the k-ϵ model, [7], with wall functions near the solid walls.

The flow equations are solved in the computational domain shown in fig. 1. The blower of the HVAC unit provides the defroster nozzle with a high temperature airflow which, in turn, directs high velocity air jets towards the windshield. The air flows through the whole cabin of the car and exits through a flap located in the rear of the passenger compartment, to the outside. Thus, the inlet (S_I) to the domain corresponds to the inlet to the defroster nozzle, the outlet (S_O) is a rectangular patch located in the rear part of the cabin and solid walls (S_W) are the duct sidewalls and the internal surface of the cabin.

At S_I, Dirichlet boundary conditions are imposed on v_i, k and ϵ and a zero Neumann condition on p. At S_O, a zero Dirichlet condition is imposed on p together with zero Neumann ones on v_i, k and ϵ. No–slip boundary is imposed on the solid walls S_W, where k, ϵ are computed using wall functions and a zero Neumann condition is imposed on p.

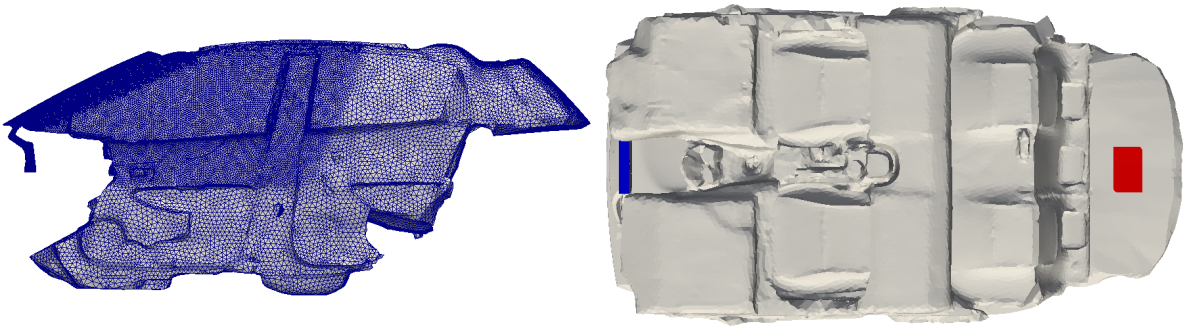


Figure 1: Side (left) and bottom (right) view of the computational domain Ω. On the left, the surface mesh is shown, highlighting the fact that the mesh is much finer inside the nozzle and in the front part of the cabin. S_I is in blue and corresponds to the inlet to the defroster nozzle. S_O, in red, is close to the trunk. The solid walls S_W correspond to the remaining surfaces of the cabin.

The flow equations are solved using the SIMPLE algorithm and a cell–centered finite–volume discretization scheme, on unstructured grids. The solution of the turbulence model PDEs is decoupled. Convection terms are discretized using a second-order upwind scheme, whereas for the computation of spatial gradients, the Green–Gauss theorem is used.

Before proceeding to the optimization, it is necessary to solve the flow problem, in the domain explained above, for the reference defroster nozzle. It should be noted that CAD data was only used for parts where high accuracy is necessary (defroster nozzle, windshield, instrument

panel close to the outlet of the nozzle, mirror). Already available laser scanned surface data was used for the remaining parts of the car cabin. Mesh refinement boxes were defined to achieve high accuracy where needed while balancing the overall computational cost, [5]. The mesh, provided by BETA CAE Systems to TME, is fine in the front cabin and coarser in the rear.

The Reynolds number of the flow is approximately 20000, based on the inlet hydraulic diameter. The flow fields over a cross section can be seen in fig. 2. The streamlines near the windshield can be seen in fig. 3.

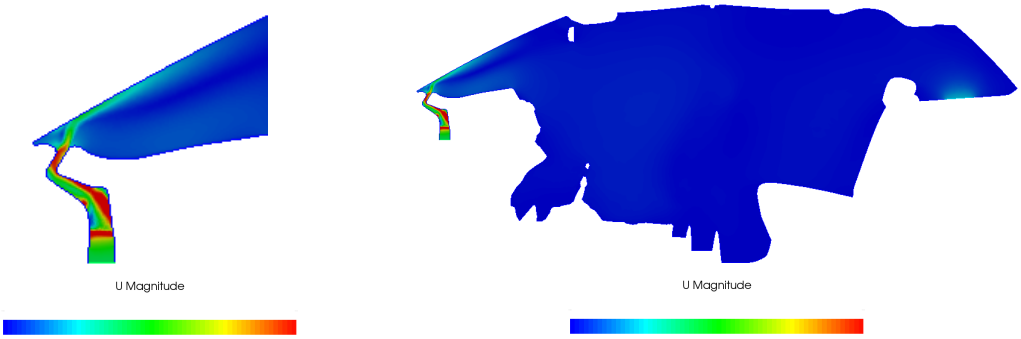


Figure 2: Velocity distribution across the symmetry plane; focus on the nozzle and windshield (left), the entire computational domain (right). The velocity magnitude is almost zero in the car cabin while, in the duct and close to the windshield, it reaches quite high values.

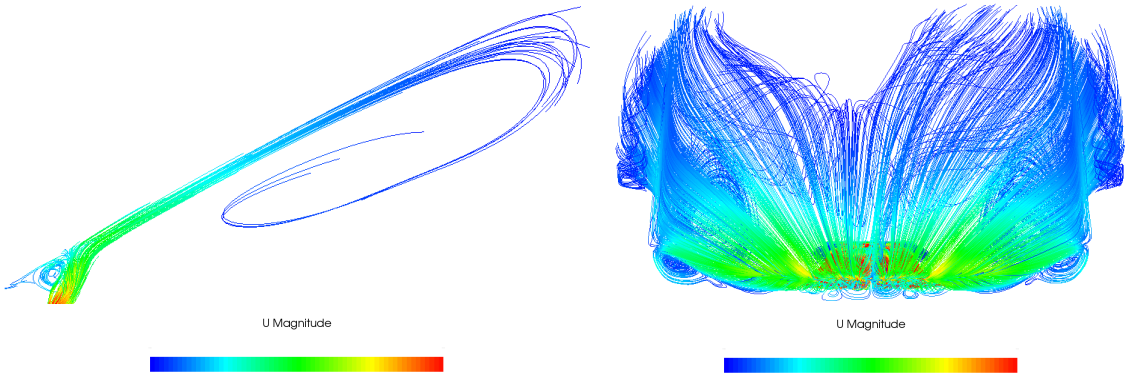


Figure 3: Streamlines emitted from seed points at the defroster inlet indicate the presence of small vortices of low velocity air at the bottom of the windshield, below the level where the jet flow starts to be attached to the windshield. The jet flow stays attached almost up to the level of the rear view mirror where recirculation occurs.

2.1 Experimental measurements–comparison

To verify the validity of the CFD results in the vicinity of the defroster nozzle jet flow and the internal surface of the windshield, [2, 3], a grid with 100mm spacing was drawn on the windshield, see fig. 4. The HVAC blower was controlled by constant voltage with an external power supply ensuring that the defroster nozzle provides constant airflow. Velocity was measured at each grid point using the hot–wire anemometer; all measurements were performed at points located on a surface 7mm away from the windshield, see also [2].

Figure 4: Set–up of the velocity pattern measurement. On the left, the 100mm–spaced grid is shown while, on the right, the use of the hot–wire anemometer on one of the grid points is demonstrated.

Measured and computed velocity patterns are compared in fig. 5. They are both non–symmetrical because the instrument panel is so. On the driver's side, the meter close to the steering wheel is creating this asymmetry that is affecting the flow exiting from the defroster nozzle, towards the windshield. Asymmetry is less intense on the CFD velocity pattern, since the HVAC blower was considered to provide uniform flow, which does not happen in reality. Regarding the measured velocity pattern, flow disturbance effects caused by the use of the measuring tool may have a small impact too. Moreover, during the measurement, the side defroster and face outlets were taped, so some leakage might be unavoidable. Last but not least, in a real car, air leakage occurs through the ducts and other parts, so the mass flow made available to the windshield through the defroster is lower than that provided from the blower, which was imposed as input to the CFD simulation.

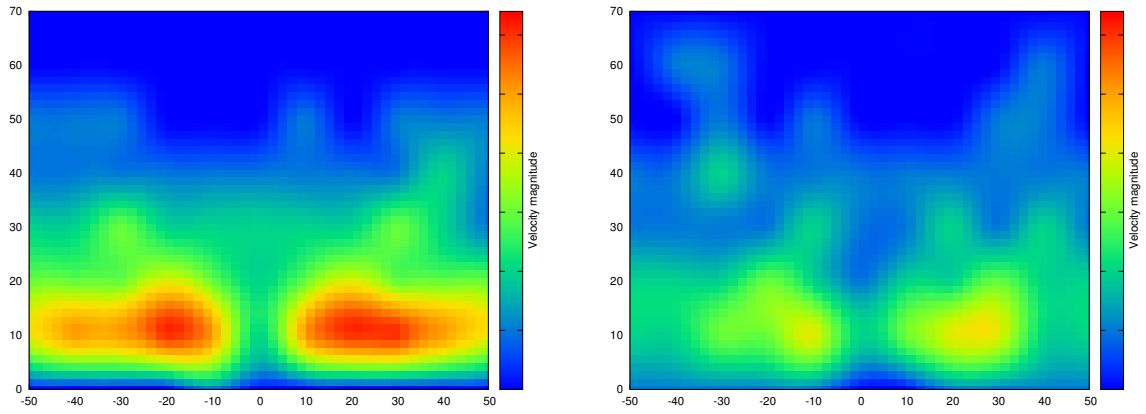


Figure 5: Comparison between CFD–predicted (left) and measured (right) velocity distributions close to the windshield. The axes are indicating the $2D$ grid used for the measurement (see fig. 4). CFD results and measurements have been similarly interpolated for better visualization. CFD analysis and measurements show similar trends.

The velocity pattern provided by the CFD run (post-processed on the basis of the computed velocity magnitude at the measurement points) gives a pattern qualitatively similar to the measured one. Quantitative differences between them can be due to the aforementioned reasons. Overall, the CFD simulation is considered to provide an acceptable flow prediction which allows to proceed to the adjoint–based optimization.

3 FORMULATION OF THE OPTIMIZATION PROBLEM–THE ADJOINT SOLVER

3.1 Objective function

It is desirable that the air velocity pattern close to the windshield meets some performance criteria pertinent to the improvement of the demisting and defrosting operation. Practically, a uniform velocity distribution close to the windshield is targeted. Given that the upper part of the windshield is dominated by low–velocity fluid flow, the objective function is confined to this upper half (hereafter, to be denoted by Ω_{tar}, figs. 6 and 7). There, target velocity v_{tar} should be higher than the flow velocity obtained with the reference defroster nozzle geometry. The objective function F is given by

$$F = \frac{1}{2}\int_{\Omega_{tar}} (v_i^2 - v_{tar}^2)^2 d\Omega \tag{3}$$

where Ω_{tar}, fig. 6 is referred to as the target volume.

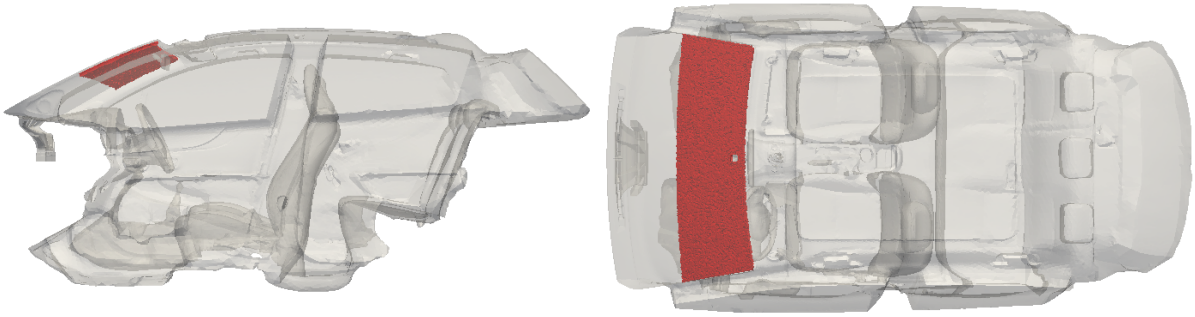


Figure 6: The thin target volume Ω_{tar} where the objective function is defined is marked in red, inside the computational domain Ω.

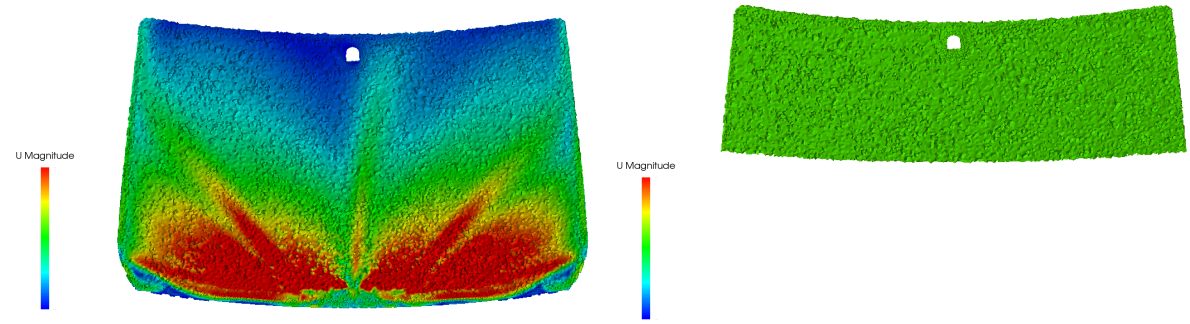


Figure 7: Velocity pattern computed by the CFD software (left). The velocity magnitude contours shown are 7mm away from the internal surface of the windshield. The air velocity distribution indicates low velocity areas in blue and high velocity areas in red. On the right, on the same color scale (indicating v_{tar}), the target volume (Ω_{tar}) is shown.

3.2 Optimization algorithm

To perform an automated CFD shape optimization loop for the defroster nozzle, the adjoint solver coupled with an in-house morpher was used, [11]. The shape morpher is based on volumetric B–splines [11], acting as a Free Form Deformation (FFD) method. The coordinates of the control points of the volumetric B–splines are the optimization variables b_n.

To start, the $3D$ morphing box enclosing the nozzle geometry to be optimized, the number of control points and the degree of the basis functions of the volumetric B–splines are defined. A structured control grid is generated. CFD mesh points residing within the control grid are identified and parameterized; by doing so, these can be displaced in conformity to the displacement of the control points during the optimization process. The flow and adjoint equations are solved and the objective function gradient w.r.t. b_n is computed. Having computed $\frac{\delta F}{\delta b_n}$, the control point coordinates are updated via steepest descent. The CFD mesh is also updated according to the nodal parametric coordinates which are assumed as constant during morphing. The previous steps are repeated until a termination criterion is met.

3.3 Adjoint equations & boundary conditions

The adjoint problem, which leads to the computation of the sensitivity derivatives consists of the adjoint mean flow equations

$$R^q = -\frac{\partial u_j}{\partial x_j} = 0 \tag{4}$$

$$R_i^u = u_j\frac{\partial v_j}{\partial x_i} - \frac{\partial (v_j u_i)}{\partial x_j} - \frac{\partial}{\partial x_j}\Big[(\nu+\nu_t)\Big(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\Big)\Big] + \frac{\partial q}{\partial x_i} + \underbrace{2(v_j^2 - v_{tar}^2)v_i}_{\textit{only in } \Omega_{tar}} = 0 \qquad i = 1,2,3 \tag{5}$$

where u_i is adjoint velocity and q is adjoint pressure as well as the adjoint boundary conditions, as explained in detail in [8, 9].

After satisfying the adjoint mean flow equations and their boundary conditions, the sensitivity derivatives are given by

$$\frac{\delta F}{\delta b_n} = -\int_{S_{W_P}} \Big[(\nu+\nu_t)\Big(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\Big)n_j - q n_i\Big]\frac{\partial v_i}{\partial x_k}\frac{\delta x_k}{\delta b_n}dS + \int_{S_{W_P}} (u_i R_i^v + q R^p)\frac{\delta x_k}{\delta b_n}dS \tag{6}$$

where, in the first term inside the brackets, there are only flow and adjoint variables while the outside part comes from the differentiation of the geometry, as computed by the morphing software. For the differentiation of the turbulence model, the reader should refer to [8, 9]

4 NUMERICAL OPTIMIZATION & VALIDATION

Before proceeding to the shape optimization of the defroster nozzle, the adjoint equations were solved in a domain associated with the reference geometry and the sensitivities of the objective function, eq. 6, w.r.t. the normal displacement of all surface mesh nodes were computed. These are shown in fig. 8, in the form of the so–called sensitivity map.

4.1 Optimization results

Several optimization runs were performed until the most suitable new shape of the defroster nozzle was obtained. The need to perform several optimization runs, instead of a single only, comes from the necessity of selecting the appropriate parameterization setup that gives more optimization potential, by also considering that the nozzle should be manufacturable and fit inside the assembly of its neighbouring parts. Manufacturing and topology constraints were manually taken into consideration during the optimization.

The designed (optimized) defroster nozzle yields 43% drop in the objective function and has the following characteristics as far as the parameterization set–up is concerned. Firstly, the

Figure 8: Sensitivity map indicating the change in the objective function F caused by the normal displacement of the boundary faces. Blue–colored areas should be pulled outwards while red areas should be pushed inwards so as to decrease the value of F.

control points were allowed to move only in the x–direction. Moreover, to get a quite smooth shape, the displacements per iso–plane were averaged, see fig. 10. In other words, all control points laying on the same iso–x plane are displaced in the x–direction using averaged sensitivity derivatives.

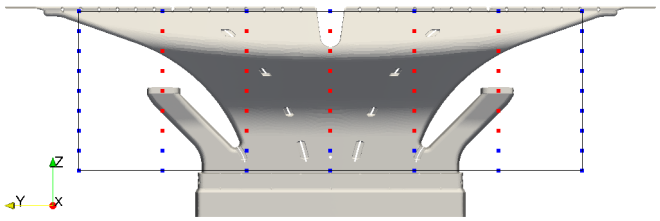


Figure 9: Control box ($3 \times 7 \times 9$) with active (red) and frozen (blue) control points. The frozen control points help to avoid mesh overlapping between the parameterized and non-parameterized areas. In this case, the side, top and two bottom rows of the control points are kept frozen.

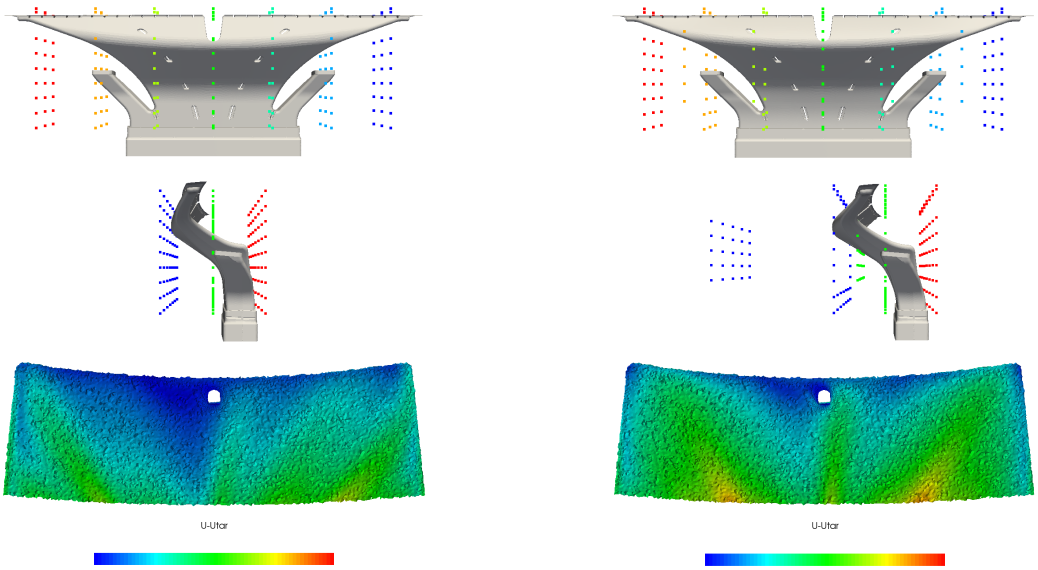


Figure 10: Reference (left) and optimized (right) defroster nozzle geometries and velocity patterns. In each row, the control grid nodes are colored based on a different coordinate. The final shape is very smooth, due to the averaging per iso–plane of the displacements of the control points and seems to be suitable for mass production. The field shown in the last row is $v - v_{tar}$ in which green areas correspond to areas where v_{tar} was reached, blue to areas with lower air velocity and red to areas with greater air velocity than the target. The comparison between the two patterns shows improvement in the coverage of the upper part of the windshield as well as increased uniformity. The target is practically reached over the majority of the cells of the target volume.

4.2 Defrost test of the optimized geometry & comparison

The optimized geometry of the defroster nozzle was manufactured using a rapid prototyping technique ($3D$ printing). Then, it was placed in the test vehicle, replacing the reference defroster nozzle and submitted to a defrost test. Windshield defrost patterns were obtained from cold room testing [3]. For the purpose of comparison, the defrosting and demisting efficiency of both the reference and optimized defroster nozzles have been tested.

To reproduce cold start condition, the vehicle soaked for several hours at a temperature of $-20°C$, in TME's climatic chamber. Following the soak, a high amount of humidity was generated in the cabin for a few minutes. Then, the defrost test commenced. The patterns were recorded regularly and the windshield was marked from the inside to indicate clearance areas.

The melting pattern for the reference and the improved defrosters, at two different time instants, are shown in fig. 11. At every instant recorded, the new defroster nozzle geometry gives a bigger clearance zone compared to the initial one. At the end, the new geometry is proved to be capable of clearing the windshield completely in 15% less time than the reference defroster nozzle.

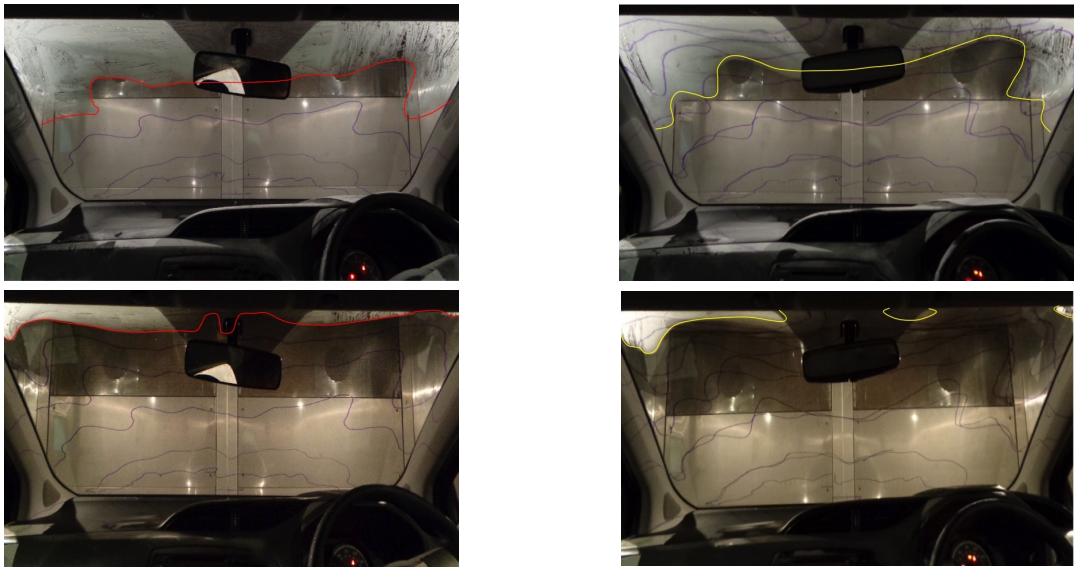


Figure 11: Defrost test of the reference (left) and improved (right) defroster nozzle shapes. The melting pattern is recorded and marked on the internal surface of the windshield. The patterns in the above figures are compared at the same time instants, proving the better defrosting ability of the optimized shape. The windshield with the optimized defroster nozzle shape is completely clear in 15% less time compared to the reference shape.

5 CONCLUSIONS

In this paper, shape optimization using the continuous adjoint method and a morphing software, both developed by the PCOpt Unit of NTUA, was applied to the defroster nozzle, part of the HVAC unit of a Toyota passenger car. A new defroster nozzle shape resulted, by maximizing the magnitude (according to a preset target value) and uniformity of the air velocity distribution on the upper–half of the windshield, leading improved defrosting performance of the vehicle. The optimized geometry, that was manually forced to comply with manufacturing and topological constraints, was manufactured with a rapid prototyping technique, placed in the vehicle and submitted to a defrost test that validated its improved defrosting performance. The optimized defroster nozzle achieves defrosting in 15% less time than the reference one.

ACKNOWLEDGMENTS

The contribution of BETA CAE Systems engineers in the pre–processing phase of the simulation is acknowledged.

REFERENCES

[1] T. Kohnotou, Y. Iwamoto, K. Hoshiawa, M. Nagataki, *Optimum Design of Defroster Nozzle.* SAE Technical Paper, 920167, 1992.

[2] B. S. AbdulNour, *Hot–Wire Velocity Measurements of Defroster and Windshield Flow.* Ford Motor Company, SAE Technical Paper, 970109, 1997.

[3] B. S. AbdulNour, *CFD Prediction of Automotive Windshield Defrost Pattern.* Ford Motor Company, SAE Technical Paper, 1999-01-1203, 1999.

[4] W. Yang, W. Shi, F. Guo, W. Yang, *Flow Field Simulation and Performance Analysis of HVAC Defrosting Duct.* 2nd International Conference on Electronic & Mechanical Engineering and Information Technology (EMEIT-2012), China, 2012.

[5] I. Goldasteh, S. Chang, S. Maaita, G. Mathur, *Numerical Simulation of Airflow Distribution on the Automobile Windshield in Defrost Mode.* SAE Technical Paper 2015-01-0330, doi:10.4271/2015-01-0330, 2015.

[6] J. Park and C. Kim, *Parametric Study on Automotive Windshield Defrost Pattern using CFD.* Hyundai MOBIS, SAE Technical Paper, 2003-01-1078, 2003.

[7] S.Patankar, D.Spalding, *Calculation Procedure for Heat, Mass and Momentum Transfer in Three–Dimensional Parabolic Flows.*, Int.J.Heat Mass Transfer Vol.15 pp.1787-1806, 1972.

[8] E.M. Papoutsis–Kiachagias, *Adjoint Methods for Turbulent Flows, Applied to Shape or Topology Optimization and Robust Design.* PhD Thesis, NTUA, 2013.

[9] E.M. Papoutsis–Kiachagias, K.C. Giannakoglou, *Continuous Adjoint Methods for Turbulent Flows, Applied to Shape and Topology Optimization: Industrial Applications.* Archives of Computational Methods in Engineering, DOI 10.1007/s11831-014-9141-9, 2014.

[10] E.M. Papoutsis–Kiachagias, A.S. Zymaris, I.S. Kavvadias, D.I. Papadimitriou, , K.C. Giannakoglou *The Continuous Adjoint Approach to the k–ϵ Turbulence Model for Shape Optimization and Optimal Active Control of Turbulent Flows.* Engineering Optimization, 47 (3), pp. 370-389, 2015.

[11] E.M. Papoutsis–Kiachagias, N. Magoulas, J. Mueller, C. Othmer, K.C. Giannakoglou, *Noise Reduction in Car Aerodynamics using a Surrogate Objective Function and the Continuous Adjoint Method with Wall Functions.* Computers and Fluids, 122, pp. 223-232, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SENSITIVITY ANALYSIS FOR FORCED RESPONSE IN TURBOMACHINERY USING AN ADJOINT HARMONIC BALANCE METHOD

Anna Engels-Putzka and Christian Frey

German Aerospace Center (DLR), Institute of Propulsion Technology
Linder Höhe, 51147 Cologne, Germany
e-mail: {Anna.Engels-Putzka,Christian.Frey}@dlr.de

Keywords: Aeroelasticity, Forced Response, Adjoint Methods, Harmonic Balance

Abstract. *The analysis of aeroelastic properties is an important aspect in the design of turbomachinery components. In this study we focus on vibrations caused by the interaction of adjacent blade rows (forced response). This is an inherently unsteady phenomenon. But due to its periodic nature it can be efficiently treated by numerical methods formulated in the frequency domain, e.g. the harmonic balance method.*

When going from the analysis of individual designs using CFD to CFD-based optimisation it is desirable to compute also sensitivities of objective functions (targets and restrictions for the optimisation) with respect to design parameters. Since in typical applications the number of design parameters is much larger than the number of objective functions, it is advantageous to use the adjoint method for the computation of these sensitivities.

An adjoint solver based on the harmonic balance method has been implemented in the framework of the flow solver TRACE. This is now extended and used to compute the sensitivities of aeroelastic objective functions to the amplitudes of a harmonic perturbation at the entry or exit of the respective blade row. These sensitivities can be validated by comparing to finite differences obtained from harmonic balance computations with different perturbations. We apply the method to model problems which are representative for turbomachinery configurations.

1 INTRODUCTION

The design of turbomachinery components is a complex and multi-disciplinary task, and computational methods are more and more extensively used in the design process. Often a compromise between conflicting design goals has to be found. For example, in aircraft propulsion it is on the one hand desired to build compact and light-weight machines, on the other hand also structural properties have to be taken into account and a reduced axial gap between blade rows leads to stronger interactions which could cause unwanted blade vibrations (forced response). Therefore, besides CFD simulations to assess the aerodynamic performance of single blade rows or entire components, also aeroelastic computations are important.

Aeroelastic phenomena in turbomachinery are inherently unsteady, but often of a periodic nature, which makes them attractive for the application of frequency domain methods. These are in most cases computationally more efficient than unsteady methods in time-domain. While some phenomena, e.g. blade flutter, can often be adequately simulated by time-linearized unsteady methods, e.g. the linear harmonic (LH) method [1], there are also important nonlinear effects, in particular for forced response (cf. [2]). An overview on the use of frequency domain methods in turbomachinery applications has been given by He [3] and in [4] Ekici and Huang compare time- and frequency-domain methods.

In our work we focus on the harmonic balance (HB) method as proposed by Hall [5]. It has been applied by Thomas et al. [6, 7] to analyse nonlinear effects in aeroelastic configurations. The same approach has been used by Huang and Ekici [8, 9, 10] and also Gopinath et al. developed a similar time-spectral method [11, 12]. In [13] different variants of HB are reviewed and several applications are presented. The form where the harmonic balance equations are formulated in the frequency domain is termed the nonlinear frequency domain (NLFD) form. It is used e.g. by McMullen et al. [14, 15, 16]. The implementation of a harmonic balance method in the DLR flow solver TRACE, which also uses the NLFD approach, and its application to aeroelastic analysis are described in [17] and [18].

Besides the numerical evaluation of given designs it is – due to increased computational resources – also possible to explore large ranges of designs in CFD-based automatic optimisations. To obtain efficient optimisation algorithms it is useful to also have access to derivatives of the functions which are employed as objectives or constraints in the optimization. There are different possibilities to compute such derivatives. The simplest is to use finite differences, but this requires at least one additional function evaluation for each partial derivative, which can be very expensive. Alternatively, derivatives can be computed directly by a linearisation of the applied computational methods. But since in typical applications the number of design parameters is large compared to the number of objectives, the application of adjoint methods [19, 20] is often more efficient than direct linearisation.

While steady adjoint methods are nowadays quite frequently used in the field of turbomachinery design (see e.g. [21, 22]), the application of unsteady adjoint CFD is very limited due to its exceedingly high computational costs, see e.g. [23, 24, 25] and references therein.

A number of adjoint solvers for frequency domain methods have already been developed. Discrete adjoint solvers for LH have been derived “by hand”, i.e. by analytic differentiation of the primal code, by Florea and Hall [26] for the Euler equations and by Duta et al. [27, 28] for the RANS equations. Since this approach is very cumbersome for nonlinear methods, many authors apply algorithmic differentiation (AD) techniques in the development of adjoint frequency domain solvers, e.g. Thomas et al. [29] for the HB method of Hall, Huang and Ekici [30] for their implementation of this method, or Mader and Martins [31] for the time-spectral

method of Gopinath et al. An implementation of an adjoint for the non-linear frequency domain method of McMullen et al. is due to Nadarajah and Jameson [32, 33]. An adjoint method for HB equations for one harmonic in the time-domain together with an application to a combined aerodynamic and aeroelastic blade optimisation has been presented by He and Wang [34].

The steady adjoint solver in TRACE [35] also employs a discrete adjoint approach, but the linearisation of the residual is obtained by computing finite differences. This approach has now been extended to the frequency domain methods harmonic balance and linear harmonic. The derivation and implementation of the adjoint methods used in this study has been described in detail in [36]. In Section 2 we recall the basic equations and present some recent extensions of the implementation. Afterwards, we apply the methods to two different test cases and compare the results to finite differences from the primal solver (see Section 3).

2 ADJOINT FREQUENCY DOMAIN METHODS

2.1 Harmonic balance method

The harmonic balance method as implemented in TRACE (see [17, 18]) approximates the time dependent flow solution as a Fourier series with a finite number of harmonics[1]

$$q(x,t) = \mathrm{Re}\left[\sum_{k=0}^{K} \hat{q}_k(x) e^{ik\omega t}\right]. \tag{1}$$

The Fourier coefficients $\hat{q}_k$ are determined by the set of equations

$$ik\omega\hat{q}_k + \widehat{R(q)}_k = 0, \quad k = 0, \ldots, K, \tag{2}$$

where the Fourier components of the residual are approximated as

$$\widehat{R(q)}_k \approx \mathscr{F}(R(\mathscr{F}^{-1}(\hat{q})))|_k, \tag{3}$$

with $\mathscr{F}$ denoting the discrete Fourier transform (DFT). This means that the solution q in the time domain is reconstructed at a finite number of sampling points, then the residual is evaluated at these points, and the resulting vector is transformed back into the frequency domain.

The characteristic of the NLFD approach is that the variables which are solved for are the Fourier coefficients $\hat{q}$, while the flow states at the sampling points in time are only computed as intermediate quantities during the solution process. In contrast, in the time-spectral approach, the states at the sampling points are stored and the equations are formulated and solved completely in the time domain.

2.2 Discrete adjoint approach

We are now interested in the sensitivity of some functional I, which is computed from the solution q, to some parameter α. We assume that I depends on α only through the Fourier coefficients of q, i.e.

$$\frac{dI}{d\alpha} = \frac{\partial I}{\partial \hat{q}} \frac{d\hat{q}}{d\alpha}. \tag{4}$$

[1] In general it could be a combination of such series for several base frequencies, but we restrict the discussion here to a single base frequency, since the adjoint solver has only been developed for this case.

To obtain an equation for $\frac{d\hat{q}}{d\alpha}$ we differentiate (2) with respect to α. Defining the harmonic balance residual as $R_{\mathrm{HB},k} := ik\omega\hat{q}_k + \widehat{R(q)}_k$, we get

$$0 = \frac{dR_{\mathrm{HB},k}}{d\alpha} = \sum_j \left(ik\omega\delta_{jk} + \frac{\partial \widehat{R(q)}_k}{\partial \hat{q}_j} \right) \frac{d\hat{q}_j}{d\alpha} + \frac{\partial R_{\mathrm{HB},k}}{\partial \alpha}. \tag{5}$$

This is a linear system of equations and we write it briefly as

$$A\frac{d\hat{q}}{d\alpha} = -\frac{\partial R_{\mathrm{HB}}}{\partial \alpha}. \tag{6}$$

The corresponding adjoint system of equations is then

$$A^*\hat{\psi} = \left(\frac{\partial I}{\partial \hat{q}} \right)^* \tag{7}$$

with the adjoint system matrix given by

$$A^* = \left(\mathscr{F}^{-1}\right)^* \left(\mathrm{diag}\left(\frac{\partial R}{\partial q}\Big|_{q(t_j)} \right) \right)^* \mathscr{F}^* - \mathrm{diag}(ik\omega), \tag{8}$$

where $\mathrm{diag}(\ldots)$ denotes a block diagonal matrix with the corresponding entries on the diagonal and t_j are the sampling points for the DFT. The adjoint operator in the frequency domain is defined with respect to the scalar product

$$\langle \hat{\psi}, \hat{q} \rangle = \mathrm{Re}\langle \hat{\psi}, \hat{q} \rangle_{\mathbb{C}}. \tag{9}$$

This means that we consider the set of (complex) Fourier coefficients as element in a real vector space, which is due to the fact that the transformations $\mathscr{F}$ and $\mathscr{F}^{-1}$ are not linear over the complex numbers.

Using (7), the expression for the sensitivities (4) can be rewritten as

$$\frac{dI}{d\alpha} = -\hat{\psi}^* \frac{\partial R_{\mathrm{HB}}}{\partial \alpha}, \tag{10}$$

which is now independent of $\hat{q}$.

2.3 Solution method and implementation

As in the steady adjoint solver in TRACE (see [35]) the discrete adjoint equations (7) are solved by a preconditioned GMRes (Generalised Minimal Residual) algorithm with restarts, as described in [36]. Due to the definition of the scalar product (cf. Eqn. (9)) all "internal" data for the GMRes algorithm is stored as real vectors and matrices. As preconditioner we use either SSOR or ILU(p), i.e. incomplete LU decomposition with variable fill level [37].

The sampling points for the DFT are chosen to be equidistant, i.e. we have $t_j = \frac{2\pi j}{\omega N}$ if ω is the base frequency and N the number of sampling points. To resolve K higher harmonics, at least $N = 2K+1$ points are needed. But due to the nonlinearity of the residual function, the problem of aliasing can occur [17]. Therefore a higher number of sampling points should be used. Since the residual is a rational function of the flow variables, it is not possible to choose the number of sampling points such that aliasing can be avoided completely. In our computations we take $N = 4K + 1$, which is sufficient to resolve cubic terms, so that the remaining errors are of fourth order and should be small in most applications.

The residual Jacobians at the sampling points are computed analogous to the steady case, i.e. finite differences are used for the fluxes while the source terms are differentiated analytically.

2.4 Functionals

The right hand side of (7) is given by the linearisation of the objective functional w.r.t. the state variables. Functionals which are relevant for aeroelastic applications are defined on the blade surfaces. We consider here the modal force for a mode Φ and frequency ω, which is given by

$$F_{\mathrm{mod}} = \langle \Phi, \hat{p}_\omega \vec{n} \rangle_{L^2} = \int_\Gamma \langle \Phi, \hat{p}_\omega \vec{n} \rangle \, dS \tag{11}$$

where Γ is the blade surface, $\vec{n}$ the surface normal and $\hat{p}_\omega$ the Fourier coefficient of the pressure on the blade for the frequency ω. In the context of this study we assume that Φ and $\vec{n}$ are fixed, so that the linearisation of F_{mod} depends only on $\hat{p}_\omega$:

$$dF_{\mathrm{mod}} = \frac{\partial F_{\mathrm{mod}}}{\partial \hat{p}_\omega} d\hat{p}_\omega. \tag{12}$$

Instead of the modal force itself, which is a complex-valued function, we consider its real and imaginary parts as functionals. In the discrete context, the integral in (11) is evaluated as a sum over the faces on the blade surface:

$$F_{\mathrm{mod}} = \sum_\sigma F_{\mathrm{mod},\sigma} \quad \text{with} \quad F_{\mathrm{mod},\sigma} = \hat{p}_{\omega,\sigma} \left\langle \Phi_\sigma, \vec{A}_\sigma \right\rangle_{\mathbb{C}^3}, \tag{13}$$

where $\vec{A}_\sigma$ is the face normal vector with $|\vec{A}_\sigma|$ equal to the area of the face σ. In the following we drop the subscripts σ and ω for simplicity. The real and imaginary parts of F_{mod} can then be written as

$$\begin{aligned} \mathrm{Re}(F_{\mathrm{mod}}) &= \mathrm{Re}(\hat{p}) A_{\Phi,\mathrm{Re}} + \mathrm{Im}(\hat{p}) A_{\Phi,\mathrm{Im}} && (14) \\ \mathrm{Im}(F_{\mathrm{mod}}) &= \mathrm{Im}(\hat{p}) A_{\Phi,\mathrm{Re}} - \mathrm{Re}(\hat{p}) A_{\Phi,\mathrm{Im}}, && (15) \end{aligned}$$

where $A_{\Phi,\mathrm{Re}}$ and $A_{\Phi,\mathrm{Im}}$ are the (real) scalar products of the real and imaginary parts of Φ with $\vec{A}$, respectively.

For the derivative with respect to the Fourier coefficients of q, which is needed for the right hand side of the adjoint equations, we get (compare (12))

$$\frac{d\mathrm{Re}(F_{\mathrm{mod}})}{d\hat{q}} = \left(\frac{\partial \mathrm{Re}(F_{\mathrm{mod}})}{\partial \mathrm{Re}(\hat{p})} \quad \frac{\partial \mathrm{Re}(F_{\mathrm{mod}})}{\partial \mathrm{Im}(\hat{p})} \right) \frac{d\hat{p}}{d\hat{q}} = \left(A_{\Phi,\mathrm{Re}} \frac{d\hat{p}}{d\hat{q}} \quad A_{\Phi,\mathrm{Im}} \frac{d\hat{p}}{d\hat{q}} \right) \tag{16}$$

and analogously

$$\frac{d\mathrm{Im}(F_{\mathrm{mod}})}{d\hat{q}} = \left(-A_{\Phi,\mathrm{Im}} \frac{d\hat{p}}{d\hat{q}} \quad A_{\Phi,\mathrm{Re}} \frac{d\hat{p}}{d\hat{q}} \right), \tag{17}$$

where we have used the fact that $\frac{d\hat{p}}{d\hat{q}}$ is real.

2.5 Sensitivity evaluation

To analyse the right hand side of (10) we assume that the flow field q depends on the parameter α only through the prescribed boundary values q_{bv}, i.e. the flow conditions at entry and exit boundaries of the computational domain. Then we get

$$\frac{dI}{d\alpha} = -\hat{\psi}^* \frac{\partial R_{\mathrm{HB}}}{\partial \alpha} = \left(-\hat{\psi}^* \frac{\partial R_{\mathrm{HB}}}{\partial q^{\mathrm{ext}}} \right) \left(T_{\mathrm{bv}} \frac{dq_{\mathrm{bv}}}{d\alpha} \right) \tag{18}$$

where q^{ext} denotes the ghost cells and T_{bv} is the inhomogeneous part of the linearised boundary operator, i.e. the linearisation of the boundary update operator

$$\mathcal{T} : (q^{\mathrm{int}}, q_{\mathrm{bv}}) \mapsto q^{\mathrm{ext}} \tag{19}$$

with respect to q_{bv}. The multiplication in (18) is a scalar product, which is defined here – as in (9) – as the real scalar product, i.e. the real part of the standard scalar product for complex vectors.

The implementation of the sensitivity evaluation is similar to the computation of boundary value sensitivities for the steady adjoint solver which is discussed in [38]. The term $\hat{\psi}^* \frac{\partial R_{\mathrm{HB}}}{\partial q^{\mathrm{ext}}}$ from (18) is evaluated as $\left(\left(\frac{\partial R_{\mathrm{HB}}}{\partial q^{\mathrm{ext}}}\right)^* \hat{\psi}\right)^*$ by applying the (adjoint) system matrix to the adjoint solution. The "outer" adjoint is then computed implicitly in the scalar product. The other factor of the product, $T_{\mathrm{bv}} \frac{dq_{\mathrm{bv}}}{d\alpha}$, is computed by applying the (inhomogeneous) linearised boundary conditions to a zero vector. In contrast to the steady case, the boundary conditions here are the gust boundary conditions, where a harmonic disturbance with a specified frequency and (circumferential) wave number is prescribed at the entry or exit of a blade row [39]. Since the parameters α we are interested in are the components of the gust perturbation themselves, $\frac{dq_{\mathrm{bv}}}{d\alpha}$ is trivial. To obtain the result of applying the linearised boundary operator for one parameter, we set the corresponding component of the gust boundary values to unity (with the appropriate de-dimensionalisation) and use these values for the inhomogeneous part of the boundary conditions. Since the gust values are prescribed band-wise, the sensitivities are also computed for each band, which means that the contributions of all faces belonging to a given band are summed up.

2.6 Linear harmonic method

The linear harmonic method is obtained if the nonlinear residual is approximated by its linearisation about the time average $\hat{q}_0$, which is justified if the amplitudes of the harmonic perturbations ($|\hat{q}_k|$ for $k > 0$) are small. Then the different harmonics decouple and the flow state is determined by the steady equation $R(\hat{q}_0) = 0$ and a linear equation

$$\left(ik\omega + \tfrac{\partial R}{\partial q}\big|_{\hat{q}_0}\right) \hat{q}_k = 0 \tag{20}$$

for each $k > 0$ [40, 39]. Accordingly, the system matrix for the adjoint LH equations is a block-diagonal matrix with the blocks

$$A_k^* = \left(\tfrac{\partial R}{\partial q}\big|_{\hat{q}_0}\right)^* - ik\omega. \tag{21}$$

If gust boundary conditions are applied, (20) takes the form $A_k \hat{q}_k = S_k$, i.e. the boundary conditions lead to an additional source term [39]. This source term depends on the boundary value parameters α and the sensitivities for the adjoint LH method are given by the expression

$$\frac{dI}{d\alpha} = \hat{\psi}^* \frac{\partial S}{\partial \alpha}, \tag{22}$$

where $\hat{\psi}$ is the adjoint solution as before.

The implementation of the adjoint LH method is analogous to adjoint HB, but the setup and application of the system matrix are simplified. Moreover, only the time-average is needed from the primal solution. Another subtle difference is that since the problem is now complex-linear, the complex scalar product can be used and all computations – including the GMRes algorithm – are carried out over the complex numbers.

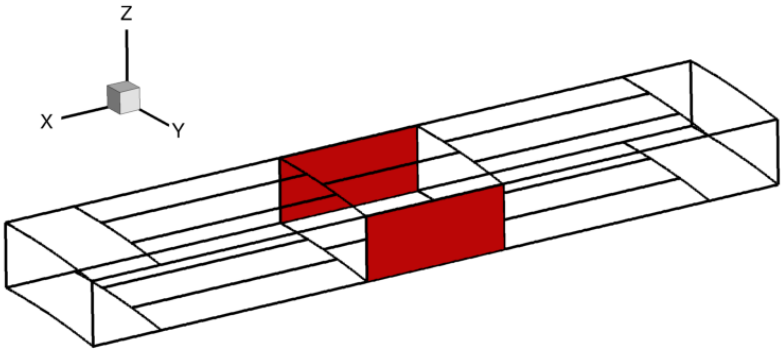


Figure 1: CAA Benchmark configuration.

3 APPLICATION

In this section we apply the described methods to single blade-row configurations with a prescribed gust perturbation at the entry surface. On the blades an (artificial) mode with the same frequency and phase shift as the gust perturbation is applied. The real and imaginary parts of the modal force for this mode are used as functionals for the adjoint solver and we evaluate the sensitivities with respect to the components $q_{\mathrm{bv},j}$ of the gust perturbation. These are the real and imaginary parts of the flow parameters at the entry, i.e. the density ρ, the velocity components (in cylindrical coordinates) u_x, u_r, u_θ and the pressure p, so that there are in total ten independent parameters.

To obtain approximate derivatives for comparison with the sensitivities computed by the adjoint solver we vary (sequentially) each component of the perturbation at the entry by a constant multiple of the corresponding reference quantity, i.e.

$$q_{\mathrm{bv},j} \mapsto q_{\mathrm{bv},j} + \delta q_{\mathrm{ref},j}, \tag{23}$$

and perform a harmonic balance (or linear harmonic) simulation with this modified gust perturbation. We choose $\delta = 10^{-4}$ as step size, but it has been verified for the used test cases that the result is not sensitive to this choice within a certain range. To compare the two methods for computing sensitivities the relative deviation of a sensitivity s is computed as

$$\Delta s = \frac{s_{\mathrm{adj}} - s_{\mathrm{FD}}}{s_{\mathrm{FD}}}. \tag{24}$$

3.1 Numerical test case

As a first test case we take a configuration that is based on the "CAA Benchmark Category 3 Problem 1" from [41], which was originally designed as a test for computational aeroacoustics methods. It consists of a cascade of (infinitely thin) flat plates (see Fig. 1). It is treated as a rotational configuration, but with a relatively large radius and therefore a high number of passages (1885). We use a scaled version of the original configuration with radius $R = 3\,\mathrm{m}$. The distance between the blades (pitch) is equal to the cord length and the computational domain extends to two times the cord length on both sides. The computational grid for one passage consists of 50000 cells ($500\times100\times1$). In circumferential direction periodic boundary conditions are applied.

The steady flow conditions are uniform with density $\rho_\infty = 1.08\,\frac{\mathrm{kg}}{\mathrm{m}^3}$ and axial velocity $u_\infty = 166\,\frac{\mathrm{m}}{\mathrm{s}}$, which corresponds to a Mach number of 0.5. At the entry plane a "vortical

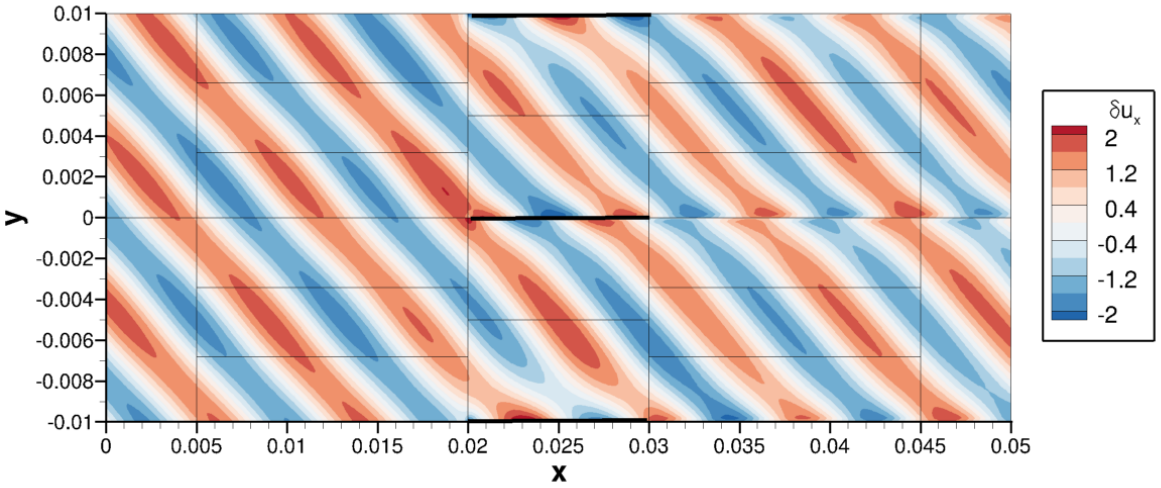


Figure 2: First harmonic of the axial velocity, reconstructed at time $t = 0$ from the harmonic balance solution including one higher harmonic.

gust" is prescribed, i.e. a perturbation of the axial and circumferential components of the velocity (real part only). It has a frequency of $20700\,\mathrm{Hz}$ and a circumferential wave number of 2356, which corresponds to a phase shift (inter-blade phase angle) of 90 degrees, and the magnitude of the perturbation is $0.01u_\infty = 1.66\,\frac{\mathrm{m}}{\mathrm{s}}$. Figure 2 visualises the flow pattern which is induced by this perturbation. For illustration the blades are drawn with a finite thickness.

The artificial mode applied on the blades in this case corresponds to a torsion. Sensitivities are only computed for the non-trivial components of the gust perturbation, i.e. the axial and circumferential velocity. The reference quantity is $u_{\mathrm{ref}} = 338.5\,\frac{\mathrm{m}}{\mathrm{s}}$.

Although the configuration is simple, the test case is numerically challenging. The convergence of the pseudo-time solver employed for the harmonic balance and linear harmonic methods is rather slow in this case and a relatively small CFL-number of 3 has to be used. The computations are stopped after 100000 pseudo-time iterations. Figure 3 shows that the values for the modal force are not entirely converged at this point, but the remaining fluctuations are much smaller than the difference between the values for two different gust perturbations.

The adjoint solver does not converge well if the SSOR preconditioner is used. Therefore we tested the newly implemented ILU preconditioner with different levels of fill-in. For lower levels (up to 4) no improvement is observed, while ILU(5) and ILU(6) work well if the restart interval is large enough. In Fig. 4 the convergence histories for different setups are compared. Based on this comparison we chose ILU(5) and a restart interval of 300 for the following computations, since this yields the fastest convergence in terms of iterations as well as with regard to simulation time. Figure 5 shows the first harmonic of the obtained adjoint flow solutions for the modal force (real and imaginary parts).

In tables 1 and 2 the sensitivities of the real and imaginary parts of the modal force with respect to the axial and circumferential velocity components of the gust perturbation computed with the adjoint solver for the harmonic balance and linear harmonic methods are compared to finite differences. The absolute value of the relative deviation is below 5% in all cases, which indicates that the implementation of the adjoint solver is in principle correct. An exact equality can not be expected, since there are differences between the primal and the adjoint solver e.g. in the implementation of the boundary conditions. For the linear harmonic computations the

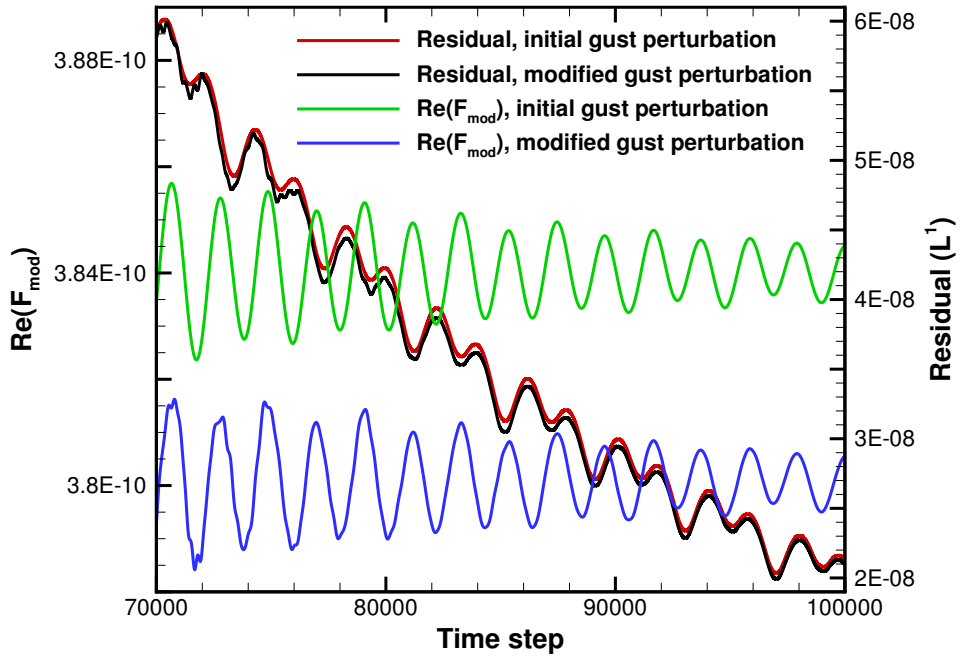


Figure 3: Convergence of the modal force for the initial gust perturbation and the perturbation with modified component $\mathrm{Re}(u_x)$ in relation to the pseudo-time solver residual.

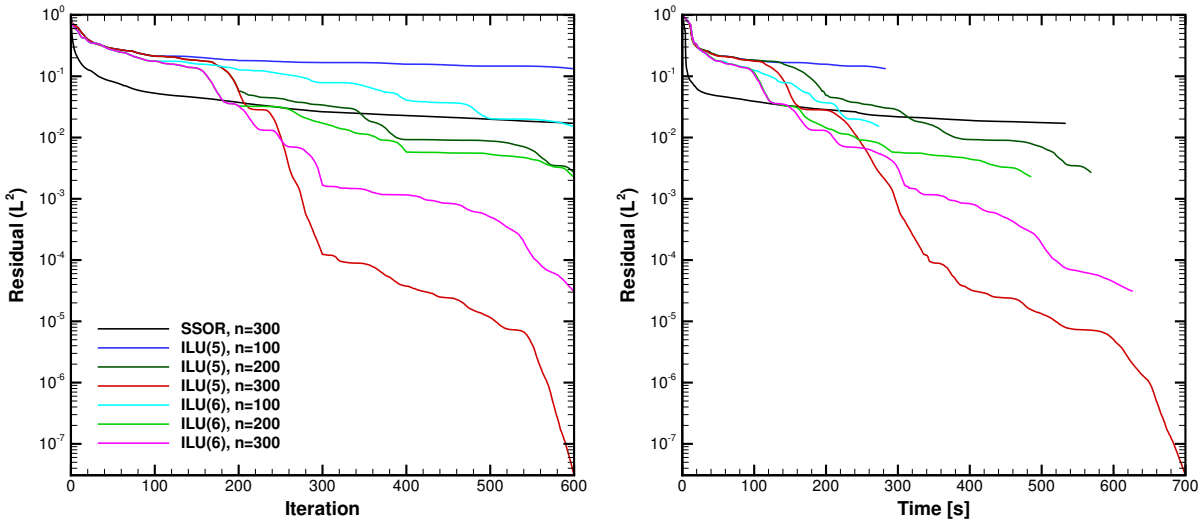


Figure 4: Convergence of the adjoint harmonic balance solver using one higher harmonic for different preconditioners and restart intervals n.

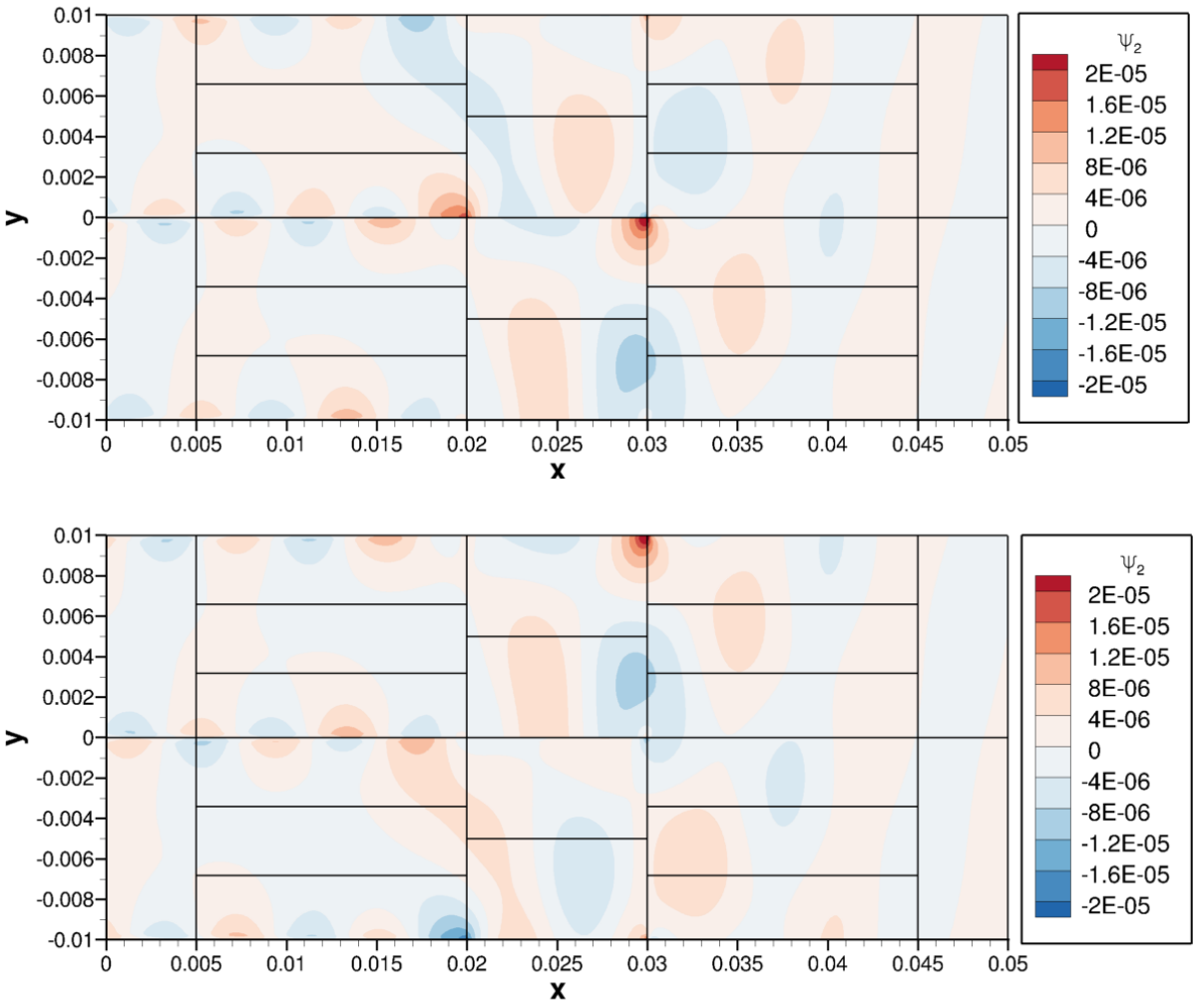


Figure 5: First harmonic of the adjoint solution for the real (top) and imaginary (bottom) parts of the modal force, reconstructed at time $t = 0$ from the adjoint harmonic balance solution including one higher harmonic. The component ψ_2 corresponds to the adjoint momentum in axial direction.

Functional	$\mathrm{Re}(F_{\mathrm{mod}})$			$\mathrm{Im}(F_{\mathrm{mod}})$		
Parameter	FD	adjoint	rel. dev.	FD	adjoint	rel. dev.
$\mathrm{Re}(u_x)$	$-1.171\cdot10^{-10}$	$-1.216\cdot10^{-10}$	3.88%	$1.813\cdot10^{-10}$	$1.863\cdot10^{-10}$	2.73%
$\mathrm{Im}(u_x)$	$-1.811\cdot10^{-10}$	$-1.859\cdot10^{-10}$	2.62%	$-1.166\cdot10^{-10}$	$-1.213\cdot10^{-10}$	4.07%
$\mathrm{Re}(u_\theta)$	$-1.165\cdot10^{-10}$	$-1.212\cdot10^{-10}$	4.00%	$1.805\cdot10^{-10}$	$1.857\cdot10^{-10}$	2.91%
$\mathrm{Im}(u_\theta)$	$-1.803\cdot10^{-10}$	$-1.857\cdot10^{-10}$	3.02%	$-1.160\cdot10^{-10}$	$-1.210\cdot10^{-10}$	4.30%

Table 1: Comparison of sensitivities for the HB method including the zeroth and first harmonic.

Functional	$\mathrm{Re}(F_{\mathrm{mod}})$			$\mathrm{Im}(F_{\mathrm{mod}})$		
Parameter	FD	adjoint	rel. dev.	FD	adjoint	rel. dev.
$\mathrm{Re}(u_x)$	$-1.260\cdot10^{-10}$	$-1.211\cdot10^{-10}$	-3.49%	$1.837\cdot10^{-10}$	$1.859\cdot10^{-10}$	1.40%
$\mathrm{Im}(u_x)$	$-1.837\cdot10^{-10}$	$-1.859\cdot10^{-10}$	1.19%	$-1.260\cdot10^{-10}$	$-1.211\cdot10^{-10}$	-3.74%
$\mathrm{Re}(u_\theta)$	$-1.255\cdot10^{-10}$	$-1.207\cdot10^{-10}$	-3.43%	$1.824\cdot10^{-10}$	$1.854\cdot10^{-10}$	1.81%
$\mathrm{Im}(u_\theta)$	$-1.824\cdot10^{-10}$	$-1.854\cdot10^{-10}$	1.81%	$-1.255\cdot10^{-10}$	$-1.207\cdot10^{-10}$	-3.57%

Table 2: Comparison of sensitivities for the LH method including only the first harmonic.

agreement is slightly better than for harmonic balance. The linearity of this method is reflected by the fact that the sensitivities satisfy the relations

$$\frac{d\mathrm{Re}(F_{\mathrm{mod}})}{d\mathrm{Re}(\alpha)} = \frac{d\mathrm{Im}(F_{\mathrm{mod}})}{d\mathrm{Im}(\alpha)} \quad \text{and} \quad \frac{d\mathrm{Re}(F_{\mathrm{mod}})}{d\mathrm{Im}(\alpha)} = -\frac{d\mathrm{Im}(F_{\mathrm{mod}})}{d\mathrm{Re}(\alpha)}. \tag{25}$$

The differences between the LH and HB results are not very large, which indicates that the nonlinear effects are not very strong in this test case.

3.2 Q3D slice of a turbine stage

As a first turbomachinery test case we take a Q3D slice of the rotor from a high pressure turbine stage with subsonic flow conditions (maximum Mach number about 0.78). The wake of the stator is extracted from a steady computation and the circumferential component with wave number $m = -70$, which corresponds to a phase shift of 60 degrees, is prescribed as gust perturbation at the entry of the rotor. The frequency in the relative system of the rotor is $11186.5\,\mathrm{Hz}$. The resulting flow is visualised in Fig. 6. The used artificial mode corresponds to a bending of the blade.

The computational grid is rather coarse with roughly 2000 cells for the rotor row. The forward computations are carried out with a CFL number of 5 and the computations are terminated when the L^1-residual of the pseudo-time solver has dropped below 10^{-8}. Figure 7 shows that the modal force is sufficiently well converged at this point. For the adjoint computations the SSOR preconditioner with a relaxation factor of 0.7 is applied and the restart interval is set to 100.

In the adjoint solver, the sensitivities are computed at each restart of the GMRes algorithm, therefore we can also monitor the convergence of these quantities. In Fig. 8 this is shown exemplarily for one component, namely the sensitivity of the real part of the modal force to the real part of the x-component of the velocity. It can be seen that after the last restart the sensitivity does not change any more, therefore in the following the adjoint computations are terminated if the (normalized) L^2-residual of the GMRes solver has dropped below 10^{-8}.

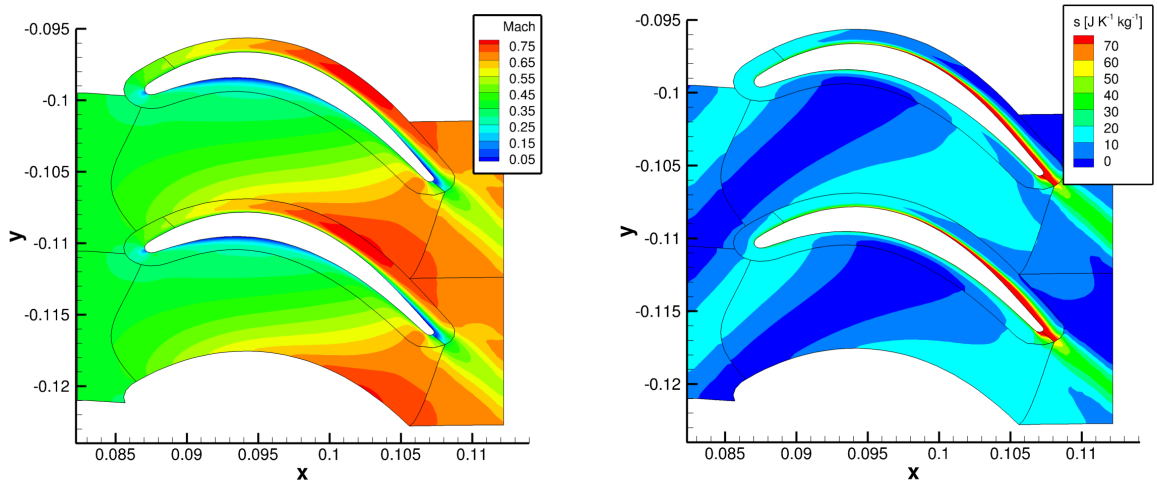


Figure 6: Mach number contours of the time averaged solution (left) and entropy contours of the reconstructed unsteady solution at time $t = 0$ (right), both from a harmonic balance computation with four higher harmonics.

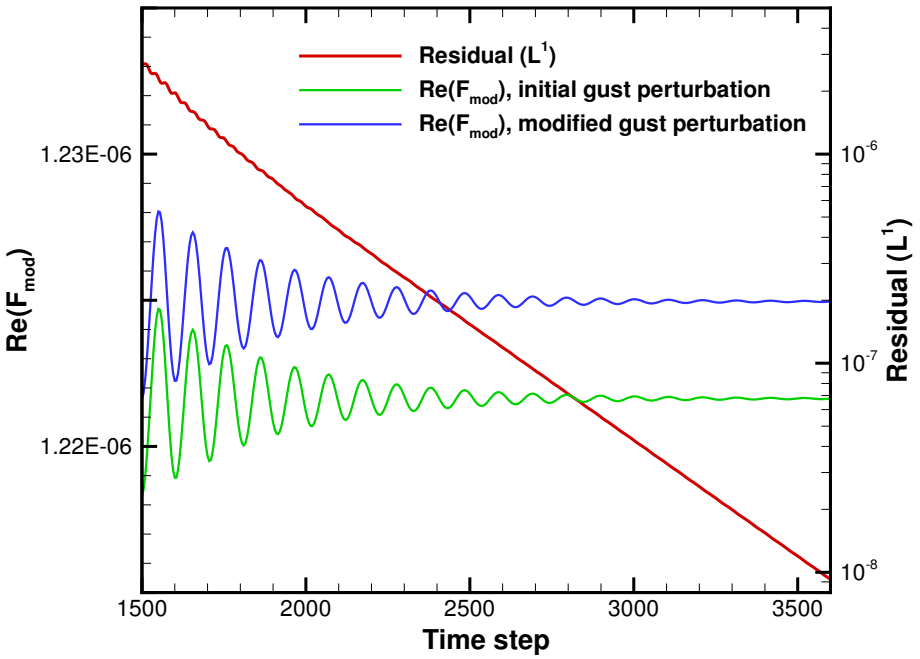


Figure 7: Convergence of the modal force for the initial gust perturbation and the perturbation with modified component $\mathrm{Re}(u_x)$ in relation to the pseudo-time solver residual.

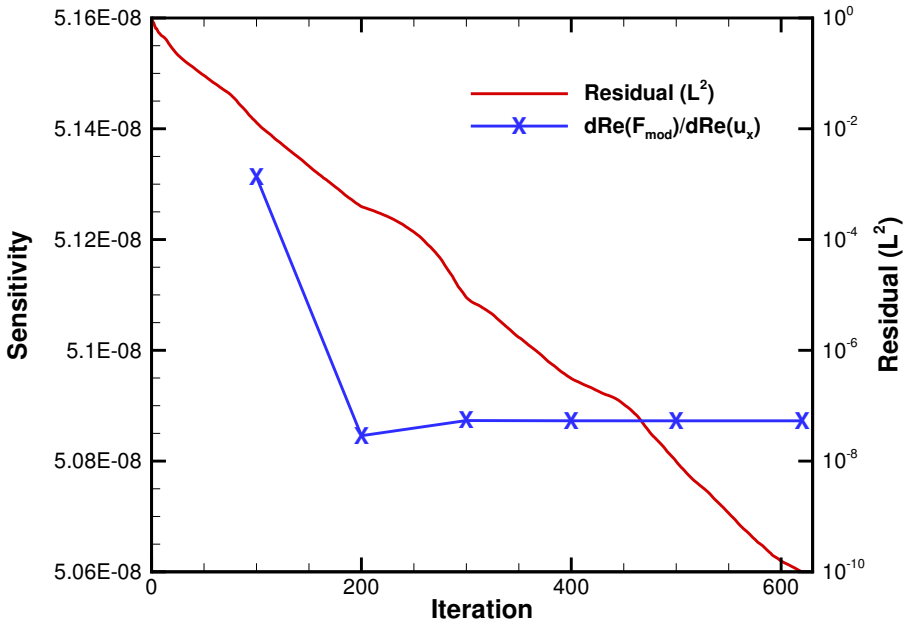


Figure 8: Convergence of sensitivity component in relation to adjoint solver (GMRes) residual.

Functional		$\mathrm{Re}(F_{\mathrm{mod}})$			$\mathrm{Im}(F_{\mathrm{mod}})$	
Parameter	FD	adjoint	rel. dev.	FD	adjoint	rel. dev.
$\mathrm{Re}(\rho)$	$-2.800 \cdot 10^{-5}$	$-2.880 \cdot 10^{-5}$	2.85%	$1.009 \cdot 10^{-5}$	$9.980 \cdot 10^{-6}$	−1.14%
$\mathrm{Im}(\rho)$	$-1.139 \cdot 10^{-5}$	$-1.028 \cdot 10^{-5}$	−9.76%	$-2.769 \cdot 10^{-5}$	$-2.701 \cdot 10^{-5}$	−2.45%
$\mathrm{Re}(u_x)$	$5.083 \cdot 10^{-8}$	$5.087 \cdot 10^{-8}$	0.08%	$2.897 \cdot 10^{-8}$	$2.626 \cdot 10^{-8}$	−9.36%
$\mathrm{Im}(u_x)$	$-2.375 \cdot 10^{-8}$	$-2.277 \cdot 10^{-8}$	−4.15%	$5.134 \cdot 10^{-8}$	$5.301 \cdot 10^{-8}$	3.25%
$\mathrm{Re}(u_\theta)$	$-7.081 \cdot 10^{-8}$	$-7.216 \cdot 10^{-8}$	1.90%	$-3.453 \cdot 10^{-8}$	$-3.374 \cdot 10^{-8}$	−2.28%
$\mathrm{Im}(u_\theta)$	$2.793 \cdot 10^{-8}$	$2.756 \cdot 10^{-8}$	−1.34%	$-7.184 \cdot 10^{-8}$	$-7.620 \cdot 10^{-8}$	6.07%
$\mathrm{Re}(p)$	$7.441 \cdot 10^{-10}$	$7.514 \cdot 10^{-10}$	0.98%	$3.215 \cdot 10^{-10}$	$3.124 \cdot 10^{-10}$	−2.84%
$\mathrm{Im}(p)$	$-2.538 \cdot 10^{-10}$	$-2.543 \cdot 10^{-10}$	0.17%	$7.550 \cdot 10^{-10}$	$7.759 \cdot 10^{-10}$	2.76%

Table 3: Comparison of sensitivities for the HB method including the zeroth and first harmonic.

Functional		$\mathrm{Re}(F_{\mathrm{mod}})$			$\mathrm{Im}(F_{\mathrm{mod}})$	
Parameter	FD	adjoint	rel. dev.	FD	adjoint	rel. dev.
$\mathrm{Re}(\rho)$	$-3.468 \cdot 10^{-5}$	$-3.465 \cdot 10^{-5}$	−0.07%	$-3.327 \cdot 10^{-6}$	$-3.330 \cdot 10^{-6}$	0.08%
$\mathrm{Im}(\rho)$	$3.327 \cdot 10^{-6}$	$3.330 \cdot 10^{-6}$	0.08%	$-3.468 \cdot 10^{-5}$	$-3.465 \cdot 10^{-5}$	−0.07%
$\mathrm{Re}(u_x)$	$3.546 \cdot 10^{-8}$	$3.568 \cdot 10^{-8}$	0.62%	$2.594 \cdot 10^{-8}$	$2.539 \cdot 10^{-8}$	−2.12%
$\mathrm{Im}(u_x)$	$-2.597 \cdot 10^{-8}$	$-2.539 \cdot 10^{-8}$	−2.22%	$3.540 \cdot 10^{-8}$	$3.568 \cdot 10^{-8}$	0.80%
$\mathrm{Re}(u_\theta)$	$-5.013 \cdot 10^{-8}$	$-5.123 \cdot 10^{-8}$	2.18%	$-3.101 \cdot 10^{-8}$	$-3.137 \cdot 10^{-8}$	1.17%
$\mathrm{Im}(u_\theta)$	$3.101 \cdot 10^{-8}$	$3.137 \cdot 10^{-8}$	1.17%	$-5.013 \cdot 10^{-8}$	$-5.123 \cdot 10^{-8}$	2.18%
$\mathrm{Re}(p)$	$5.684 \cdot 10^{-10}$	$5.719 \cdot 10^{-10}$	0.61%	$3.202 \cdot 10^{-10}$	$3.229 \cdot 10^{-10}$	0.84%
$\mathrm{Im}(p)$	$-3.202 \cdot 10^{-10}$	$-3.229 \cdot 10^{-10}$	0.84%	$5.684 \cdot 10^{-10}$	$5.719 \cdot 10^{-10}$	0.61%

Table 4: Comparison of sensitivities for the LH method including only the first harmonic.

Functional	$\mathrm{Re}(F_{\mathrm{mod}})$			$\mathrm{Im}(F_{\mathrm{mod}})$		
Parameter	FD	adjoint	rel. dev.	FD	adjoint	rel. dev.
$\mathrm{Re}(\rho)$	$-3.476 \cdot 10^{-5}$	$-3.470 \cdot 10^{-5}$	-0.17%	$-3.345 \cdot 10^{-6}$	$-3.318 \cdot 10^{-6}$	-0.81%
$\mathrm{Im}(\rho)$	$3.132 \cdot 10^{-6}$	$3.108 \cdot 10^{-6}$	-0.78%	$-3.522 \cdot 10^{-5}$	$-3.519 \cdot 10^{-5}$	-0.09%
$\mathrm{Re}(u_x)$	$3.551 \cdot 10^{-8}$	$3.585 \cdot 10^{-8}$	0.97%	$2.602 \cdot 10^{-8}$	$2.555 \cdot 10^{-8}$	-1.82%
$\mathrm{Im}(u_x)$	$-2.587 \cdot 10^{-8}$	$-2.547 \cdot 10^{-8}$	-1.57%	$3.506 \cdot 10^{-8}$	$3.541 \cdot 10^{-8}$	1.00%
$\mathrm{Re}(u_\theta)$	$-5.018 \cdot 10^{-8}$	$-5.150 \cdot 10^{-8}$	2.64%	$-3.101 \cdot 10^{-8}$	$-3.154 \cdot 10^{-8}$	1.69%
$\mathrm{Im}(u_\theta)$	$3.111 \cdot 10^{-8}$	$3.154 \cdot 10^{-8}$	1.39%	$-4.947 \cdot 10^{-8}$	$-5.082 \cdot 10^{-8}$	2.74%
$\mathrm{Re}(p)$	$5.687 \cdot 10^{-10}$	$5.742 \cdot 10^{-10}$	0.98%	$3.205 \cdot 10^{-10}$	$3.245 \cdot 10^{-10}$	1.26%
$\mathrm{Im}(p)$	$-3.199 \cdot 10^{-10}$	$-3.239 \cdot 10^{-10}$	1.25%	$5.644 \cdot 10^{-10}$	$5.697 \cdot 10^{-10}$	0.95%

Table 5: Comparison of sensitivities for the HB method including only the first harmonic.

We apply both the harmonic balance and the linear harmonic method, at first using only one harmonic. For the finite differences, all components are varied except for the radial velocity, which is not relevant in a Q3D configuration. The reference quantities are $\rho_{\mathrm{ref}} = 0.439\,\frac{\mathrm{kg}}{\mathrm{m}^3}$, $u_{\mathrm{ref}} = 653.3\,\frac{\mathrm{m}}{\mathrm{s}}$ and $p_{\mathrm{ref}} = 133870\,\mathrm{Pa}$. Tables 3 and 4 show the sensitivities for the two methods computed by finite differences (FD) as described above and by the adjoint solver, and the relative deviation between them. For LH, there is a very good agreement between both solvers. Almost all deviations are below 5%, most of them even below 1% (see Tab. 4). For HB the deviations between finite differences and adjoint computations are somewhat larger (see Tab. 3). Most deviations are still below 5%, but some are also up to 10% large. These differences could be due to inconsistencies between the two solvers, as mentioned above, but should be further investigated. It can also be observed that the HB sensitivities differ considerably from the LH results. For comparison we also performed a harmonic balance computation with only the first harmonic, i.e. neglecting the coupling between this harmonic and the time-mean solution (zeroth harmonic). Table 5 shows that in this case the results are much closer to those of the LH method, indicating that there is a strong interaction between the zeroth and first harmonic. Comparing the different sensitivity components as in Eqn. (25) we see that they are closer to a linear behaviour than in the coupled computation, but the nonlinearity is still present, although there is only one solution component. A comparison of all methods is shown in Fig. 9 for the sensitivities w.r.t. the angular component of the velocity.

We also investigated the effect of including higher harmonics in the computations. Since only the first harmonic is used for the evaluation of the objective functionals, there is only an indirect influence of higher harmonics through their interaction with the first, and it turns out that the impact on the modal force - and therefore the sensitivities - is indeed small. In Fig. 10 the values of the real and imaginary parts of the modal force are shown over the number of higher harmonics included in the HB computation and it can be seen that only the second harmonic yields a significant contribution. The convergence of the GMRes solver becomes slower if more harmonics are included, as can be seen in Fig. 11, but we were able to obtain converged solutions for all sets of harmonics considered here. The agreement between forward and adjoint results is qualitatively similar to the case with only one higher harmonic, although the maximal deviations become somewhat larger.

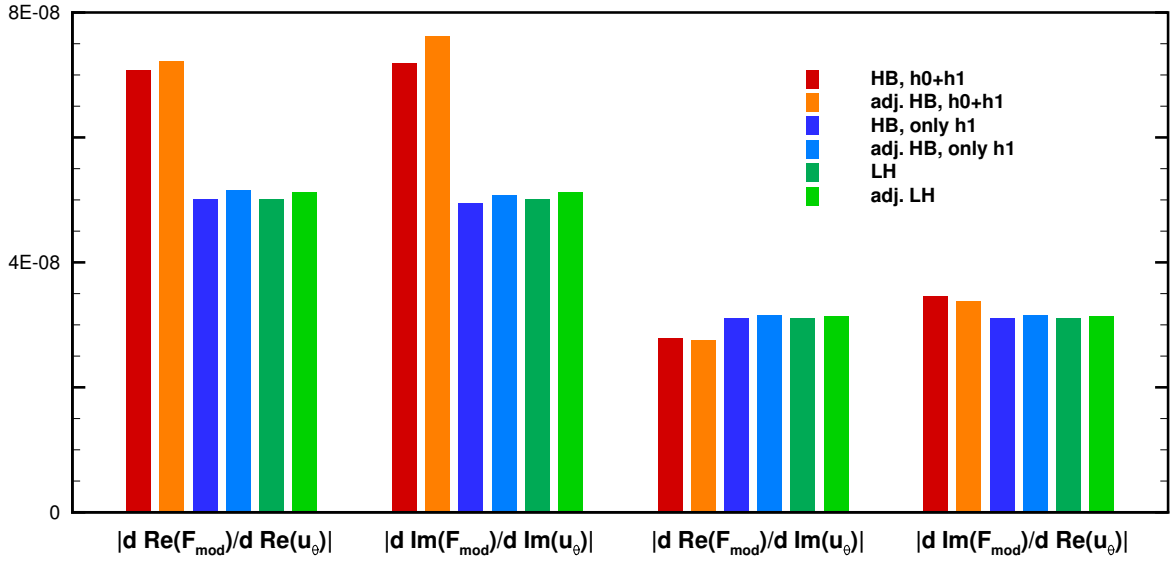


Figure 9: Comparison of four sensitivity components for different methods

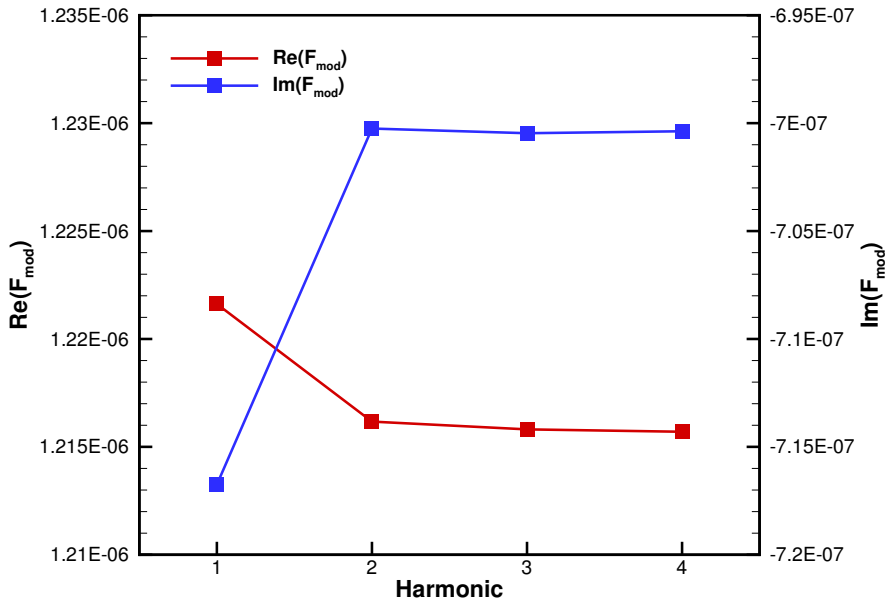


Figure 10: Real and imaginary parts of modal force for harmonic balance computations with different sets of harmonics ("Harmonic" is the order of the highest harmonic which is included).

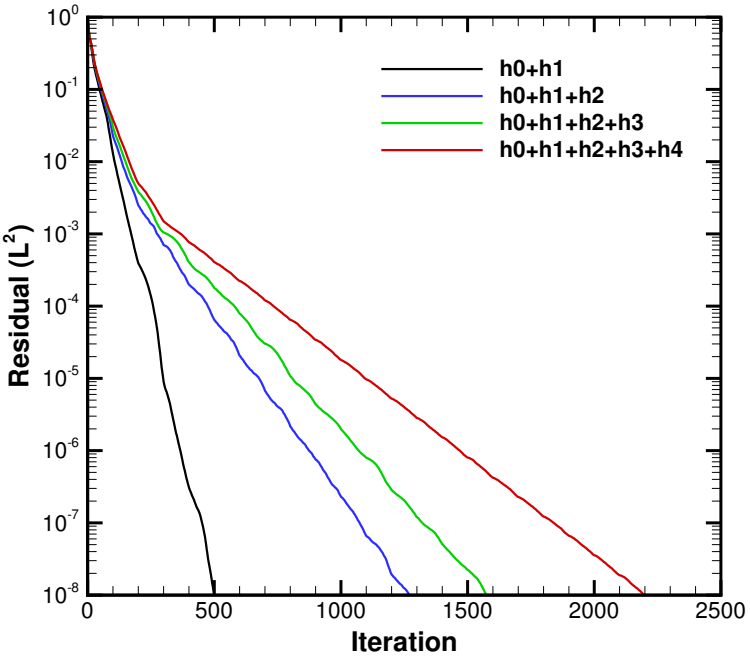


Figure 11: Convergence of the adjoint solver for the functional $\mathrm{Re}(F_{\mathrm{mod}})$ with different sets of harmonics included.

4 CONCLUSION

We have described the development and implementation of an adjoint solver for frequency domain methods (harmonic balance and linear harmonics) within an industrial turbomachinery CFD solver. It is based on the extension of an existing steady adjoint solver. In particular, the implementation of the modal force on blades as a functional for aeroelastic applications and the evaluation of sensitivities with respect to the components of a flow perturbation at inflow or outflow boundaries have been discussed. The method has been extensively tested on Q3D configurations which are representative for turbomachinery applications. The overall good agreement of the sensitivities obtained by the adjoint solver with finite differences shows the validity of the approach. The reasons for some larger deviations have to be investigated in the future. Also the extension to 3D configurations has to be tested in detail.

Another important issue for future work is the coupling of different blade rows, since we are interested in the influence of the geometry of a neighbouring blade row on e.g. the modal forces for the blade row under consideration. In the current process, where the influence of e.g. a stator wake on the subsequent blade row is simulated only for selected wave numbers, the next step would be to use the sensitivities from the adjoint harmonic balance solver as input, i.e. objective functional, for a following steady adjoint computation. This would then give the geometric sensitivities for the upstream blade row. To obtain an adjoint process for a complete forced-response analysis using harmonic balance, the blade-row coupling and a geometric sensitivity evaluation – analogous to the steady adjoint solver – have to be implemented within the adjoint HB solver.

Acknowledgements The authors would like to thank J. Backhaus and H.-P. Kersken for helpful discussions.

REFERENCES

[1] K. C. Hall and E. F. Crawley, "Calculation of unsteady flows in turbomachinery using the linearized Euler equations," *AIAA Journal*, vol. 27, no. 6, pp. 777–787, 1989.

[2] R. Kielb, "CFD for turbomachinery unsteady flows - an aeroelastic design perspective," in *Aerospace Sciences Meetings*, American Institute of Aeronautics and Astronautics, Jan. 2001.

[3] L. He, "Fourier methods for turbomachinery applications," *Progress In Aerospace Sciences*, vol. 46, pp. 329–341, Nov. 2010.

[4] K. Ekici and H. Huang, "An assessment of frequency-domain and time-domain techniques for turbomachinery aeromechanics," in *30th AIAA Applied Aerodynamics Conference 2012*, pp. 1807–1823, June 2012.

[5] K. C. Hall, J. P. Thomas, and W. S. Clark, "Computation of unsteady nonlinear flows in cascades using a harmonic balance technique," *AIAA Journal*, vol. 40, pp. 879–886, May 2002.

[6] J. P. Thomas, E. H. Dowell, and K. C. Hall, "Nonlinear inviscid aerodynamic effects on transonic divergence, flutter, and limit-cycle oscillations," *AIAA Journal*, vol. 40, pp. 638–646, Apr. 2002.

[7] J. P. Thomas, E. H. Dowell, and K. C. Hall, "Modeling viscous transonic limit-cycle oscillation behavior using a harmonic balance approach," *Journal of Aircraft*, vol. 41, no. 6, pp. 1266–1274, 2004.

[8] H. Huang and K. Ekici, "A harmonic balance method for the analysis of unsteady flows in cascades," in *47th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit*, American Institute of Aeronautics and Astronautics, July 2011.

[9] H. Huang and K. Ekici, "An efficient harmonic balance method for unsteady flows in cascades," *Aerospace Science and Technology*, vol. 29, no. 1, pp. 144–154, 2013.

[10] H. Huang and K. Ekici, "Stabilization of high-dimensional harmonic balance solvers using time spectral viscosity," *AIAA Journal*, vol. 52, no. 8, pp. 1784–1794, 2014.

[11] A. Gopinath and A. Jameson, "Time spectral method for periodic unsteady computations over two- and three- dimensional bodies," in *Proceedings of the 43rd AIAA Aerospace Sciences Meeting & Exhibit*, 2005.

[12] E. van der Weide, A. Gopinath, and A. Jameson, "Turbomachinery applications with the time spectral method," in *35th AIAA Fluid Dynamics Conference and Exhibit, AIAA paper 05-4905*, American Institute of Aeronautics and Astronautics, June 2005.

[13] K. C. Hall, K. Ekici, J. P. Thomas, and E. H. Dowell, "Harmonic balance methods applied to computational fluid dynamics problems," *International Journal of Computational Fluid Dynamics*, vol. 27, no. 2, pp. 52–67, 2013.

[14] M. McMullen, A. Jameson, and J. Alonso, "Acceleration of convergence to a periodic steady state in turbomachinery flows," in *AIAA 39th Aerospace Sciences Meeting, AIAA paper 01-0152*, American Institute of Aeronautics and Astronautics, Jan. 2001.

[15] M. McMullen, A. Jameson, and J. Alonso, "Application of a non-linear frequency domain solver to the Euler and Navier–Stokes equations," in *AIAA 40th Aerospace Sciences Meeting, AIAA paper 02-0120*, American Institute of Aeronautics and Astronautics, Jan. 2002.

[16] M. S. McMullen, *The Application of Non-Linear Frequency Domain Methods to the Euler and Navier-Stokes Equations*. PhD thesis, Stanford University, 2003.

[17] C. Frey, G. Ashcroft, H.-P. Kersken, and C. Voigt, "A harmonic balance technique for multistage turbomachinery applications," in *Proceedings of ASME Turbo Expo 2014*, 2014.

[18] G. Ashcroft, C. Frey, and H.-P. Kersken, "On the development of a harmonic balance method for aeroelastic analysis," in *6th European Conference on Computational Fluid Dynamics (ECFD VI)*, 2014.

[19] A. Jameson, "Aerodynamic design via control theory," *Journal of Scientific Computing*, vol. 3, no. 3, pp. 233–260, 1988.

[20] M. B. Giles and N. A. Pierce, "An introduction to the adjoint approach to design," *Flow, Turbulence and Combustion*, vol. 65, pp. 393–415, 2000.

[21] D. X. Wang and L. He, "Adjoint aerodynamic design optimization for blades in multistage turbomachines—part I: Methodology and verification," *Journal of Turbomachinery*, vol. 132, no. 2, p. 021011, 2010.

[22] D. X. Wang, L. He, Y. S. Li, and R. G. Wells, "Adjoint aerodynamic design optimization for blades in multistage turbomachines—part II: Validation and application," *Journal of Turbomachinery*, vol. 132, no. 2, p. 021012, 2010.

[23] S. K. Nadarajah and A. Jameson, "Optimum shape design for unsteady flows with time-accurate continuous and discrete adjoint methods," *AIAA Journal*, vol. 45, pp. 1478–1491, July 2007.

[24] E. J. Nielsen and B. Diskin, "Discrete adjoint-based design for unsteady turbulent flows on dynamic overset unstructured grids," *AIAA Journal*, vol. 51, pp. 1355–1373, June 2013.

[25] G. Ntanakas and M. Meyer, "Towards unsteady adjoint analysis for turbomachinery applications," in *Proceedings of the 6th. European Conference on Computational Fluid Dynamics - ECFD VI*, pp. 5071–5081, 2014.

[26] R. Florea and K. Hall, "Sensitivity analysis of unsteady inviscid flow through turbomachinery cascades," *AIAA Journal*, vol. 39, pp. 1047–1056, June 2001.

[27] M. C. Duta, M. B. Giles, and M. S. Campobasso, "The harmonic adjoint approach to unsteady turbomachinery design," *International Journal for Numerical Methods in Fluids*, vol. 40, no. 3-4, pp. 323–332, 2002.

[28] M. Duta, M. Campobasso, M. Giles, and L. Lapworth, "Adjoint harmonic sensitivities for forced response minimization," *Journal of Engineering for Gas Turbines and Power*, vol. 128, pp. 183–189, Jan. 2006.

[29] J. Thomas, K. Hall, and E. Dowell, "Discrete adjoint approach for modeling unsteady aerodynamic design sensitivities," *AIAA Journal*, vol. 43, pp. 1931–1936, Sep. 2005.

[30] H. Huang and K. Ekici, "A discrete adjoint harmonic balance method for turbomachinery shape optimization," *Aerospace Science and Technology*, vol. 39, pp. 481–490, 2014.

[31] C. Mader and J. Martins, "Derivatives for time-spectral computational fluid dynamics using an automatic differentiation adjoint," *AIAA Journal*, vol. 50, no. 12, pp. 2809–2819, 2012.

[32] S. K. Nadarajah, M. S. McMullen, and A. Jameson, "Aerodynamic shape optimization for unsteady three-dimensional flows," *International Journal of Computational Fluid Dynamics*, vol. 20, pp. 533–548, Sep. 2006.

[33] S. K. Nadarajah and A. Jameson, "Optimum shape design for unsteady three-dimensional viscous flows using a nonlinear frequency-domain method," *Journal of Aircraft*, vol. 44, no. 5, pp. 1513–1527, 2007.

[34] L. He and D. X. Wang, "Concurrent blade aerodynamic-aero-elastic design optimization using adjoint method," *Journal of Turbomachinery*, vol. 133, p. 011021, Jan. 2011.

[35] C. Frey, D. Nürnberger, and H.-P. Kersken, "The discrete adjoint of a turbomachinery RANS solver," in *Proceedings of ASME-GT2009*, 2009.

[36] A. Engels-Putzka and C. Frey, "Adjoint harmonic balance method for forced response analysis in turbomachinery," in *Proceedings of the VI International Conference on Coupled Problems in Science and Engineering*, pp. 645–656, 2015.

[37] Y. Saad, *Iterative methods for sparse linear systems. 2nd ed.* SIAM Society for Industrial and Applied Mathematics, Philadelphia., 2003.

[38] A. Engels-Putzka and C. Frey, "Adjoint mesh deformation and adjoint-based sensitvitities with respect to boundary values," in *ECCOMAS 2012 - European Congress on Computational Methods in Applied Sciences and Engineering, e-Book Full Papers*, pp. 1789–1808, 2012.

[39] C. Frey, G. Ashcroft, H.-P. Kersken, and C. Weckmüller, "Advanced numerical methods for the prediction of tonal noise in turbomachinery — Part II: Time-linearized methods," *Journal of Turbomachinery*, vol. 136, no. 2, p. 021002, 2013.

[40] H.-P. Kersken, C. Frey, C. Voigt, and G. Ashcroft, "Time-linearized and time-accurate 3D RANS methods for aeroelastic analysis in turbomachinery," *Journal of Turbomachinery*, vol. 134, no. 5, p. 051024, 2012.

[41] C. K. W. Tam and J. C. Hardin, eds., *Second Computational Aeroacoustics (CAA) Workshop on Benchmark Problems*, NASA Conference Publication 3352, NASA Langley Research Center, June 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A TWO-LEVEL HYBRID APPROACH FOR OPTIMAL ACTIVE FLOW CONTROL ON A THREE-ELEMENT AIRFOIL

Anil Nemili[1], Emre Özkaya[2], Nicolas R. Gauger[2], Felix Kramer[3] and Frank Thiele[3]

[1]Department of Mathematics, BITS Pilani - Hyderabad Campus, Hyderabad 500078, India
e-mail: anil@hyderabad.bits-pilani.ac.in

[2] Chair for Scientific Computing, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany
e-mail: {emre.oezkaya, nicolas.gauger}@scicomp.uni-kl.de

[3] CFD Software Entwicklungs- und Forschungsgesellschaft mbH, 10625 Berlin, Germany
e-mail: {felix.kramer, frank.thiele}@cfd-berlin.com

Keywords: Active Flow Control, Unsteady RANS, Hybrid optimization, Discrete adjoints, Evolutionary Algorithms, Algorithmic Differentiation, Multi-element airfoil

Abstract. *In this paper we present a two-level approach that combines an adjoint-based gradient search method with an evolutionary algorithm for optimal active flow control. The suggested method effectively combines the advantages of both approaches and achieves a good compromise between the computational effort and the degree of freedom used in optimization. In the first level, a global optimization is performed with few design parameters using an evolutionary algorithm. In the second level, the global optimal solution from the first level is taken as the initial setting for the adjoint based local optimization using a large number of design parameters. The unsteady discrete adjoint solver required for the second level is developed based on Algorithmic Differentiation techniques for the unsteady incompressible flowsgoverned by Unsteady Reynolds-Averaged Navier Stokes (URANS) equations. In this way, the discrete adjoint solver is robust and has exactly the same functionality with the underlying URANS flow solver. The applicability of the two-level method is demonstrated by finding the optimal parameters of the active flow control mechanism on a three element airfoil configuration at a Reynolds number of* $Re = 10^6$ *and an angle of attack of* $AoA = 6^\circ$*. Numerical results have shown that the hybrid approach completely suppressed the separation and very significantly increased the mean-lift coefficient by* 67% *compared to the un-actuated baseline flow.*

1 INTRODUCTION

Application of active flow control methods [1] to delay the flow separation on the flap of modern transport aircraft has been the subject of research for many years. An active flow control technique that is of particular interest to us is the predetermined or the so-called open-loop active flow control approach. Typically, these methods supply energy or auxiliary power to the boundary layer through an actuator. The additional longitudinal momentum due to actuation delays or even completely suppresses the flow separation. Various researchers have performed several experimental [2, 3, 4, 5, 6] and numerical investigations [7, 8, 9] to study the effect of actuation in delaying the separation on the flap and enhancing the overall aerodynamic performance of high-lift configurations. In the present work, synthetic jets, also known as zero-net-mass-flux actuators [10, 11] are employed to control the flow separation on the suction side of the flap. Common parameters of these actuators are amplitude, blowing angles, frequency and the phase shift between two adjacent actuators. Effective separation control and thus delaying aerodynamic stall to enhance maximum lift can be achieved by finding the optimal set of actuation parameters.

The optimal values of the actuation parameters can be efficiently determined by employing gradient-based optimization algorithms combined with adjoint approaches. A nice feature of the adjoint approaches is that they enable the evaluation of gradient vectors at a fixed computational effort. Therefore, optimization studies with a higher degree of freedom become viable and better control mechanisms can be achieved. Broadly, we can classify the adjoint approaches into continuous and discrete adjoint methods. In the continuous adjoint method [12, 13], one first derives the optimality system from the continuous optimization problem and the resulting adjoint partial differential equations (PDEs) are then discretised and solved using numerical methods. Although being computationally efficient, development of continuous adjoint flow solvers requires much effort and their maintenance becomes a problem as the underlying non-linear flow solvers are subject to continuous modifications, e.g., new boundary conditions, new turbulence models etc. On the other hand, in the discrete adjoint method, one first discretizes the state PDEs that govern the fluid flow, and then derives the discrete adjoint equation based on the optimization problem in discrete form. In general, discrete adjoint solvers are more straightforward to implement, and therefore they have found a wider acceptance for the applications of practical relevance.

In general, a discrete adjoint solver for optimal active flow control can be developed either by using the so-called hand-discrete approach [14] or by employing Algorithmic Differentiation (AD) techniques [15, 16]. In the hand-discrete approach, the governing flow equations are first discretized as usual. The adjoint equations are then derived by linearizing various terms, which comprise the discrete cell residuals. These terms involve convective and viscous fluxes as well as source terms. Finally a computer code is implemented to solve the unsteady adjoint equations and to compute the design sensitivities using the adjoint solution. On the other hand, in the AD based approach the adjoint code is generated by applying AD techniques [17] directly to the source code of the corresponding flow solver.

Accurate computation of sensitivity information requires exact differentiation of all residual terms in discretized state equations. However, the exact linearization of all terms in the discrete residual is quite complex, laborious and might be often prone to errors. To simplify this tedious

effort and ease the development of the adjoint solvers, various Jacobian approximations [18] have been proposed in the past. In these approximations, linearization of certain terms in the flux Jacobian is omitted. Examples of these are, turbulent models, flux limiters, reconstruction operator for higher order accuracy, model parameters of the convective schemes etc. Jacobian approximations may result in inaccurate computation of sensitivities [19]. In fact, in unsteady flows, the effect of these approximations on the accuracy of sensitivities is much more significant as the errors generated in the adjoint solution tend to accumulate rapidly while solving the adjoint equations backward-in-time [16]. On the other hand, using the AD techniques the exact differentiation of all residual terms can be performed with much ease.

Although being computationally very efficient, adjoint-based gradient search methods may perform poorly in some cases as their solutions tend to be trapped to local optima. In problems with fairly noisy response surfaces, it is typical that the optimization method yields some improvement after few cycles. However, further cycles may not yield any significant improvement. Moreover, gradient-based algorithms are highly sensitive to the initial values chosen for the design parameters. Therefore, a "bad" initial setting for the design parameters may result in a significantly lower performance than expected. As a remedy to this problem, a global optimization method may provide a better improvement within a specified design space. For example, Evolutionary Algorithms (EAs) [20, 21], which try to imitate the natural evolution processes, are widely accepted as global optimization tools. EAs are generally computationally expensive and require a prohibitive amount of computational resources if the number of design parameters is large. For this reason, EAs are applicable to limited problems with a few design variables and they are usually implemented in conjunction with surrogate modeling techniques such as response surface methods [22], radial basis functions [23] or Kriging method [24].

In the present study, an efficient two-level design strategy is employed by taking advantage of both the global optimization method based on EA and the local optimization from the adjoint-based gradient search. Since very few design parameters are used in the first level, the evolutionary algorithm is constructed directly based on the non-linear URANS simulations. In the second level, a gradient-based optimization is started using the best initial values provided by the global optimization from the first level. In order to fully exploit the capabilities of the adjoint solver, the control parameters of the actuation are allowed to vary spatially in the flap region so that largest degree of freedom is taken for the optimization, thus allowing the maximum design space dimension. A similar design strategy for aerodynamic shape optimization is suggested in [25].

This paper is organized as follows. Section 2 is devoted to the development of a robust and accurate discrete adjoint incompressible URANS solver, meant for optimal active flow control. In Section 3, we present the EA based optimization method, which is employed to specify the initial point for the gradient-based optimization stage. In Section 4, we introduce the configuration used in the present study along with the features of flow and adjoint solvers used in the present work. In Section 5, numerical results are presented based on optimal synthetic jet actuation on the three-element airfoil high-lift configuration. Finally, conclusions are drawn in Section 6.

2 A DISCRETE ADJOINT APPROACH FOR OPTIMAL ACTIVE FLOW CONTROL

Consider the problem of finding an optimal active flow control mechanism that delays or suppresses the flow separation and thus increases the stall angle of the three-element airfoil. This amounts to the solution of a PDE-constrained optimal control problem. Since we are interested in enhancing the overall lift coefficient, the objective function is defined as the time-averaged lift coefficient. The PDE constraints are the incompressible URANS equations, that govern the low speed flows during the landing and take-off phases of modern transport aircraft. The control variables are the parameters of actuation.

In the discrete form, the time-averaged or the mean lift coefficient $\overline{C}_l$ over the time interval $[0, T]$ can be defined as

$$\overline{C}_l = \frac{1}{N} \sum_{n=1}^{N} I^n \left(\mathbf{U}^n, \alpha \right), \quad I^n = C_l^n \tag{1}$$

where $\mathbf{U}^n$ and C_l^n are the state vector and the lift coefficient at time iteration n respectively. α is the vector of control variables comprising the actuation parameters amplitude, frequency, phase shift and blowing angle. $N = T/\Delta t$ is the number of time iterations that span the given time interval and Δt is the step size of the time discretisation scheme. At each time iteration n, the discretised URANS equations are solved for the solution of the state vector $\mathbf{U^n}$ by a contractive fixed-point iterative scheme of the form

$$\mathbf{U}_{i+1}^n = G^n \left(\mathbf{U}_i^n, \mathbf{U}^{n-1}, \mathbf{U}^{n-2}, \alpha \right), \; n = 1, \ldots, N. \tag{2}$$

Here, G^n represents an iteration of the pressure-velocity coupling scheme based on the SIMPLE algorithm, employed for solving the incompressible URANS equations. $\mathbf{U}^{n-1}$ and $\mathbf{U}^{n-2}$ are the converged state vectors at time iterations $n-1$ and $n-2$ respectively. Note that the transient terms in the governing equations are approximated by a second-order implicit backward difference formula. The Lagrangian associated with the constrained optimization problem given by Eqs. (1) and (2) is defined as

$$L = \frac{1}{N} \sum_{n=1}^{N} \left\{ I^n \left(\mathbf{U}^n, \alpha \right) \right\} - \sum_{n=1}^{N} \left\{ \left(\overline{\mathbf{U}}^n \right)^T \left(\mathbf{U}^n - G^n \left(\mathbf{U}^n, \mathbf{U}^{n-1}, \mathbf{U}^{n-2}, \alpha \right) \right) \right\}, \tag{3}$$

where $\overline{\mathbf{U}}^n$ is the vector of Lagrangian multipliers or the adjoint state vector at time iteration n. From the first order necessary conditions for optimality (KKT conditions), the discrete adjoint URANS equations can be derived in the fixed-point form as

$$\overline{\mathbf{U}}_{i+1}^n = \left[\frac{\partial G^n}{\partial \mathbf{U}^n} \right]^T \overline{\mathbf{U}}_i^n + \left[\frac{\partial G^{n+1}}{\partial \mathbf{U}^n} \right]^T \overline{\mathbf{U}}^{n+1} + \left[\frac{\partial G^{n+2}}{\partial \mathbf{U}^n} \right]^T \overline{\mathbf{U}}^{n+2} + \frac{1}{N} \left[\frac{\partial I^n}{\partial \mathbf{U}^n} \right]^T. \tag{4}$$

Here, $\overline{\mathbf{U}}^{n+1}$ and $\overline{\mathbf{U}}^{n+2}$ are the converged adjoint solutions at time iterations $n+1$ and $n+2$ respectively. Unlike the flow equations, the adjoint equations are solved backward-in-time starting from $n = N$ to $n = 1$. The solution of the adjoint equations is then used to compute the actuation sensitivities as

$$\frac{dL}{d\alpha} = \sum_{n=1}^{N} \left\{ \frac{1}{N} \frac{\partial I^n}{\partial \alpha} + \left(\overline{\mathbf{U}}^n \right)^T \frac{\partial G^n}{\partial \alpha} \right\}. \tag{5}$$

From Eqs. (4) and (5) it can be observed that an accurate computation of actuation sensitivities requires the exact differentiation of the I^n and G^n. Note that the fixed-point iterator G^n consists of residuals due to turbulence models, higher-order convective scheme and terms related to implicit time marching algorithm. The exact differentiation of these terms by hand is laborious and prone to errors. On the other hand, Algorithmic Differentiation techniques [17] can be used to perform the exact differentiation of G^n with much ease. Since all terms in the fixed-point iterator can be differentiated exactly, the adjoint code based on AD computes sensitivities that are always accurate and consistent to the solutions obtained by URANS simulations. Furthermore, the adjoint solver inherits the robustness and the asymptotic rate of convergence of the underlying URANS solver [17]. The ability of the AD based approach in accurate computation of sensitivities in incompressible URANS simulations is successfully demonstrated on practical configurations [26].

3 AN EVOLUTIONARY ALGORITHM FOR THE INITIAL CONTROL

Since the results obtained from gradient search algorithms are highly sensitive to the starting values of the actuation parameters, the initial values chosen for these parameters play a crucial role on the success of the optimization study. In general, values obtained from previous simulations and/or experimental results can be specified for the initial values of the actuation parameters. These values are, however, likely to be far away from the optimal values. If initially the actuation parameters are not allowed to vary spatially, application of an EA based optimization using direct URANS simulations is viable since in this case the design vector has only four parameters, namely amplitude, frequency, phase shift of the actuation and the angle of blowing/suction. The evolutionary algorithm used in this work includes the following basic ingredients:

- Selection of the initial sample points using the Latin Hypercube Sampling (LHS) method.
- URANS simulations at the initial sample points to form an initial population for the EA.
- Selection of a parent pair for the reproduction process based on the fitness values of each member in the population.
- The cross-over process between the parent designs using random Gaussian distribution to determine the design parameters of the children (new designs).
- URANS simulation performed at each new design.
- Removal of the member with the worst fitness from the population in each 10 iterations.

The initial population used for the EA algorithm is formed using the LHS method obtained with initial URANS simulations. The selection of the parent pair for the reproduction step is performed by a randomized selection procedure using a roulette wheel containing reproduction likelihoods of each member in the population. The cross-over process between the parent members are then performed using random Gaussian distributions with a mutation likelihood close to 5% (chosen to be 2σ event). As an example, in Figure 1, the probability distribution function obtained by random 10^7 realizations of a cross-over process between two designs with $x = 1.0$ and $x = 5.0$ is shown. As one may notice, the majority of the designs (close to 95%) lie between 1.0 and 5.0. Only a small amount of new designs lie outside this range (in shaded

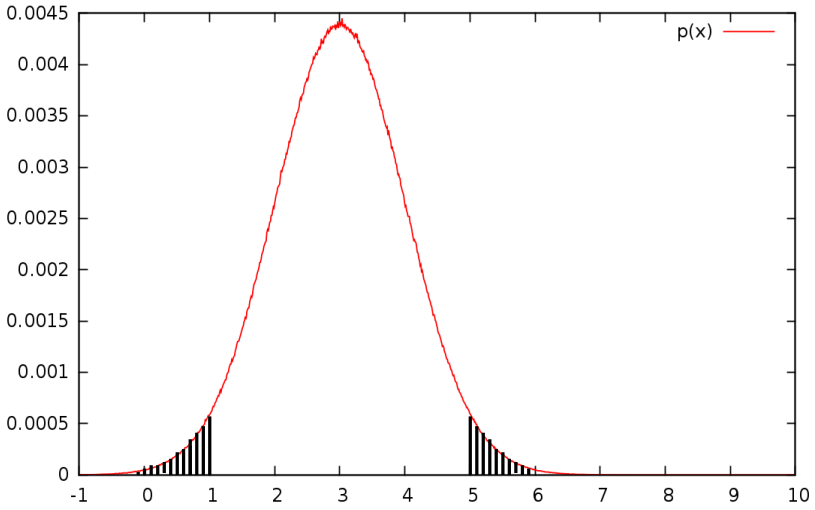


Figure 1: Gaussian distribution of new designs for $x_1 = 1.0$ and $x_2 = 5.0$ obtained from 10^7 different cross-overs.

area). These designs are considered as a result of mutations.

Once the new design point is determined as a result of the cross-over process, a new URANS simulation is performed at this design point and the fitness values of the population members and the reproduction likelihoods of each member in the population (roulette wheel) are accordingly adjusted. After every 10 reproductions, one member with the worst fitness value is removed from the population. The stopping criteria in the EA based optimization is specified as the maximum number of URANS simulations and is set to 250. Therefore, the global optimization in the first level terminates after 250 iterations. The global optimum result obtained by the global optimization is then used to specify the initial values for the actuation parameters in the gradient-search optimization. In contrast to the first level, in the second level, the actuation parameters are allowed to vary in each actuation slot. Therefore, the dimension of the design vector increases from 4 to 162. In Figure 2, the two-level approach is illustrated schematically.

4 COMPUTATIONAL SETUP

4.1 Configuration Details

The geometry of the three-element high-lift configuration is shown in Figure 3. The high-lift configuration consists of a main airfoil with extended slat and flap. The relative chord lengths of slat and flap are given by $l_{slat} = 0.158l_c$ and $l_{flap} = 0.254l_c$, where l_c is the chord length. The flap deflection angle is set to $\delta_{flap} = 37^o$ while the angle of attack is fixed at $AoA = 6^o$. This angle of attack is within the typical range of angle of attack during landing scenarios of modern commercial aircraft. It may be noted that with these settings, the flow over the flap is detached from the shoulder region, while the flow over the slat and main airfoil is still fully attached. Therefore, the above settings are best suited for the application of an active flow control mechanism to delay the separation on the suction side of the flap.

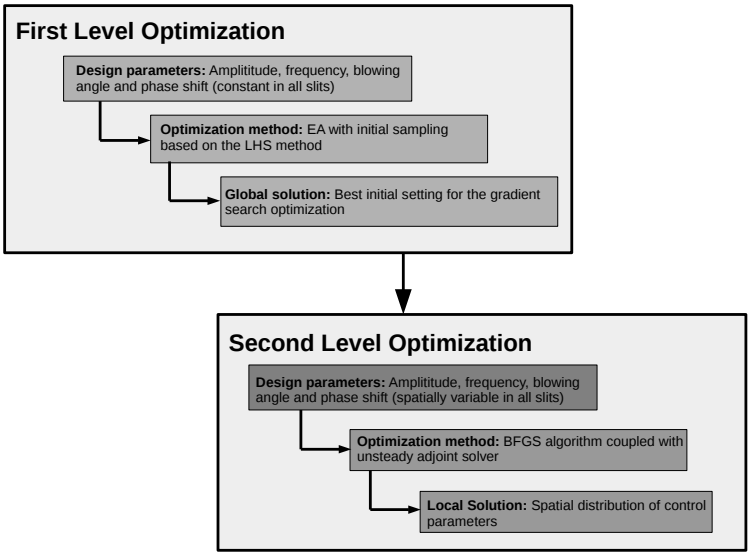


Figure 2: Illustration of the two-level method for optimal active flow control on a three-element airfoil.

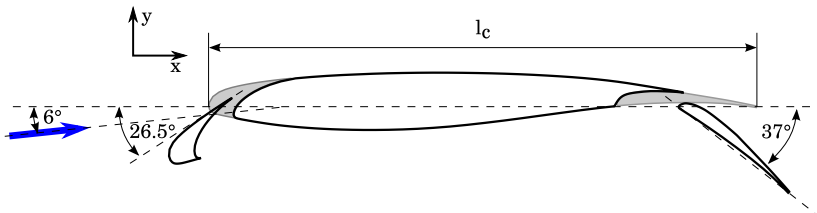


Figure 3: Sketch of a three-element airfoil high-lift configuration.

4.2 Computational Domain

The computational domain consists of a block-structured grid with $172,392$ finite volumes. Figure (4) shows the grid around the flap of the high-lift configuration. The flap boundary has 370 grid points. Note that the non-dimensional wall distance of the first cell center remains below $y^+ = 1$ on the entire surface.

4.3 Boundary Conditions

At the inlet boundary all flow quantities are prescribed. At the outflow, a convective boundary condition is used that allows unsteady flow structures to be transported outside the domain. On the airfoil-flap-slat surface no-slip boundary condition is used. In the present work, the flow control is realized by applying synthetic jet actuation at 27 slots on the suction side of the flap, as shown in Figure (5). Numerically, each slot is resolved by two cell faces on the flap boundary. It can be observed that the actuation is distributed almost on the entire flap starting from the shoulder region. The actuation boundary condition at a flap face is defined by

$$\begin{pmatrix} u \\ v \end{pmatrix} = A \cdot u_\infty \begin{pmatrix} \frac{\cos\theta}{\tan\beta} - \sin\theta \\ \frac{\sin\theta}{\tan\beta} + \cos\theta \end{pmatrix} \sin[2\pi F \cdot (t - t_0)] \tag{6}$$

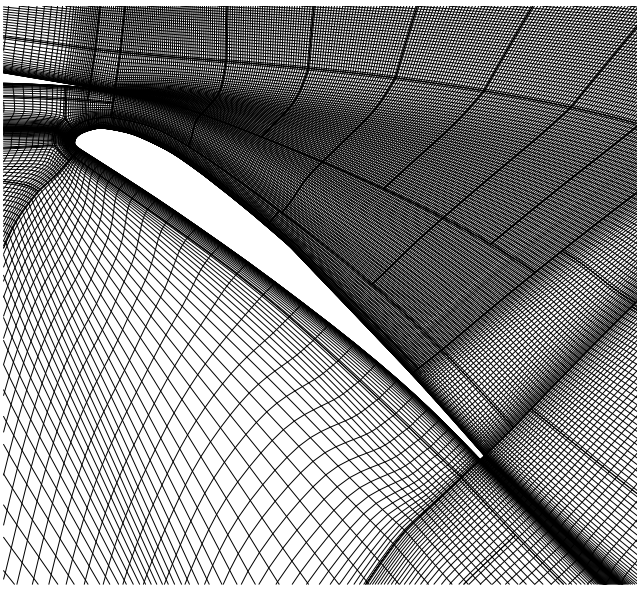


Figure 4: Block-structured grid around the flap of a three-element airfoil.

Here u and v are components of the actuation velocity vector, A is the non-dimensional amplitude of actuation, u_∞ is the free-stream velocity, β is the blowing angle, F is the non-dimensional frequency, t is the non-dimensional physical time and t_0 is the non-dimensional phase shift. The angle of the slit face, θ is fixed by the geometry of the airfoil. The named quantities are appropriately non-dimensionalised by u_∞.

4.4 Numerical Solvers

In the present work, the URANS calculations are performed using the multi-purpose parallel CFD code ELAN [7, 8]. ELAN is an incompressible finite volume solver based on the SIMPLE algorithm. The code is fully implicit and is of second order accuracy both in space and time. The convective fluxes are approximated by a total variation diminishing (TVD) scheme. In addition, the code offers various options for RANS/LES turbulence models.

The discrete adjoint URANS code ELAN-A [16, 26] is developed by applying the AD tool Tapenade [27] to the underlying URANS solver. ELAN-A retains all features of the ELAN code.

5 RESULTS AND DISCUSSION

In this section, we present the numerical results based on optimal active flow control mechanism on the three-element airfoil high-lift configuration. We first consider the base flow without actuation. The incompressible URANS simulations are performed with the SST $k-\omega$ turbulence model at a Reynolds number of $Re = 10^6$ and angle of attack $AoA = 6^o$. The contours of velocity magnitude in Figure (6) show a massive flow separation on the suction side of the flap starting from the shoulder region. Note that the early separation reduces the circulation around the flap and hence the lift contribution from the flap decreases rapidly.

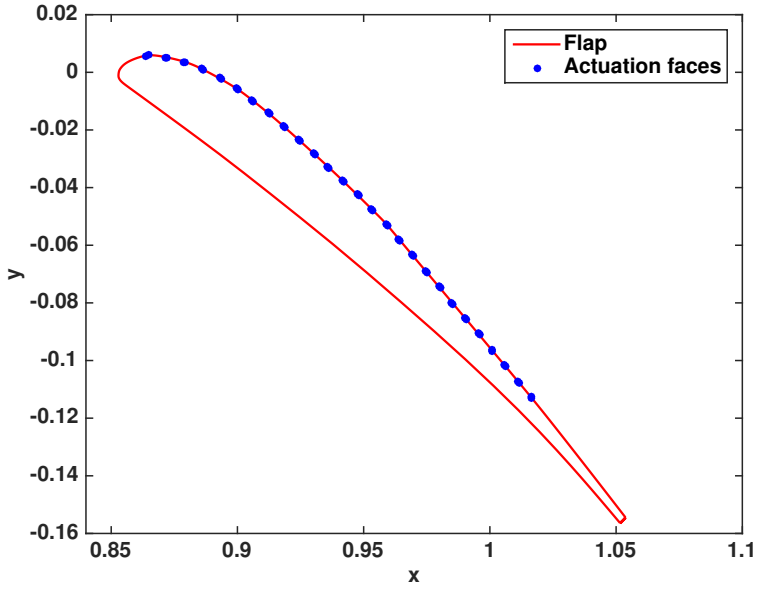


Figure 5: Actuation slot distribution on the suction side of the flap of a three-element airfoil.

In order to delay the separation and to enhance the lift coefficient of the flap, synthetic jet actuation is applied at 54 faces on the suction side of the flap. To start the optimization, initially the non-dimensional values of the actuation parameters at all the faces are chosen as $A = 0.2$, $F = 2.0$, $\beta = 90^o$ and $t_0 = -0.125$. At a blowing angle of 90° the flow over the flap is perturbed perpendicular to the local faces. The frequency of the initial actuation is fixed to the estimated frequency of the vortex shedding of the un-actuated baseline flow. Time accurate simulations are performed with the initial control starting from a fully developed baseline flow until the non-dimensional time $T = 28$, which amounts to $40,000$ unsteady time iterations. Figure 7 shows the time history of the lift coefficient. Compared to the baseline flow with $\overline{C}_l = 2.2356$, the initial actuation has resulted in reasonable improvement in the lift coefficient with $\overline{C}_l = 2.4936$.

Our objective is to find the optimal actuation that results in maximum lift coefficient. It can be observed that the lift profile for the actuated flow shows a typical initial transient behavior, after which the flow settles down to a periodic state. For the objective function evaluation, we neglect this initial transient and compute the time-averaged lift coefficient over a sufficiently large interval of fully developed flow starting from $30,000$ time iterations to $40,000$ iterations. It can be observed that the objective function interval consists of around 14 cycles of actuation, which are good enough for a meaningful average. The discrete objective function and the corresponding Lagrangian can then be written as

$$\begin{aligned} J = \overline{C}_l &= \frac{1}{10,000} \sum_{n=30,001}^{40,000} C_l^n \left(\boldsymbol{U}^n, \boldsymbol{\alpha} \right) \\ L &= \frac{1}{10,000} \sum_{n=30,001}^{40,000} C_l^n - \sum_{n=1}^{40,000} \left[\left(\overline{\boldsymbol{U}}^n \right)^T \left(\boldsymbol{U}^n - \boldsymbol{G}^n \right) \right] \end{aligned} \tag{7}$$

The control variables are the actuation amplitudes, blowing angles and phase shifts at 54 actua-

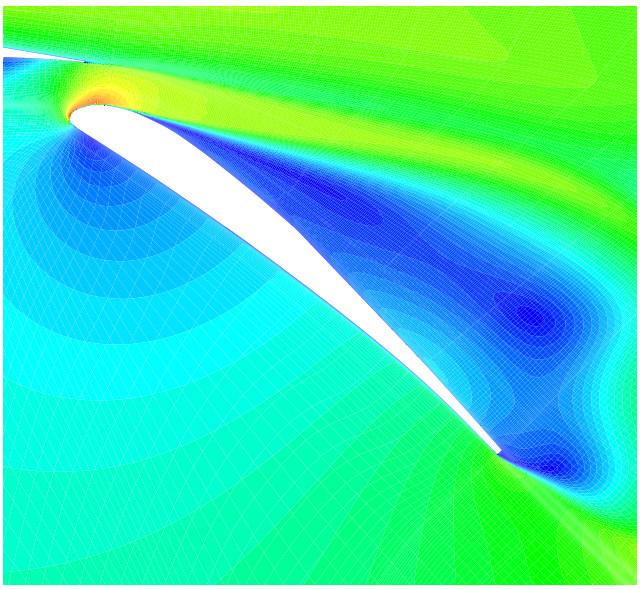


Figure 6: Baseline flow: Contours of the velocity magnitude on the flap of the three-element airfoil.

tion faces. This results in a total of 162 control variables. We now present the numerical results for optimal active flow control based on a gradient based algorithm, evolutionary algorithm and hybrid optimization approach.

5.1 Optimal Actuation based on the Gradient-based Optimization

To find the optimal actuation, the discrete adjoint URANS solver is combined with the BFGS optimization algorithm. Figure 8 shows the time-history of the lift coefficient for the optimal actuation. Compared to the un-actuated baseline flow, optimal actuation enhanced the mean lift by around 19% to a value of $\overline{C}_l = 2.6534$. Figure 9 shows the contours of velocity magnitude for the optimal case at the crest of a periodic oscillation in the lift coefficient. It can be observed that the separation point has moved further downstream of the shoulder region.

5.2 Optimal Actuation based on the Evolutionary Algorithm

Figure 8 shows the variation of the lift coefficient based on the the Evolutionary Algorithm. This approach resulted in very significant enhancement in the lift coefficient. Compared to the baseline flow, the mean lift is increased by 63% to a value of $\overline{C}_l = 3.6402$. Figure 10 shows the contours of velocity magnitude at the crest of a periodic oscillation in the lift coefficient. It can be observed that the flow is fully attached as the separation is completely suppressed.

5.3 Optimal Actuation based on the Two-level Hybrid Approach

The optimal actuation obtained from the Evolutionary Algorithm is then used as an initial condition at all the actuation faces on the suction side of the flap. Using this initial setting, the discrete adjoint URANS solver is then employed along with the BFGS algorithm to further enhance the lift coefficient. Figure 8 shows that the two-level approach has yielded a reasonable

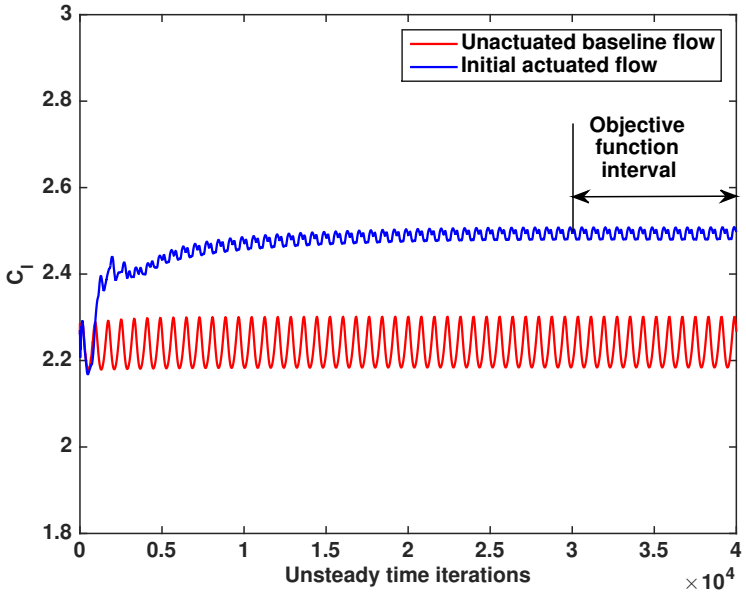


Figure 7: Comparison of the lift coefficient for the un-actuated baseline flow and the initial actuated flows.

improvement in the lift coefficient compared to Evolutionary Algorithm. The mean lift coefficient is observed to be 3.7266, which is almost 67% more than the baseline value. Figure 10 shows the corresponding contours of velocity magnitude. As a result, the two-level approach yields an improvement of the mean drag around 40% compared to the classical gradient-search. It can be concluded that, better initial values for the actuation variables, which are obtained by the global optimization brings a significant improvement as far as the optimization results are concerned. Table 1 shows a comparison of the mean-lift coefficient for the un-actuated, initial actuated and optimal actuated flows.

In the present work, the levels of the hybrid approach were only performed once such that the adjoint based gradient search stage has no effect on the initial global search using the EA. It must also be noted that the EA uses a reduced design space that is case-specific and somehow arbitrary. Therefore, choosing the reduced design space influences largely each stage's potential and emphasizes the case dependency of each stage's individual performance. More complex cases should show a stronger benefit of the hybrid algorithm.

Flow	Mean lift coefficient $\overline{C}_l$
Un-actuated baseline flow	2.2356
Initial actuated flow	2.4936
Optimal actuation based on gradient search approach	2.6534
Optimal actuation based on EA	3.6402
Optimal actuation based on hybrid approach	3.7266

Table 1: Comparison of the mean lift for the un-actuated, initial actuated and optimal actuated flows.

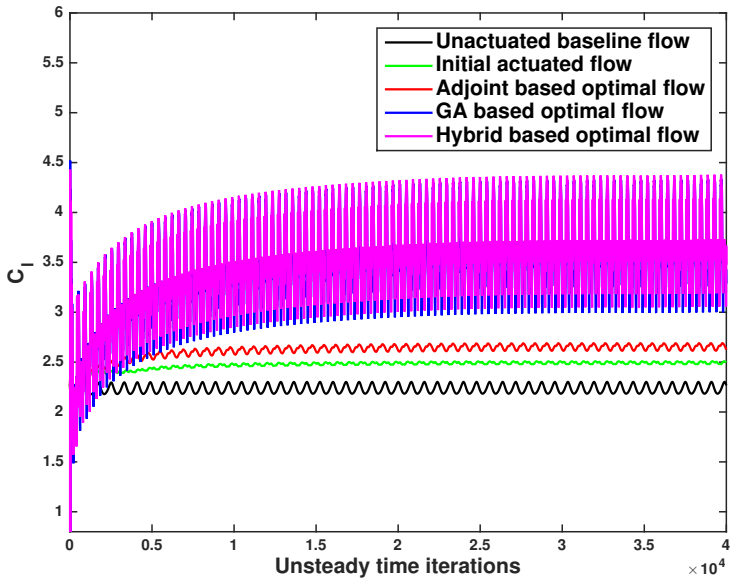


Figure 8: Comparison of the lift coefficient for the un-actuated baseline flow, initial actuated flow, optimal actuated flows based on discrete adjoint approach, evolutionary and hybrid optimization algorithms respectively.

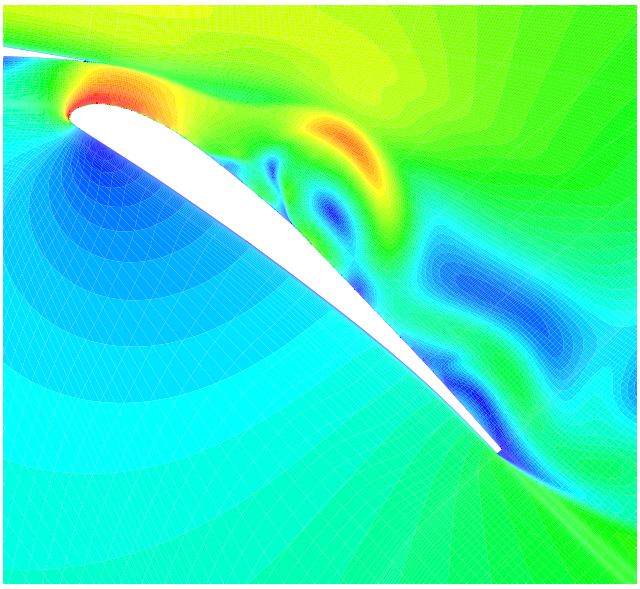


Figure 9: Optimal active flow control based on a gradient based optimization algorithm combined with discrete adjoints. Contours of the velocity magnitude on the flap of the three-element airfoil.

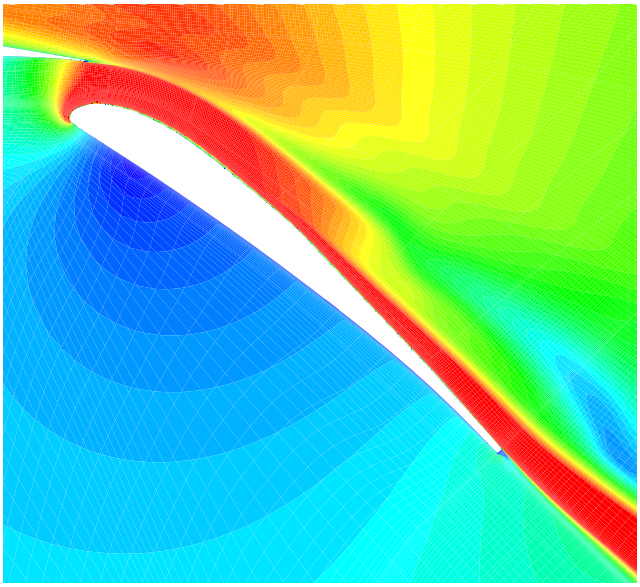


Figure 10: Optimal active flow control based on an evolutionary algorithm. Contours of the velocity magnitude on the flap of the three-element airfoil.

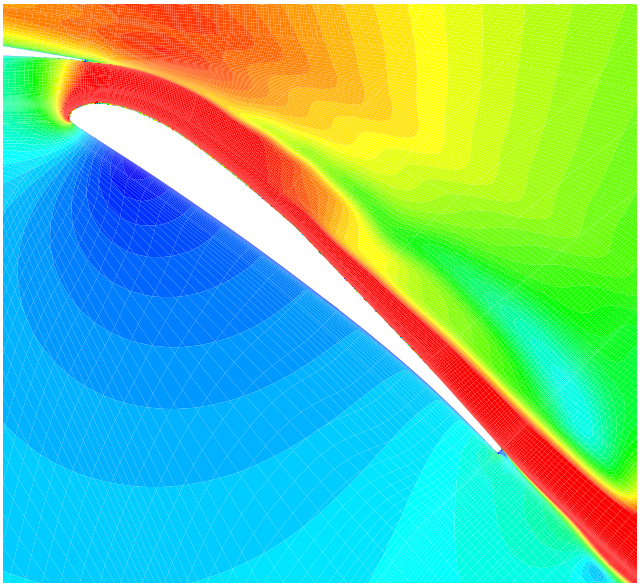


Figure 11: Optimal active flow control based on the two-level hybrid optimization algorithm. Contours of the velocity magnitude on the flap of the three-element airfoil.

6 CONCLUSIONS

In this paper, we presented a two-level approach for the optimal active flow control by combining an evolutionary algorithm based global optimization method with a gradient search method using unsteady discrete adjoints. The discrete adjoint solver for the governing incompressible URANS equations was developed by applying Algorithmic Differentiation (AD) techniques. The adjoint solver retains the full functionality of the underlying non-linear URANS solver. The suggested two-level method was then applied to find the optimal synthetic jet actuation of a three-element airfoil configuration at a Reynolds number of 10^6 and an angle of attack of 6°. Numerical results have shown that the two-level approach had yielded 67% improvement in the mean lift compared to the un-actuated baseline flow and 40% increase over the adjoint based gradient search algorithm.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the computing time granted by the Allianz für Hochleistungsrechnen Rheinland-Pfalz (AHRP), Germany.

REFERENCES

[1] M. Gad-el Hak. Modern developments in flow control. *Applied Mechanics Reviews*, 49:365–379, 1996.

[2] E. Tinapp and W. Nitsche. Separation control on a high-lift configuration by periodic excitation. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design (NNFM)*, 77:42–49, 2002.

[3] M. Schatz, E. Thiele, R. Petz, and W. Nitsche. Separation control by periodic excitation and its application to a high lift configuration. *AIAA Paper 2004-2507*, 2004.

[4] L. Pack, C. Yao, and A. Seifert. Application of excitation from multiple locations on a simplified high-lift system. *AIAA Paper 2004-2324*, 2004.

[5] R. Petz, W. Nitsche, M. Schatz, and F. Thiele. Increasing lift by means of active flow control on the flap of a generic high-lift configuration. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design (NNFM)*, 92:257–264, 2006.

[6] M. Bauer, I. Peltzer, W. Nitsche, and B Gölling. Active flow control on an industry-relevant civil aircraft half model. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design (NNFM)*, 108:95–107, 2010.

[7] B. Günther, F. Thiele, W. Petz, R. Nitsche, J. Sahner, T. Weinkauf, and H.C. Hege. Control of seperation on the flap of a three-element high-lift configuration. *AIAA paper 2007-0265*, 2007.

[8] T. Höll, A. K. V. Job, P. Giacopinelli, and F. Thiele. Numerical study of active flow control on a high-lift configuration. *Journal of Aircraft*, 49(5):1406–1422, 2012.

[9] V. Ciobaca, T. Kühn, R Rudnik, M. Bauer, B. Gölling, and W. Breitenste. Active flow-separation control on a high-lift wing-body configuration. *Journal of Aircraft*, 1:56–72, 2013.

[10] A. Seifert, T. Bachar, D. Koss, M. Shepshelovich, and I. Wyganski. Oscillatory blowing: A tool to delay boundary-layer separation. *AIAA Journal*, 31(11):2052–2060, 1993.

[11] B. L. Smith and A. Glezer. The formation and evolution of synthetic jets. *Physics of Fluids*, 10(9):2281–2297, 1998.

[12] A. Jameson, L. Martinelli, and N. Pierce. Optimum aerodynamic design using the Navier-Stokes equations. *Theor. Comput. Fluid Dyn.*, 10(1-4):213–237, 1998.

[13] A. Carnarius, B. Günther, F. Thiele, D. Wachsmuth, F. Tröltzsch, and J. C. de los Reyes. Numerical study of the optimization of separation control. *AIAA Paper 2007–58*, 2007.

[14] E. J. Nielsen and W. T. Jones. Integrated design of an active flow control system using a time-dependent adjoint method. *Mathematical Modelling of Natural Phenomena*, 6(3):141–165, 2011.

[15] A. Nemili, E. Özkaya, N. Gauger, A. Carnarius, and F. Thiele. Optimal control of unsteady flows using discrete adjoints. *AIAA Paper 2011–3720*, 2011.

[16] A. Nemili, E. Özkaya, N. Gauger, F. Kramer, A. Carnarius, and F. Thiele. Discrete adjoint based sensitivity analysis for optimal flow control of a 3D high-lift configuration. *AIAA Paper 2013–2585*, 2013.

[17] A. Griewank and A. Walther. *Evaluating derivatives: Principles and techniques of Algorithmic Differentiation*. SIAM, 2008.

[18] J. E. V. Peter and R. P. Dwight. Numerical sensitivity analysis for aerodynamic optimization: A survey of approaches and applications. *Computers & Fluids*, 39(3):373391, 2010.

[19] E. J. Nielsen and W. K. Anderson. Aerodynamic design optimization on unstructured meshes using the Navier-Stokes equations. *AIAA Journal*, 37(11):185–191, 1999.

[20] D. Quagliarella and A. Della Cioppa. Genetic algorithms applied to the aerodynamic design of transonic airfoils. *Journal of Aircraft*, 32(4):889–891, 1995.

[21] A. Oyama, S. Obayashi, and K. Nakahashi. Transonic wing optimization using genetic algorithm. *AIAA Paper 97-1854*, 1997.

[22] J. Ahn, H.-J. Kim, D.-H. Lee, and O-H. Rho. Response surface method for airfoil design in transonic flow. *Journal of Aircraft*, 38(2):231–238, 2001.

[23] M. J. D. Powell. Algorithms for approximation. chapter Radial Basis Functions for Multivariable Interpolation: A Review, pages 143–167. Clarendon Press, New York, NY, USA, 1987.

[24] Z.-H. Han, K.-S. Zhang, W.-P. Song, and Z.-D. Qiao. Optimization of active flow control over an airfoil using a surrogate-management framework. *Journal of Aircraft*, 47:603–612, 2010.

[25] J.-W. Yim, B.-J. Lee, and C. Kim. Exploring multi-stage shape optimization strategy of multi-body geometries using Kriging-based model and adjoint method. *Computers & Fluids*, 68:71 – 87, 2012.

[26] A. Nemili, E. Özkaya, N. Gauger, F. Kramer, T Höll, and F. Thiele. Optimal design of active flow control for a complex high-lift configuration. *AIAA Paper 2013–2515*, 2014.

[27] L. Hascoët and V. Pascual. The Tapenade Automatic Differentiation tool: Principles, Model, and Specification. *ACM Transactions On Mathematical Software*, 39(3), 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TRANSITION FROM 2D CONTINUOUS ADJOINT LEVEL SET TOPOLOGY TO SHAPE OPTIMIZATION

J.R.L. Koch[1], **E.M. Papoutsis-Kiachagias**[1], **and K.C. Giannakoglou**[1]

[1]National Technical University of Athens (NTUA),
School of Mech. Eng., Parallel CFD & Optimization Unit
Athens, Greece
e-mail: kochjamesr@gmail.com, vaggelisp@gmail.com, kgianna@central.ntua.gr

Keywords: Adjoint-Based Optimization, Constrained Topology Optimization, Shape Optimization, Level Set Method, Topology to Shape Transition

Abstract. *The processes of topology and shape optimization are well known methods in the field of fluid mechanics. Although successful in their own rights, it is conceivable that the two methods will find choicest solutions in tandem: i.e. if shape optimization were able to improve a topological solution. Conjoining the two methods in this manner is not straightforward, however, since there is no existing process to connect one to the other. Toward this goal, a novel transitional process is proposed to process level set topology solutions obtained using the continuous adjoint method such that a shape optimization loop using the continuous adjoint can be initialized, run and ultimately produce a refined, parameterized solution. First, the topology optimization process is enhanced using the level set method to both maintain an explicit description of the interface between the solid and fluid topological domains and prevent the formation of fluid or solid islands which would not be viable for manufacturing. The interface is then fitted with Non-Uniform Rational B-Splines (NURBS) through application of sensitivities garnered from the solution of an auxiliary optimization problem which aims at reducing the difference between the signed distance fields generated about each NURBS curve and its corresponding interface section. A body-fitted mesh is generated for the geometry defined by the fitted NURBS, allowing a shape optimization loop to be initiated. The parameterized result of the topology to shape transition process will be compared to that of shape optimization in two 2D cases with internal, incompressible fluid flows.*

This research was funded by the People Programme (ITN Marie Curie Actions) of the European Union's H2020 Framework Programme (MSCA-ITN-2014-ETN) under REA Grant Agreement no. 642959 (IODA project). The first author is an IODA Early Stage Researcher.

1 INTRODUCTION

Topology optimization (TopO) was first proposed in [1] for the design of structural mechanics and was expanded to the field of fluid mechanics through the introduction of a blockage term (β) into the Stokes [2, 3] and laminar [4] flow equations which is treated as a design variable per grid cell and varied to minimize an objective function. On the other hand, shape optimization (ShpO) was initially introduced as shape control technique to reduce an objective function, [5], and is expanded to fluid mechanics using boundary mesh displacement to alter the original boundary shape of the case and, thus, its flow solution.

Both the TopO and ShpO processes require the ability to affect the current solution of a problem through design variable alteration in order to decrease (increase) the provided objective function(s). To this end, the adjoint method is employed to supply objective function sensitivities, effectively telling the design variable set in which way they should be altered. Continuous adjoint has been extensively developed for both ShpO and TopO for laminar and turbulent flows, [5, 6, 7, 8, 9], and is applied for the TopO and ShpO processes used in this work.

While TopO is commonly used in fluid mechanics to develop optimal flow paths between prescribed inlets and outlets according to the objective function(s) defined by the designer, ShpO is only capable of altering known boundaries of a geometry in order to obtain an optimal solution, but uses far fewer design variables (i.e., the boundary nodal or control point coordinates instead of an internal field) to do so. Ideally, a solution found using TopO could be used to begin a ShpO process which would produce a parameterized, manufacturable CAD model. However, solutions of TopO do not, on their own, contain the surface information required for ShpO to initialize; information which would be available if a fluid-solid interface (FSI) were explicitly defined by the TopO solution. Furthermore, TopO solutions found can also suffer from island formation and the development of grey-zones in which the FSI is ambiguous.

To address these issues, a process which conjoins the theories of both the continuous adjoint topology and level set (LS) optimization methods for incompressible viscous fluid simulations is devised and hereon associated with the term 'TopO'. The LS method is a conceptual framework, first proposed in [10], which maps moving interfaces through the introduction of a LS field (LSF) that corresponds to a signed-distance field. Integration of the LS method into TopO is achieved by making β a function of the LSF: the domain is modeled as solidified where the LSF is negative and fluidized when positive, making the zero-value contour of the LSF the FSI by definition. FSI progression toward an optimal solution occurs via adjoint sensitivity convection through alteration of cell-center LSF values and subsequent application of an accurate, fast-marching reinitialization algorithm.

The TopO-to-ShpO transition process begins by subdividing the FSI from TopO such that each division pertains to paired inlet-outlet connections. These FSI segments are used to initialize new individual LSFs which act as target solutions to a field-matching algorithm which tries to iteratively fit a NURBS curve to each segment. This is done by building a narrow band (NB) LSF about each NURBS curve, comparing it to the target LSF and, then, generating sensitivities for moving each NURBS control point (CP) based on their difference. A body-fitted mesh is generated from this NURBS fitting and is used to initialize an adjoint-driven ShpO process, which in turn generates a parameterized solution. This paper will apply the proposed method for conjoining the TopO and ShpO processes to two 2D test cases with the goal of minimizing total pressure losses in ducted flow using volumetric constraints. Both the cases and the in-house code pertaining to the coupling process are implemented within OpenFOAM2.2.1 [11].

2 TOPOLOGY OPTIMIZATION

2.1 Flow Equations

In this paper, the modeled fluid flow is governed by the steady-state Navier-Stokes equations for incompressible, laminar flows. The blockage formulation of TopO requires the addition of a term which contains a function of the blockage variable $(0 \leq \beta \leq 1)$ into the momentum equation (eq. 1b). If $f(\beta)$ becomes comparatively large, the added term becomes dominant, forcing the velocity toward zero and effectively rendering the domain solid. Conversely, if this term is comparatively small, the velocity is uninhibited and the domain remains fluid. The flow equations are

$$R_p = -\frac{\partial v_j}{\partial x_j} = 0 \tag{1a}$$

$$R_{v_i} = v_j \frac{\partial v_i}{\partial x_j} + \frac{\partial p}{\partial x_i} - \frac{\partial}{\partial x_j}\left[\nu\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] + \underbrace{f(\beta)v_i}_{Blockage\ Term} = 0 \tag{1b}$$

where $f(\beta) = \beta\beta_{MAX}$ and β_{MAX} is a large and positive user-defined value. The usual primal boundary conditions associated with internal aerodynamics–constant inlet velocity, constant outlet pressure and no-slip condition along the wall–are imposed for all cases.

2.2 Adjoint Equations, Boundary Conditions and Sensitivities

In order to properly discuss the formulation of the adjoint equations, the basic TopO optimization problem must be examined. In this paper, the objective function to be minimized is that of the volume-averaged total pressure losses between the inlet(s), S_I, and outlet(s), S_O,

$$F = -\int_{S_I}\left(p + \frac{1}{2}v_k^2\right)v_i n_i dS - \int_{S_O}\left(p + \frac{1}{2}v_k^2\right)v_i n_i dS \tag{2}$$

where n_i are the components of the outward normal vector. To account for the possible inclusion of a constraint C into the TopO process, the objective function F to be minimized is re-defined as the Lagrangian objective function L according to the Augmented Lagrange Multiplier (ALM) method [12]. Constraints are discussed in Section 2.4. The optimization problem and corresponding Lagrangian function are defined as

$$\begin{aligned} &\min(F)\ ,\ \text{subject to}\ \ c = 0 \\ &L = F - \lambda c + wc^2 \end{aligned} \tag{3}$$

where c and λ are the constraint function and Lagrangian multiplier, respectively, and w is a dynamically changing weighting scalar [12]. If the constraint is not imposed $L \equiv F$. The Lagrangian function to be minimized is further augmented by the flow equation residuals as follows

$$L_{aug} = L + \int_\Omega q R_p d\Omega + \int_\Omega u_i R_{v_i} d\Omega \tag{4}$$

in which q and u_i are the adjoint pressure and velocity components, respectively. After lengthy derivation, [13], the variation of eq. 4 with respect to (w.r.t.) the design variable(s) β becomes

$$\begin{aligned} \frac{\delta L_{aug}}{\delta \beta_m} = &\int_\Omega R_q \frac{\partial p}{\partial \beta_m} d\Omega \ + \int_\Omega R_{u_i}\frac{\partial v_i}{\partial \beta_m} d\Omega \ + \int_\Omega v_i u_i \frac{\partial f(\beta)}{\partial \beta_m} d\Omega \\ &+ \int_S \mathcal{BC}_1 \frac{\partial p}{\partial \beta_m} dS + \int_S \mathcal{BC}_{2,i}\frac{\partial v_i}{\partial \beta_m} dS - \int_S \frac{\partial}{\partial \beta_m}\left[\nu\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] u_i n_j dS \end{aligned} \tag{5}$$

where m is the cell index and S is the boundary of the computational domain Ω. The R and BC terms in eq. (5) will become the adjoint field equations and boundary conditions, respectively, and are derived through eliminating all terms with field integrals containing partial derivative(s) dependent on the β design variable(s) by setting their multipliers, or, in the case of the BC terms, portions of their multipliers, against zero. It should be noted that these partial derivatives are technically total derivatives: $\delta \equiv \partial$ for TopO since the computational grid remains unchanged when a design variable is changed [13].

The adjoint continuity and momentum field equations are defined as

$$R_q = \frac{\partial u_j}{\partial x_j} = 0 \tag{6a}$$

$$R_{u_i} = -v_j \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \frac{\partial q}{\partial x_i} - \frac{\partial}{\partial x_j} \left[\nu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right] = 0 \tag{6b}$$

with appropriate boundary conditions derived from $\mathcal{BC}_1$ and $\mathcal{BC}_2$ (see [13]). After satisfaction of the field equations and their boundary conditions, the remaining terms of eq. 5 define the sensitivity derivatives of L_{aug} (here-on referred to as simply L) for laminar flows w.r.t. the blockage design variable(s)

$$\frac{\delta L}{\delta \beta_m} = \int_\Omega v_i u_i \frac{\partial f(\beta)}{\partial \beta_m} d\Omega = v_i^m u_i^m \Omega^m \beta_{MAX} \tag{7}$$

where Ω^m is the volume of the m-th grid cell.

2.3 Level Set Topology Formulation

Use of the blockage formulation of TopO creates two dilemmas w.r.t. the process of transitioning to ShpO described in Section 3. The first is the lack of an explicit definition between the solid and fluid topological domains. Depending on the value of β, the topological field will be defined as solid ($\beta = 1$), fluid ($\beta = 0$), or, more likely, slushy grey-scale β values that are neither solid or fluid ($0 < \beta < 1$). Such grey-scale regions are problematic from a manufacturing point of view, since there is no way to know where the blockage-based TopO process would choose the FSI within the grey-scale domain. Generally, the issue of β grey-scale has been treated through a cut-off approach: if β is below a certain (small) value then it is considered fluid while for larger values it is considered solidified. While convenient, this approach does not address the larger issue at hand: that the solution has not defined a true FSI. The second dilemma inherent to the blockage-based TopO formulation is the development of fluid or solid islands within otherwise homogeneous domains of the opposite type, and can be attributed to the fact that the adjoint sensitivities are applied across the entire domain. Since they have no relevance w.r.t. manufacturing, islands should not be considered when defining the FSI.

To address both these issues simultaneously, the level set (LS) method was employed as a β filter. The level set is a differential equation-driven method in which the design domains and their interfaces Γ (here Γ_{FSI}) are represented in terms of implicit functions which are monotonic and differentiable on the interface [14]. The LS method is used to manipulate its field variable ϕ, i.e. a signed distance field which must have a gradient magnitude of 1 everywhere,

$$\|\nabla \phi\| = 1 \tag{8}$$

such that $\phi_i = \{\text{negative } \forall i \in \Omega_{\text{Solid}},\ \text{zero } \forall i \in \Gamma_{\text{FSI}},\ \text{positive } \forall i \in \Omega_{\text{Fluid}}\}$

To move the FSI toward the optimal solution, the LSF is subjected to a process of convection such that the total derivative remains zero for the Lagrangian reference system. If eq. 8 holds and the convecting velocity V_i of Γ_{FSI} is defined as its unit normal multiplied by a scalar quantity, then the convection equation is derived as follows

$$\frac{\delta\phi}{\delta t} = \frac{\partial\phi}{\partial t} + \frac{\partial\phi}{\partial x_i}\frac{\partial x_i}{\partial t} = \frac{\partial\phi}{\partial t} + V_i\frac{\partial\phi}{\partial x_i} = \frac{\partial\phi}{\partial t} + V_n = 0$$

where V_n, the 'normal' convecting velocity of Γ_{FSI}. V_n is redefined as G_{Tot} and is comprised of two different contributions: those from the TopO sensitivities derived from the objective function and those from the constraint C (if imposed). The discrete form of the equation of convection is

$$\phi_m^{n+1} = \phi_m^n - (\Delta\tau)G_{Tot_m} \tag{9}$$

where

$$G_{Tot_m} = \underbrace{G_{F_m}}_{F\ Term} + \underbrace{G_{C_m}}_{C\ Term} = \frac{\delta L}{\delta\phi_m} \tag{10}$$

and $\Delta\tau$ and m are a convection step size and cell index, respectively. The F and C terms found in eq. 10 will be discussed in Section 2.4.

After eq. 9 is explicitly applied to Γ_{FSI}, eq. 8 does not necessarily hold: the ϕ field must undergo a signed-distance correction procedure (SDCorr) after each convection application. The method employed for the work in this paper corrects $\forall\phi_i \in \Gamma_{\text{FSI}}$ using the original ϕ values for those cells and then conducts a Fast Marching Method (FMM) to solve for the remainder of the ϕ domain [15]. If the computational domain is large, performing this FMM for the entire ϕ domain can be expensive. Thus, a narrow-band (NB) technique is employed in this work: the ϕ values are subject to SDCorr only if they are within the NB. Although this means that the SDCorr must be performed after each convection step (Γ_{FSI} moves and thus the NB cell set must be reassessed), the overall computational cost of the procedure becomes negligible if the NB is small.

The LS field can be trivially initialized from any given β field through the relation

$$\phi_m = 1 - 2\beta_m \tag{11}$$

which defines the zero-ϕ isoline as the FSI. Eq. 11 is an auxiliary, one-way expression: once the LSF has been convected, an inverse application of this equation will not give an accurate β field since the LS field has undergone a SDCorr and $0 < \beta < 1$ must hold (the corrected ϕ values would create β values above and below 0 and 1). Thus, after initialization the relationship between the LSF and β is inverted: β depends on ϕ according to a piecewise continuously differentiable NB sigmoidal Heavyside relationship ($\beta(\phi_m) = H(\phi_m)$). This Heavyside function differentiates into the smoothed Dirac function $D(\phi_m)$, i.e.,

$$\frac{\delta\beta(\phi_n)}{\delta\phi_m} = D(\phi_m)\delta_m^n \tag{12}$$

where δ_m^n is a Kronecker delta. This relationship is used in Section 2.4 to recast the objective function's TopO adjoint sensitivities into TopO convecting velocities.

The entire TopO process can be summarized as follows:

1. Initialize ϕ from a given β field via eq. 11.
2. Solve eqs. 1 and 6; get new field sensitivities via eq. 7.
3. Compute the LS convecting velocities G_{Tot} via eq. 10 (see Section 2.4) and convect within Γ_{FSI} via eq. 9.
4. Conduct the SDCorr process and get the new NB ϕ values based on the newly convected zero-ϕ isoline.
5. Compute $D(\phi_m)$ via eq. 12 for the next iteration's convection and set the new β field with $\beta(\phi_m) = H(\phi_m)$; return to step 2.

2.4 Level Set Convecting Velocity Term Formulation

The TopO formulation discussed in Section 2.3 requires the set of normalized convecting velocities G_{Tot} which are defined in eq. 10 and expanded as follows.

$$G_{Tot_m} = \frac{\delta L}{\delta \phi_m} = \underbrace{\frac{\partial L}{\partial \beta_n}\frac{\delta \beta_n}{\delta \phi_m}}_{F\ Term} + \underbrace{\frac{\partial L}{\partial c}\frac{\delta c}{\delta \phi_m}}_{C\ Term} \tag{13}$$

G_{Tot} is effectively comprised of two different contributions: those from the TopO adjoint sensitivities derived from the defined objective function and those from the constraint. These two contributions are discussed below and recast for the LS formulation as necessary.

Convecting Contribution *G_F: Objective Function*
Without the constraint applied, only the $F\ Term$ remains in eq. 13, which is eq. 7: the TopO sensitivities become the LS convecting velocity. To recast the TopO sensitivities as LS sensitivities, the fact that β is a function of ϕ is exploited through the relationship found in eq. 12 to give $\frac{\delta L}{\delta \phi_m} = D(\phi_m)\frac{\partial L}{\partial \beta_m}$: the TopO sensitivities are simply multiplied by the LS Dirac function. Physically, this means that the TopO sensitivities are only valid as LS convecting velocities within the NB. Thus, the $F\ Term$ in eq. 10 is

$$G_{F_m} = D(\phi_m) v_i^m u_i^m \Omega^m \beta_{MAX} \tag{14}$$

Convecting Contribution *G_C: Fluid Volume Constraint*
LS convecting velocity contributions generated from the constraint is contained within the $C\ Term$ of eq. 13. The constraint considered in this paper is an ALM equality constraint which limits the total volume-averaged sum of the β field, requiring that a certain percentage $\forbar{V}_{Fluid,Tar}$ of Ω be Ω_{Fluid}. This constraint is useful from a production perspective since limitation of the allowed fluid volume can help meet material usage restrictions. Without TopO sensitivities, only the $C\ Term$ of eq. 13 remains and is defined as

$$G_{C_m} = (-\lambda + 2wc)\frac{\delta c}{\delta \phi_m} \tag{15}$$

$$c = \frac{1}{2}(\forbar{V}_{Fluid} - \forbar{V}_{Fluid,Tar})^2, \quad \forbar{V}_{Fluid} = \frac{\int_\Omega (1 - H(\phi_m))d\Omega}{\int_\Omega d\Omega} \tag{16}$$

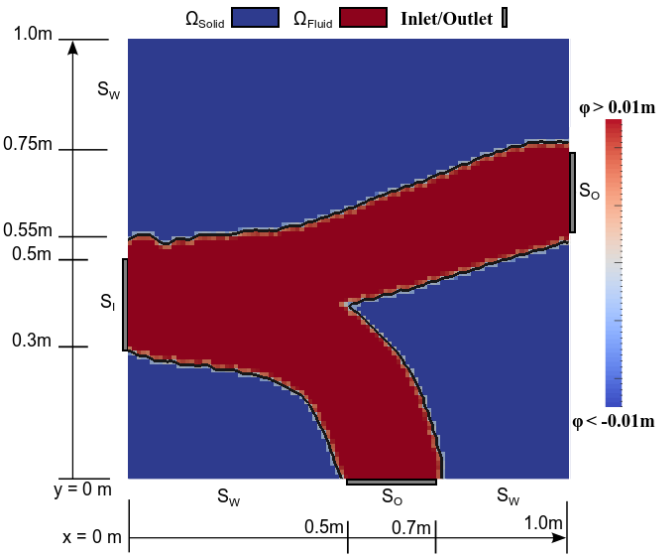


Figure 1: The TopO minimal total pressure loss ϕ solution and geometry of the TToST example case, using a NBHW of 0.01m. The Γ_{FSI} of the solution can be seen as a black line.

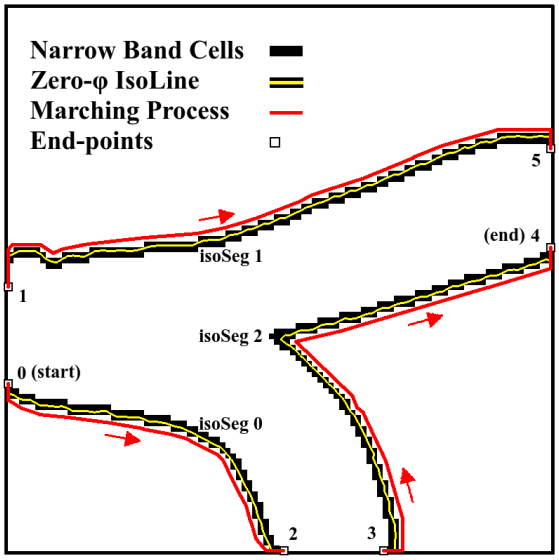


Figure 2: The marching segmentation process starts with end-point 0 and ends at end-point 4, generating points along Γ_{FSI} and assigning them to the appropriate isoPts list of each isoSeg.

3 2D TOPOLOGY-TO-SHAPE TRANSITION

The process of ShpO requires a parameterized curve or surface input in order to build a computational mesh and generate a refined, parameterized solution. In the present work, these inputs are comprised of NURBS curves. The goal of the proposed transition process is to automatically generate and fit, as closely and efficiently as possible, NURBS curves to TopO solutions so as to have a ShpO input which matches the TopO case geometry and can be fed directly into a meshing program to begin the ShpO process. This approach is only feasible due to the explicitly defined FSI found by the TopO formulation described in Section 2.3, as it provides a known target curve (the zero-ϕ isoline) to be fitted. The procedure for the proposed 2D TopO-to-ShpO transition (TToST) is explained through its application to a representative test case.

3.1 TToST Explanatory Test Case: 2D Single Inlet, Double Outlet

The case used to exemplify the TToST process is a laminar, 2D OpenFOAM (one cell in the z direction) case with one inlet to be connected to two outlets. The case has an orthogonal Cartesian mesh of 100x100 cells. β is initially specified as 0 at the cells which owned by (are adjacent to) S_I and S_O, and as 1 everywhere else. A narrow band half width (NBHW), i.e. the normal distance from Γ_{FSI} to the limits of the NB, of $0.01m$ was selected for the LS process, and the fluid volume constraint was specified with $V\!\!\!-_{Fluid,Tar} = 0.3$ (see Section 2.4). The case geometry and ϕ field solution found by the TopO process can be seen in fig. 1.

3.2 TToST Step1: Isoline Segmentation

In order to obtain a set of target curves for the NURBS fitting processes, the TopO solution's Γ_{FSI} must be decomposed into as many isoline segments (isoSegs) as there are NURBS end-point pairs to be connected. The NURBS end-points are defined at the centers of the edges

shared by boundary face pairs which have a member in both the S_W and $S_I||S_O$ face sets. The normal vectors of each point (used to define the initial, positive normal direction for a NURBS curve starting there) are defined as pointing from the S_W cell to the $S_I||S_O$ cell. For the considered case there are six end-points, requiring three isoSegs to be generated. The isoSegs are generated by interpreting an auxiliary, binary field which is 1 for cells which contain the Γ_{FSI} and 0 otherwise. During the generation process, points which lie on the zero-ϕ isoline are gathered so as to have an ordered set of points against which to fit NURBS curves in Section 3.4. The process for generating the isoSeg and their isoPts starts from an end-point and traverses the case's boundaries and Γ_{FSI} until another, unused end-point is reached, and then repeated until no end-points remain to be considered. This process is sketched in fig. 2.

3.3 TToST Step 2: ϕ Target Field Generation

The next step in the TToST is to generate a set of 'target' signed distance fields ϕ_{Tar} from each isoSeg's Γ_{FSI} cells through a SDCorr process similar to that used in the LS NB calculation (see Section 2.3). Unlike during the TopO process, the SDCorr must be performed throughout the entire domain: these target fields are used as a basis for distance comparison in step 4 of the TToST (Section 3.5) to quantify how far away from its Γ_{FSI} segment the generated NURBS is, and must therefore exist wherever the NURBS could exist within Ω. An optimization problem is solved in steps 3 and 4 of the TToST to minimize the difference between these ϕ_{Tar} fields and the NB ϕ fields generated about each NURBS curve. The use of the ϕ_{Tar} fields during this fitting process means that each NURBS curve only sees the isoSeg it is to be matched with and will not attempt to fit the other isoSegs of the case.

3.4 TToST Step 3: NURBS Least Square Fit and Parameter Search Algorithm

Once the end-points, starting normals, ordered isoPt lists and ϕ_{Tar} fields have been found, the unique initialization and fitting process for each NURBS can begin. To initiate the iterative Field Matching process described in Section 3.5 P_I points are generated on the NURBS curve, where P_I is the number of isoPts for the corresponding isoSeg to be matched, and a Least Squares Fitting problem (LSqF) is solved. The LSqF is supplied with a user-defined number of control points (nCPs) and degree which are identical for all NURBS to be fitted; 8 CPs and degree 4 for the considered example case. Performed with fixed initial nCPs and degree, the LSqF can be poor if the isoSeg has high curvature (isoSeg2) or begins away from the specified end-points (isoSeg1). A plot of the resulting NURBS curves, their CPs and the isoPts after this LSqF process can be seen in fig. 3. To mitigate the effect of these matching errors, a Parametrization Search Fit (PSF) is performed for each NURBS curve after its initialization in which the nCPs and degree are altered to decrease the error of the LSqF. The PSF algorithm set degrees 4, 3 and 2 and nCPs 9, 10 and 14 for NURBS 0, 1 and 2, respectively. A plot of the resulting NURBS curves, their CPs and the isoPts after this PSF process can be seen in fig. 4.

3.5 TToST Step 4: NURBS Field Matching Fit Algorithm

The final step in the TToST process is an iterative field comparison between a generated NURBS NB distance field, ϕ_N and the portion of its ϕ_{Tar} field within its NB which attempts to minimize $\Delta\phi_{NB}$, the aggregate difference between the two, through CP, and thus NB, displacement. Displacement for each CP is dependent on sensitivities which quantify how the current $\Delta\phi_{NB}$ relationship would change if the CP were to be moved. Thus, to properly explain the Field Matching Fit (FMF) algorithm, the formulation of its CP sensitivity generation must

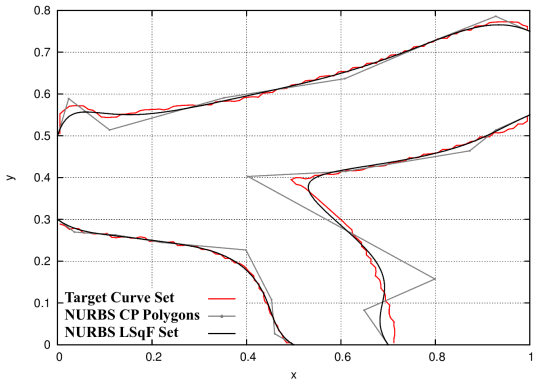


Figure 3: The TToST NURBS and their CP polygons plotted against the target TopO Γ_{FSI} solution after the LSqF process.

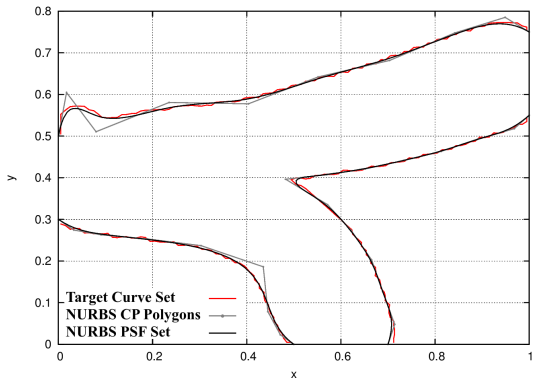


Figure 4: The TToST NURBS and their CP polygons plotted against the target TopO Γ_{FSI} solution after the PSF Algorithm.

be discussed in brief. The Field Matching (FM) objective function to be minimized and its sensitivities w.r.t. the design variables b_i are

$$F_{FM} = \frac{1}{2}\int_{\Omega} HH(\phi_N)(\phi_N - \phi_{Tar})^2 d\Omega \tag{17}$$

$$\frac{\delta F_{FM}}{\delta b_i} = \int_{\Omega} \left[\underbrace{\frac{1}{2}\frac{\partial HH(\phi_N)}{\partial \phi_N}(\phi_N - \phi_{Tar})^2}_{\Delta \mathbb{V}_{NB}\,Term} + \underbrace{HH(\phi_N)(\phi_N - \phi_{Tar})}_{\in \mathbb{V}_{NB}\,Term} \right] \frac{\delta \phi_N}{\delta b_i} d\Omega \tag{18}$$

where $HH(\phi_N)$ is a double Heavyside function which allows only the NB to be considered for integration. The $\Delta\mathbb{V}_{NB}$ term in eq. 18 describes how altering the shape of a NURBS curve through CP displacement affects a change in the NB volume, while the $\in\mathbb{V}_{NB}$ term describes the change of $\Delta\phi_{NB}$ within the current NB due the same CP displacement.

Each design variable b_i is a degree of freedom of a CP. Recast, $b_i = B_q^n$, where q is the index of the CP and n is the component of the CP which b_i pertains to. Note that both the number of degrees of freedom and the number of all components considered here is the number of dimensions of the system, D ($D=2$ for this paper). The NURBS curve has a total set of points P that can be broken into subsets P_i which pertain to the points affected by design variable b_i. Within each subset, points are denoted as p_j^k, where j is the index of the point within P_i and k is the point's component. Furthermore, each point p_j has a set $M_{i,j}$ of cells c_m within the NB which it 'owns' (i.e., which it defines the ϕ value for due to the SDCorr process), each with components denoted k due to the SDCorr relation to its curve point. Here, m is the the index of cell c within the entire NB; an index which is interchangeable with that of its ϕ value. If b_i moves, the set P_i of these NURBS points which are affected by it will also move, requiring a sensitivity relationship between b_i and $p_j \in P_i$ and between $p_j \in P_i$ and $c_m \in M_{i,j}$. The term within eq. 18 that requires this b_i to c_m coupling is $\frac{\partial \phi_N}{\partial b_i}$. If N is the set of basis functions of the NURBS defined by the parametric variable u, $\frac{\partial \phi_N}{\partial b_i}$ can be recast and found, after some derivation, to be

$$\frac{\delta \phi_N}{\delta b_i} \Longrightarrow \frac{\delta \phi_m}{\delta B_q^n} = \sum_{j=0}^{P_1-1}\sum_{k=0}^{D-1} \frac{\delta \phi_m}{\delta p_j^k}\frac{\delta p_j^k}{\delta B_q^n} = \sum_{j=0}^{P_1-1} \frac{p_l^n - c_{m(l)}^n}{\phi_m}\delta_j^l N_q(u_j) \tag{19}$$

where l is an iterate over the set P_i, subscript $m(l)$ denotes that a cell of index m is owned by the point with index l and $N_q(u_j)$ is the NURBS' derivative contribution at point u_j of the curve. The calculation of the FM CP sensitivities therefore requires a summation over all cells c_m in the NB, adding each's sensitivity contributions to all relevant CPs according to eq. 19.

The FMF algorithm, initialized with the NURBS curves found with the PSF (see Section 3.4), can now be summarized as follows for each NURBS curve to be fitted:

1. Generate points on the NURBS curve and perform the SDCorr for a defined NBHW.
2. Find F_{FM} via summation between the ϕ_N and ϕ_{Tar} fields within the NB.
3. Calculate the CP sensitivities via eq. (18) and move the CPs accordingly.
4. $IF\ |F^n_{FM} - F^{n-1}_{FM}| < \Delta F_{FM_{MIN}}, BREAK.$
5. Return to Step 1.

The FMF algorithm was run for the example case with parameters of $\Delta F_{FM_{MIN}} = 0.0001$ (a relatively high matching tolerance) and NBHW $= 0.1m$. A plot of the resulting NURBS curves, their CPs and the isoPts after the FMF process can be seen in fig. 5.

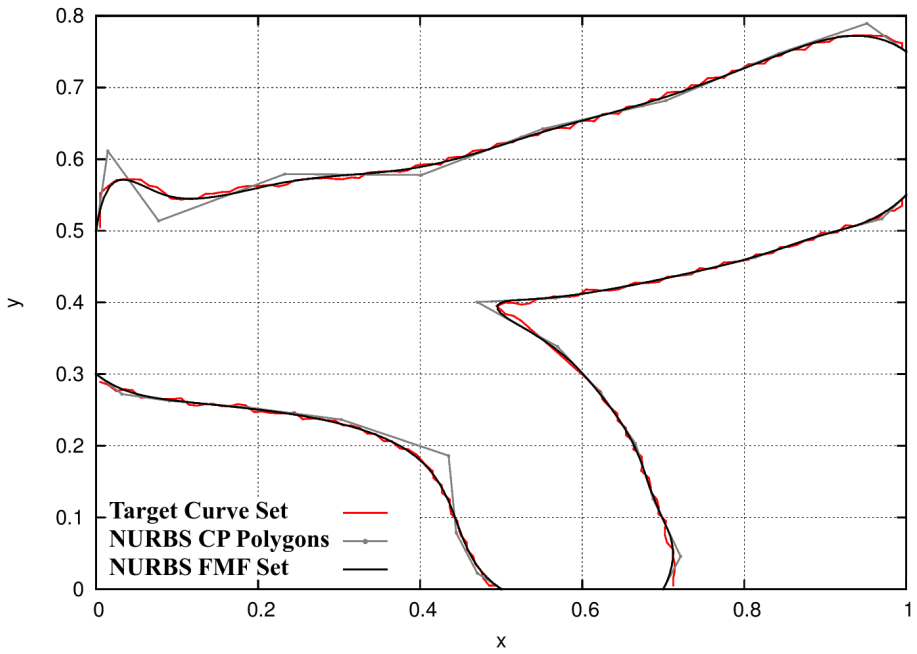


Figure 5: The TToST NURBS and their CP polygons plotted against the target TopO Γ_{FSI} solution after the FMF Algorithm.

4 SHAPE OPTIMIZATION

The TToST process (see Section 3) generates a set of NURBS curves which represent the Γ_{FSI} of the TopO solution and which can be treated as a solid wall boundary (S_W) for the primal and adjoint ShpO problem once a mesh has been generated from them. The flow equations for the ShpO are identical to eqs. 1a and 1b if the blockage term is excluded. Similarly, the

boundary conditions of the ShpO are identical to those of TopO. The design variables b_l for ShpO are, much like for the FMF optimization problem in Section 3.5, the degrees of freedom of the CPs which define S_W. The sensitivities for the ShpO CPs are computed by integrating over S_W via

$$\frac{\delta F}{\delta b_l} = -\int_{S_W} \left[\nu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) n_j - q n_i \right] \frac{\partial v_i}{\partial x_k} \frac{\delta x_k}{\delta b_l} dS \tag{20}$$

where $\frac{\delta x_k}{\delta b_l}$ is computed analytically through NURBS. It is beyond the scope of this paper to present ShpO in detail; the interested reader will find material pertinent to the ShpO method and its corresponding continuous adjoint formulation in [9]. The found ShpO solution compared to the parameterized TToST NURBS solution can be seen in fig. 6. Aside from deeming the initial bulging of isoSeg1 to be insignificant on the final solution, ShpO shows preference for widening the lower and narrowing the upper exit ducts, decreasing the objective function by over 3.0% through doing so.

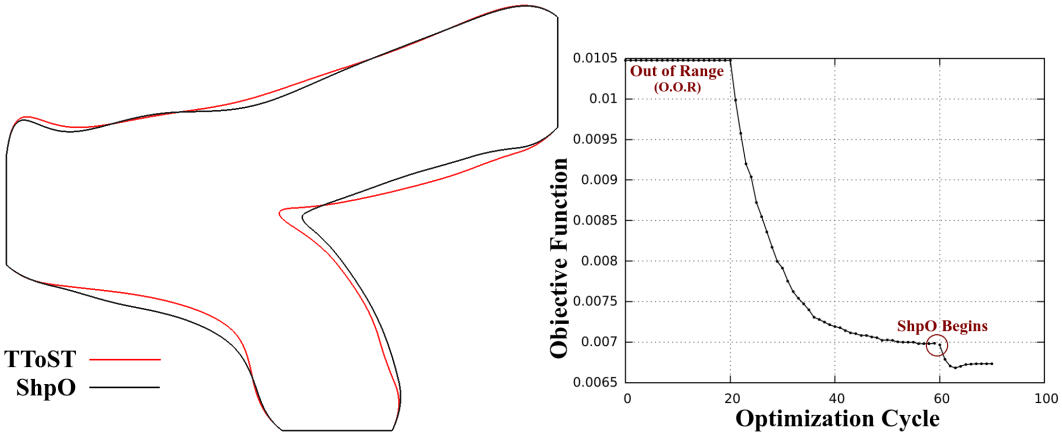


Figure 6: Example case's initial and final ShpO boundary solution (**left**) and its total pressure losses objective function's convergence history of and transitional behavior between TopO and ShpO (**right**).

The objective function trend comparison between TopO and ShpO is also shown in fig. 6. The TopO case was run to convergence to show that the continuous adjoint TopO and ShpO find similar solutions. In many applications, and as is done in Section 5, TopO would be used simply to initialize a flow channel, then let ShpO converge the solution as ShpO is considered to be faster than TopO since the zero-ϕ isoline moves only within the NB.

5 2D APPLICATION: Dual Bottle Neck Case

To demonstrate the robustness of the TToST method developed in this work, a more complex 'dualBottleNeck' case is presented. The case has a single inlet in the upper left of the case domain, two outlets in close proximity at the bottom right and two 'bottle neck' locations between them which fix the area of flow passage. The dualBottleNeck has a uniform Cartesian grid of 15200 cells and the same flow conditions, NBHW and objective function as the example case presented in Section 3. The volume constraint is similarly set to enforce $V_{Fluid,Tar} = 0.3$. The

case geometry, ϕ field solution found by the TopO process, TToST fitting solution and the final ShpO solution are shown in fig. 7.

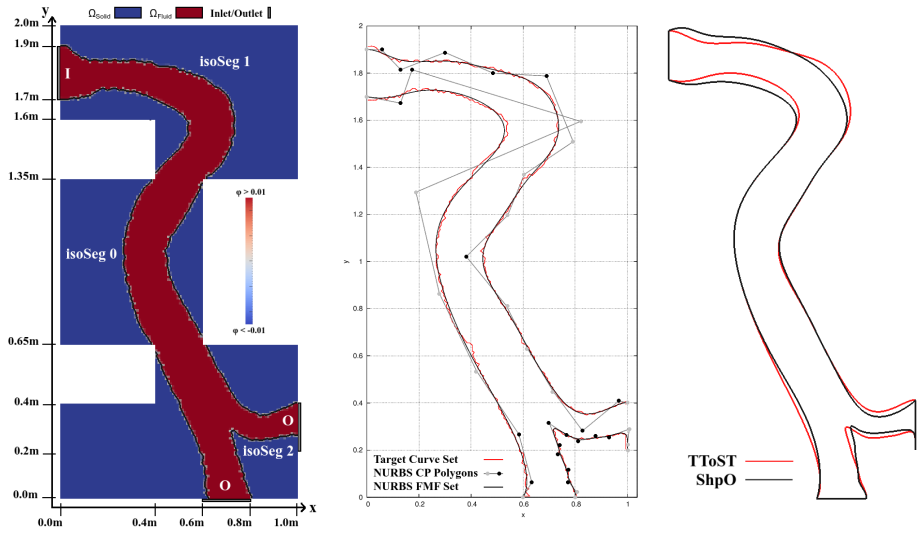


Figure 7: dualBottleNeck case solution set: (**left**): Case geometry and TopO ϕ solution with its known Γ_{FSI}. (**center**): TToST solution. Black and light grey dots indicate which control points were allowed to move and kept fixed, respectively. (**right**): The ShpO boundary solution overlaid that of the TToST.

Unlike the example case in Section 3, the dualBottleNeck case was not run to convergence during the TopO process but prematurely stopped at iteration 25 for the reasons discussed at the end of Section 4. Thus, the γ and ω ALM parameters were increased in strength to ensure the constraint be respected by the time of the TopO process' termination. Given initial degree of 4 and 10 CPs, the PSF process chose to define degrees 4, 4 and 3 and nCPs 10, 16 and 14 for NURBS curves 0,1 and 2, respectively. The FMF solution found using these parameters is shown in fig. 7, center. The two bottle necks in the case present an obstacle to the ShpO process, as its S_W boundaries are desired to remain inside the original boundary of the TopO case even though ShpO does not know where this boundary was. Thus, CPs which could affect boundary movement outside of the original TopO domain were fixed and are colored grey in the FMF solution. The ShpO process enlarges the entrance channel and carves out its lower wall in order to decrease pressure losses. It also expands the main lower exit channel. This enlarging of the lower channel was seen in the example case as well and occurs because a pressure loss is incurred when flow is diverted to the right (or upper) exit channel. Inversely, and due to the need to maintain the volume constraint, the ShpO process narrows the right exit channel.

The TopO and ShpO objective functions are plotted in fig. 8. The initial oscillation of the TopO objective is due to the severity of the imposed volume constraint since the dominance of the adjoint and constraint sensitivities can switch between iterations. A sizable objective function decrease is also seen between the end of TopO and the beginning of the ShpO: the TopO solution has not yet converged and the sensitivities are still quite large, resulting in a jagged Γ_{FSI} (see fig. 7, center) which generates pressure losses along its entire length. In contrast, all boundaries of ShpO are smooth. The ShpO process decreases the objective function by over 50%, which is not surprising due to the poor quality of the TopO's premature solution.

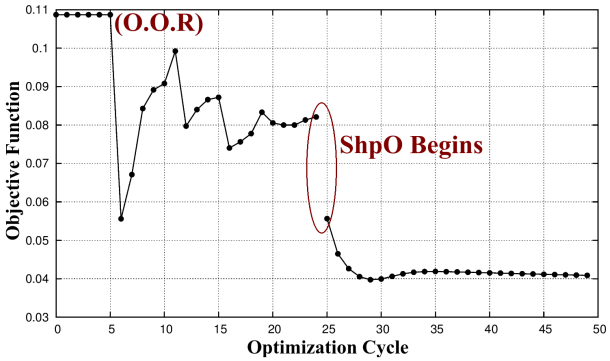


Figure 8: The convergence history of and transitional behavior between TopO and ShpO for the dualBottleNeck case's total pressure losses objective function.

6 CONCLUSIONS

The novel transitional TToST process between level-set topology and shape optimization was developed in this paper. The narrow band implementation of the continuous adjoint level set topology optimization process, TopO, ensures that the interface between the solid and fluid topological domains is explicitly defined and prevents topological island formation. Although the TopO method works well for 3D cases, its application was limited to 2D in order to showcase the TToST process, as it is currently only applicable to 2D geometries. A topological constraint for fluid volume was developed and can be used to ensure a topological solution which meets manufacturing specifications.

The 2D TToST process itself was presented, with its fitting algorithms and sensitivity calculation being discussed in detail. The cases shown in this paper demonstrate that the TToST process can accurately fit the interface defined by the TopO solution, thereby finding a parameterized NURBS set which can be used to initialize a continuous adjoint shape, ShpO, optimization problem. TToST has been shown to capture optimized solutions for both converged and pre-converged TopO problems, allowing it to decrease the overall computational cost of the linked TopO-ShpO optimization problem since ShpO's boundary movement is not limited by a narrow band like TopO. For all presented cases, ShpO allowed for a significant decrease in the objective function w.r.t. the solution found by TopO, with the improvement being more significant the sooner ShpO was allowed to take over.

Future research regarding the TToST algorithm includes expansion to 3D and the inclusion of user-defined topological islands: the current TToST algorithm would not consider floating solidified (or liquidized) islands as viable interfaces to be fitted. Validity of such an inclusion into the TToST process is under investigation.

REFERENCES

[1] M. Bendsoe and N. Kikuchi, Generating Optimal Topologies in Structural Design using a Homogenization Method. *Journal of Computer Methods in Applied Mechanics and Engineering*, **71**, 197-224, 1988.

[2] T. Borvall and J. Peterson, Topology Optimization of Fluids in Stokes Flow. *International Journal for Numerical Methods in Fluids*, **41**, 77-107, 2003.

[3] J.K. Guest and J.H. Prevost, Topology Optimization of Creeping Fluid Flows using a Darcy-Stokes Finite Element. *International Journal for Numerical Methods in Engineering*, **66**, 461-484, 2006.

[4] A. Gersborg-Hansen, O. Sigmund and R. Haber, Topology Optimization of Channel Flow Problems. *Structural and Multidisciplinary Optimization*, **30**, 181-192, 2005.

[5] O. Pironneau, *Optimal Shape Design for Elliptic Systems*. Springer-Verlag, 1984.

[6] A. Jameson, Aerodynamic Design via Control Theory. *Journal of Scientific Computing*, **3**, 233-260, 1988.

[7] W.K. Anderson and V. Venkatakrishnan, Aerodynamic Design Optimization on Unstructured Grids with a Continuous Adjoint Formulation. *Computers & Fluids*, **28**, 443-480, 1999.

[8] D.I. Papadimitriou and K.C. Giannakoglou, Aerodynamic Shape Optimization Using First and Second Order Adjoint and Direct Approaches. *Archives of Computational Methods in Engineering*, **15**, 447-488, 2008.

[9] E.M. Papoutsis-Kiachagias and K.C. Giannakoglou, Continuous Adjoint Methods for Turbulent Flows, applied to Shape and Topology Optimization: Industrial Applications. *Arch. Comput. Methods Eng.*, 1-45, 2014.

[10] S. Osher and J.A. Sethian, Fronts Propagating with Curvature Dependent Speed: Algorithms based on Hamilton-Jacobi Formulations. *Journal of Computational Physics*, **79**, 12-49, 1988.

[11] `http://www.openfoam.com`

[12] J. Nocedal and S.J. Wright, *Numerical Optimization*. Springer, 1999.

[13] E.A. Kontoleontos, E.M. Papoutsis-Kiachagias, A.S. Zymaris, D.I. Papadimitriou and K.C. Giannakoglou, Adjoint–Based Constrained Topology Optimization for Viscous Flows, Including Heat Transfer. *Engineering Optimization*, **45**, 941-961, 2013.

[14] G.K. Karpouzas and E.D. Villiers, Level-Set Based Topology Optimization using the Continuous Adjoint Method. M. Papadrakakis, M.G. Karlaftis, N.D. Lagaros (eds). *An International Conference on Engineering and Applied Sciences Optimization*, Kos Island, Greece, 4-6, June 2014.

[15] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 2003.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

AERODYNAMIC SHAPE OPTIMIZATION USING THE TRUNCATED NEWTON METHOD AND CONTINUOUS ADJOINT

Mehdi Ghavami Nejad[1], **Evangelos M. Papoutsis-Kiachagias**[1], **and Kyriakos C. Giannakoglou**[1]

[1] National Technical University of Athens (NTUA), School of Mechanical Engineering, Parallel CFD & Optimization Unit, Greece
e-mail: mehdi@mail.ntua.gr, vaggelisp@gmail.com, kgianna@central.ntua.gr

Keywords: Truncated Newton, Continuous Adjoint, Aerodynamic Optimization, Losses Minimization

Abstract. *This paper presents the development and application of the Truncated Newton (TN) method for shape optimization problems based on continuous adjoint. The method is presented for laminar, incompressible flows. OpenFOAM® is chosen as the CFD toolbox in which the method is developed. The Newton equations are solved using the restarted linear GMRES algorithm which requires only the product of the Hessian matrix of the objective function (with respect to the design variables) with a vector. This overcomes the cost for computing the Hessian matrix itself, which unfortunately scales with the number of design variables. The computation of Hessian-vector products is conducted via the combination of continuous adjoint and direct differentiation that gives the minimum cost. The developed method is used for the shape optimization of two 3D ducts and the speed-up gained compared to rival methods is showcased.*

This research was funded from the People Programme (ITN Marie Curie Actions) of the European Union's 7th Framework Programme (FP7/2007-2013) under REA Grant Agreement 317006 *(AboutFLOW). The first author is an AboutFLOW Early Stage Researcher.*

1 INTRODUCTION TO THE TRUNCATED NEWTON METHOD

An unconstrained optimization problem, in which the target is to minimize the objective function F by controlling the design variables b_i , $i = 1,...,N$ can be solved by means of the Newton method, according to which the design variables are updated ($b_i^{n+1} = b_i^n + \delta b_i$) after solving the Newton equations

$$\frac{\delta^2 F}{\delta b_i \delta b_j}^n \delta b_j = -\frac{\delta F}{\delta b_i}^n \tag{1}$$

where n is the Newton iteration counter, to be omitted hereafter. The direct solution of eq. 1 requires the computation of the Hessian of F, with a computational cost that scales with N [4].

Considering eq. 1 as a linear system of equations of the form $Ax = q$, a possible way to solve it is through an iterative solver which requires only the computation of matrix-vector products. Since the Hessian matrix is symmetric, a popular choice is the Conjugate Gradient (CG) method, [5, 1]. For reasons to be discussed in sections 8 and 9.1, the linear restarted GMRES method, [9], schematically given in Algorithm 1, is used herein instead.

Algorithm 1 : The Linear Restarted GMRES Method for the Solution of $Ax = q$

$r^0 = Ax^0 - q$, $s^1 = \frac{r^0}{\|r^0\|_2}$
for $j = 1,2,\ldots,M$ **do**
  $w^j = As^j$
  for $i = 1,2,\ldots,j$ **do**
    $h_{i,j} = (w^j, s^i)$
  end for
  $s^{j+1} = w^j - \sum_{i=1}^{j} h_{i,j} s^i$
  $h_{j+1,j} = \|s^{j+1}\|_2$
  $s^{j+1} = \frac{s^{j+1}}{h_{j+1,j}}$
end for
Compute $\beta_1,\ldots,\beta_M$ by solving the minimization problem $min\|Ax^M - q\|_2$
$x^M = x^0 + \sum_{i=1}^{M} \beta_i s^i$

Based on Algorithm 1, the cost of each GMRES iteration is dominated by the cost of computing the matrix–vector product (As), M times during the Arnoldi process, where M is the chosen number of basis vectors. Regarding eq. 1, since the Hessian matrix stands for A, the use of the Truncated Newton (TN) method in aerodynamic shape optimization problems means that the Hessian matrix itself is no more needed and only its product with a vector must be computed. On the other hand, for the r.h.s. of eq. 1, the gradient of F must be available and the (continuous) adjoint method, [6], is the less expensive way to compute it, at a CPU cost which is practically independent of N.

2 FLOW MODEL & OBJECTIVE FUNCTION

3D laminar flows of incompressible fluids are governed by the continuity and momentum equations,

$$R^p = -\frac{\partial v_j}{\partial x_j} = 0 \tag{2}$$

$$R_i^v = v_j \frac{\partial v_i}{\partial x_j} - \frac{\partial \tau_{ij}}{\partial x_j} + \frac{\partial p}{\partial x_i} = 0\,, \quad i = 1,2,3 \tag{3}$$

where v_i are the velocity components, p the static pressure divided by the constant density, $\tau_{ij} = \nu\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)$ the stress tensor and ν the constant viscosity.

In this paper, the development of the TN method will be demonstrated for the objective function

$$F = -\int_{S_{I,O}} \left(p + \frac{1}{2} v_k^2\right) v_i n_i dS \tag{4}$$

where $S = S_I \cup S_O \cup S_W$ is the domain boundary with S_I being the inlet, S_O the outlet, S_W the solid wall and $\mathbf{n}$ the outward unit normal vector to the surface. F stands for the volume–averaged total pressure losses of the flow inside a duct; the optimal duct shape is the one yielding the minimal value that F may take on, given the parameterization of S_W.

3 COMPUTATION OF $\frac{\delta F}{\delta b_i}$ VIA CONTINUOUS ADJOINT

It is beyond the scope of this paper to present the continuous adjoint method for the computation of $\delta F/\delta b_i$; the interested reader should refer to [6]. The adjoint continuity and adjoint momentum equations are

$$R^q = -\frac{\partial u_j}{\partial x_j} = 0 \tag{5}$$

$$R_i^u = u_j \frac{\partial v_j}{\partial x_i} - \frac{\partial \left(u_i v_j\right)}{\partial x_j} - \frac{\partial \tau_{ij}^a}{\partial x_j} + \frac{\partial q}{\partial x_i} = 0\,,\ i = 1,2,3 \tag{6}$$

where u_i are the adjoint velocity components, q the adjoint pressure and $\tau_{ij}^a = \nu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$ the adjoint stress tensor. By satisfying eqs. 5 and 6, $\delta F/\delta b_n$ becomes independent of $\delta v_i/\delta b_n$ and $\delta p/\delta b_n$ at the interior of the computational domain, [6].

Using the continuous adjoint method, the gradient of the F w.r.t. b_n becomes

$$\frac{\delta F}{\delta b_n} = \int_\Omega A_{jk} \frac{\partial}{\partial x_j}\left(\frac{\delta x_k}{\delta b_n}\right) d\Omega \tag{7}$$

where

$$A_{jk} = -u_i v_j \frac{\partial v_i}{\partial x_k} - u_j \frac{\partial p}{\partial x_k} - \tau_{ij}^a \frac{\partial v_i}{\partial x_k} + u_i \frac{\partial \tau_{ij}}{\partial x_k} + q \frac{\partial v_j}{\partial x_k} \tag{8}$$

A few comments on eq. 7 are due. According to [2], $\delta F/\delta b_i$ can either be expressed exclusively in terms of surface integrals or may also include field integrals. The two formulations are referred to as *SI* (*Surface Integral*) and *FI* (*Field Integral*), respectively. Eq. 7 is obviously based

on the FI formulation. A noticeable advantage of the latter is that it avoids the computation of second-order spatial derivatives along S which might become a source of error. The proposed TN method relies upon the FI formulation since, following this approach, the computation of even higher spatial gradients at the boundary is avoided during the evaluation of $\frac{\delta^2 F}{\delta b_n \delta b_m} s_m$.

4 BACKGROUND EXPRESSIONS

First of all, a clear distinction between total and partial derivatives should be made. For any flow quantity Φ, the total derivative $\delta\Phi/\delta b_n$, which represents the total change in Φ caused by variations in b_n, is

$$\frac{\delta\Phi}{\delta b_n} = \frac{\partial\Phi}{\partial b_n} + \frac{\partial\Phi}{\partial x_k}\frac{\delta x_k}{\delta b_n} \tag{9}$$

where the partial derivative $\partial\Phi/\partial b_n$ includes only the variation in Φ caused due to changes in the design variables, without considering space deformations.

The TN method makes extensive use of the products of total derivatives and any vector s_m. So, it is convenient to define

$$\overline{\Phi} = \frac{\delta\Phi}{\delta b_m} s_m \tag{10}$$

Eq. 10 is also valid for the grid coordinates, so $\overline{x_k} = \frac{\delta x_k}{\delta b_m} s_m$. Starting from

$$\frac{\delta}{\delta b_m}\left(\frac{\partial\Phi}{\partial x_j}\right) s_m = \frac{\partial}{\partial b_m}\left(\frac{\partial\Phi}{\partial x_j}\right) s_m + \frac{\partial}{\partial x_k}\left(\frac{\partial\Phi}{\partial x_j}\right)\frac{\delta x_k}{\delta b_m} s_m = \frac{\partial}{\partial x_j}\left(\frac{\partial\Phi}{\partial b_m}\right) s_m + \frac{\partial}{\partial x_k}\left(\frac{\partial\Phi}{\partial x_j}\right)\overline{x_k}$$

it can easily be proved that

$$\overline{\frac{\partial\Phi}{\partial x_j}} = \frac{\delta}{\delta b_m}\left(\frac{\partial\Phi}{\partial x_j}\right) s_m = \frac{\partial\overline{\Phi}}{\partial x_j} - \frac{\partial\Phi}{\partial x_k}\frac{\partial\overline{x_k}}{\partial x_j} \tag{11}$$

It can also be proved that, if Φ, Ψ is any pair of quantities, the following equation is also valid

$$\frac{\delta}{\delta b_m}\left(\Psi\frac{\partial\Phi}{\partial x_j}\right) s_m = \overline{\Psi}\frac{\partial\Phi}{\partial x_j} + \Psi\frac{\partial\overline{\Phi}}{\partial x_j} - \Psi\frac{\partial\Phi}{\partial x_k}\frac{\partial\overline{x_k}}{\partial x_j} \tag{12}$$

Also, as shown in [4], for either structured or unstructured grids,

$$\frac{\delta(d\Omega)}{\delta b_m} = \frac{\partial}{\partial x_\lambda}\left(\frac{\delta x_\lambda}{\delta b_m}\right) d\Omega \tag{13}$$

from which we get

$$\frac{\delta(d\Omega)}{\delta b_m} s_m = \frac{\partial}{\partial x_\lambda}\left(\frac{\delta x_\lambda}{\delta b_m} s_m\right) d\Omega = \frac{\partial\overline{x_\lambda}}{\partial x_\lambda} d\Omega \tag{14}$$

In what follows, the following abbreviation

$$\overline{\overline{x_{k,n}}} = \frac{\delta^2 x_k}{\delta b_n \delta b_m} s_m \tag{15}$$

will also be used.

5 COMPUTATION OF HESSIAN(F)–VECTOR PRODUCTS

The TN method requires the computation of $\frac{\delta^2 F}{\delta b_n \delta b_m} s_m$. Based on the background expressions presented in section 4, it is a matter of a rather lengthy development to show that

$$\begin{aligned}\frac{\delta^2 F}{\delta b_n \delta b_m} s_m &= \int_\Omega \overline{A_{jk}} \frac{\partial}{\partial x_j}\left(\frac{\delta x_k}{\delta b_n}\right) d\Omega + \int_\Omega A_{jk} \frac{\delta}{\delta b_m}\left[\frac{\partial}{\partial x_j}\left(\frac{\delta x_k}{\delta b_n}\right)\right] s_m d\Omega \\ &+ \int_\Omega A_{jk} \frac{\partial}{\partial x_j}\left(\frac{\delta x_k}{\delta b_n}\right) s_m \frac{\delta(d\Omega)}{\delta b_m} \end{aligned} \tag{16}$$

where

$$\begin{aligned}\overline{A_{jk}} &= -\overline{u_i} v_j \frac{\partial v_i}{\partial x_k} - u_i \overline{v_j} \frac{\partial v_i}{\partial x_k} - u_i v_j \overline{\frac{\partial v_i}{\partial x_k}} + u_i v_j \frac{\partial v_i}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_k} - \overline{u_j} \frac{\partial p}{\partial x_k} - u_j \frac{\partial \overline{p}}{\partial x_k} \\ &+ u_j \frac{\partial p}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_k} - \nu \left(\frac{\partial \overline{u_i}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i}\right) \frac{\partial v_i}{\partial x_k} + \nu \left(\frac{\partial u_i}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_j} + \frac{\partial u_j}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_i}\right) \frac{\partial v_i}{\partial x_k} \\ &- \nu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \frac{\partial \overline{v_i}}{\partial x_k} + \nu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \frac{\partial v_i}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_k} \\ &+ \overline{u_i} \frac{\partial}{\partial x_k}\left[\nu \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] + u_i \frac{\partial}{\partial x_k}\left[\nu \left(\frac{\partial \overline{v_i}}{\partial x_j} + \frac{\partial \overline{v_j}}{\partial x_i}\right)\right] \\ &- u_i \frac{\partial}{\partial x_k}\left[\nu \left(\frac{\partial v_i}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_j} + \frac{\partial v_j}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_i}\right)\right] - u_i \frac{\partial}{\partial x_\lambda}\left[\nu \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] \frac{\partial \overline{x_\lambda}}{\partial x_k} \\ &+ \overline{q} \frac{\partial v_j}{\partial x_k} + q \frac{\partial \overline{v_j}}{\partial x_k} - q \frac{\partial v_j}{\partial x_\lambda} \frac{\partial \overline{x_\lambda}}{\partial x_k} \end{aligned} \tag{17}$$

Based on eq. 15, the second integral on the r.h.s. of eq. 16 becomes

$$\int_\Omega A_{jk} \frac{\delta}{\delta b_m}\left[\frac{\partial}{\partial x_j}\left(\frac{\delta x_k}{\delta b_n}\right)\right] s_m d\Omega = \int_\Omega A_{jk} \frac{\partial \overline{x_{k,n}}}{\partial x_j} d\Omega - \int_\Omega A_{jk} \frac{\partial}{\partial x_\lambda}\left(\frac{\delta x_k}{\delta b_n}\right) \frac{\partial \overline{x_\lambda}}{\partial x_j} d\Omega \tag{18}$$

Computing $\overline{v_i}$ and $\overline{p}$ is straightforward since these are equal to the product of the directly differentiated flow variables and s_m. So, $\overline{v_i}$ and $\overline{p}$ result from

$$\overline{R^p} = \frac{\partial \overline{v_j}}{\partial x_j} - \frac{\partial v_j}{\partial x_k} \frac{\partial \overline{x_k}}{\partial x_j} = 0 \tag{19}$$

and

$$\begin{aligned}\overline{R_i^v} &= \frac{\partial(\overline{v_i} v_j)}{\partial x_j} + \frac{\partial(v_i \overline{v_j})}{\partial x_j} - \frac{\partial}{\partial x_j}\left[\nu \left(\frac{\partial \overline{v_i}}{\partial x_j} + \frac{\partial \overline{v_j}}{\partial x_i}\right)\right] + \frac{\partial \overline{p}}{\partial x_i} \\ &- \frac{\partial(v_i v_j)}{\partial x_k} \frac{\partial \overline{x_k}}{\partial x_j} + \frac{\partial}{\partial x_j}\left[\nu \left(\frac{\partial v_i}{\partial x_k} \frac{\partial \overline{x_k}}{\partial x_j} + \frac{\partial v_j}{\partial x_k} \frac{\partial \overline{x_k}}{\partial x_i}\right)\right] \\ &+ \frac{\partial}{\partial x_k}\left[\nu \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] \frac{\partial \overline{x_k}}{\partial x_j} - \frac{\partial p}{\partial x_k} \frac{\partial \overline{x_k}}{\partial x_i} = 0 \end{aligned} \tag{20}$$

Also, the product of the DD of the adjoint equations and s_m yields

$$\overline{R^q} = \frac{\partial \overline{u_j}}{\partial x_j} - \frac{\partial u_j}{\partial x_k} \frac{\partial \overline{x_k}}{\partial x_j} = 0 \tag{21}$$

and

$$\begin{aligned}\overline{R_i^u} &= \overline{u_j}\frac{\partial v_j}{\partial x_i}+u_j\frac{\partial \overline{v_j}}{\partial x_i}-\frac{\partial(\overline{u_i}v_j)}{\partial x_j}-\frac{\partial(u_i\overline{v_j})}{\partial x_j}\\ &- \frac{\partial}{\partial x_j}\left[\nu\left(\frac{\partial \overline{u_i}}{\partial x_j}+\frac{\partial \overline{u_j}}{\partial x_i}\right)\right]+\frac{\partial \overline{q}}{\partial x_i}-u_j\frac{\partial v_j}{\partial x_k}\frac{\partial \overline{x_k}}{\partial x_i}\\ &+ \frac{\partial(v_ju_i)}{\partial x_k}\frac{\partial \overline{x_k}}{\partial x_j}+\frac{\partial}{\partial x_j}\left[\nu\left(\frac{\partial u_i}{\partial x_k}\frac{\partial \overline{x_k}}{\partial x_j}+\frac{\partial u_j}{\partial x_k}\frac{\partial \overline{x_k}}{\partial x_i}\right)\right]\\ &+ \frac{\partial}{\partial x_k}\left[\nu\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)\right]\frac{\partial \overline{x_k}}{\partial x_j}-\frac{\partial q}{\partial x_k}\frac{\partial \overline{x_k}}{\partial x_i}=0\end{aligned} \tag{22}$$

from which $\overline{q}$ and $\overline{u}_i$ can be computed.

6 COMPUTATION OF $\overline{x_k}$ AND $\overline{\overline{x_{k,n}}}$

In order to compute the Hessian-vector product of eq. 16, the grid sensitivities $\delta x_k/\delta b_n$ as well as their first-($\overline{x_k}$) and second-order ($\overline{\overline{x_{k,n}}}$) projections to $\mathbf{s}$ must be computed. These computations depend upon the method used to deform the computational grid after the update of the design variables. In [1], the Laplace equation was used as the grid displacement model and the corresponding PDEs for computing the aforementioned terms were presented. Here, a different grid displacement model is employed, based on volumetric B–Splines, details for which can be found in [7, 3]. In brief, the grid points coordinates x_l are given by

$$x_l(u,v,w) = U_{i,pu}(u)V_{j,pv}(v)W_{k,pw}(w)B_l^{ijk} \tag{23}$$

Here, $B_l^{ijk}, l\in[1,3], i\in[0,I], j\in[0,J], k\in[0,K]$ are the Cartesian coordinates of the ijk-th control point of a $3D$ structured control grid (acting also as the design variables of the optimization problem), I,J and K stand the number of control points per control grid direction, $\mathbf{u}=[u_1,u_2,u_3]^T=[u,v,w]^T$ are the CFD grid point parametric coordinates, U,V,W are the B–Splines basis functions and pu,pv,pw their respective degrees, which may be different per control grid direction. Details about B–Splines basis definitions and properties can be found in [8].

Obtaining grid sensitivities and their projections to $\mathbf{s}$ is just a matter of analytically differentiating eq. 23 w.r.t. the coordinates of the control grid points. Let $b_m = B_t^{\lambda\mu\xi}$. Then, the grid sensitivities are given by

$$\frac{\delta x_l(u,v,w)}{\delta b_m} = U_{\lambda,pu}(u)V_{\mu,pv}(v)W_{\xi,pw}(w)\delta_l^t \tag{24}$$

where δ_l^t is the Kronecker symbol. Eq. 24 states that grid sensitivities for each CFD grid point with parametric coordinates $\mathbf{u}$ are given by the product of the basis functions, evaluated at $\mathbf{u}$, corresponding to the $\lambda\mu\xi$ control point. After computing the N components of $\delta x_l/\delta b_m$, computing $\overline{x_l}$ is a matter of a simple summation. It should be noted that $\overline{\overline{x_{l,n}}}=0$, since the grid displacement model depends linearly on the design variables. This further simplifies eq. 18, by eliminating the first term on its r.h.s.

7 THE TN ALGORITHM – COMMENTS ON THE CPU COST

Using eqs. 14 to 18, eq. 16 can be written as

$$\frac{\delta^2 F}{\delta b_n \delta b_m}s_m = \int_\Omega\left[\overline{A_{jk}}+A_{jk}\frac{\partial \overline{x_\lambda}}{\partial x_\lambda}-A_{\lambda k}\frac{\partial \overline{x_j}}{\partial x_\lambda}\right]\frac{\partial}{\partial x_j}\left(\frac{\delta x_k}{\delta b_n}\right)d\Omega \tag{25}$$

where $\overline{A_{jk}}$ is given by eq. 17. To compute $\overline{A_{jk}}$, apart from the flow and adjoint fields, the "overbar" fields $(\overline{v_i}, \overline{u_i}, \overline{p}, \overline{q})$, as well as $\overline{x_i}$ and their spatial derivatives must be available.

So, in each Newton cycle, the numerical solution of $R^p = 0$ and $R^v_i = 0$ (eqs. 2 and 3) yields the flow fields (p, v_i). The solution of $R^q = 0$ and $R^u_i = 0$ (eqs. 5 and 6) yields the adjoint fields (q, u_i). So far, the computational cost is approximately equal to that of twice solving the flow equations or 2 EFS (EFS stands for an Equivalent Flow Solution, i.e. the cost for solving the flow equations).

Before solving for $\overline{p}$ and $\overline{v_i}$, $\overline{x_k}$ must be computed by evaluating eq. 24 for $m \in [1, N]$ and contracting with the components of the projection vector $\mathbf{s}$. The latter has a cost of N GDE (GDE stands for Grid Displacement Evaluations, i.e. the cost of evaluating $\delta x_k / \delta b_m$ for a single m), since $\delta x_k / \delta b_m$ has to be evaluated separately for each design variable. It should be mentioned that 1 GDE is significantly cheaper than 1 EFS since $\delta x_k / \delta b_m$ is computed analytically through eq. 24. $\overline{x_k}$ has to be evaluated once per GMRES iteration, contributing a total cost of MN GDE per optimization cycle.

Computing $\overline{p}$ and $\overline{v_i}$ requires the numerical solution of eqs. 19 and 20. Similarly, to compute $\overline{q}$ and $\overline{u_i}$ requires the numerical solution of eqs. 21 and 22. Both systems of equations should be solved within the GMRES loop (i.e. M times) and contribute $2M$ EFS to the overall cost of a Newton iteration or cycle.

Within each GMRES iteration, the computation of $\overline{A_{jk}}$ also requires the availability of the $\delta x_k / \delta b_n$. These fields, however, have already been computed for the evaluation of $\overline{x_k}$ and contribute no extra cost.

Based on the above, the overall CPU cost per Newton iteration is equal to $2+2M$ EFS plus NM GDE. However, since the cost of a GDE is significantly lower than that of an EFS, the GDE part can be considered negligible for a moderate number of design variables. This leads to a cost per Newton cycle that is, practically, independent of the number of design variables N.

8 CHOICE OF THE LINEAR SOLVER

The TN method can be coupled with any iterative linear solver that relies on the computation of matrix-vector products, without requiring the knowledge of the Hessian matrix itself. In previous publications, [5, 1], the CG method was used as the linear solver, since the Hessian matrix is symmetric in theory. However, the Hessian expression obtained through the use of the AV-DD approach (i.e., use the adjoint variable (AV) method for the computation of $\delta F / \delta b_n$ and DD for the computation of the variations of the primal and adjoint fields; the equivalent of tangent-on-reverse in the Automatic Differentiation terminology) is not symmetric (eq. 16, neglecting the multiplication with s_m) and produces a symmetric matrix only upon the convergence of all equations to machine accuracy (as discussed in Appendix A). This non-symmetry of the Hessian expression is essential for the application of TN methods, since it allows the computation of Hessian-vector products at a cost which is independent of N. In CFD-based optimization, it is quite common not to converge the primal and adjoint equations to machine accuracy in each optimization cycle in order to reduce the total CPU cost. This can deteriorate, to an extent, the symmetry of the Hessian matrix, rendering CG inappropriate for the solution of the Newton equations. To avoid this inconsistency, the linear GMRES solver can be used in the context of TN methods. The impact of the linear solver is investigated in section 9.1.

9 APPLICATIONS

The applications section consists of two parts. In the first, the impact of the linear solver used to iteratively solve the Newton equations is investigated. In the second, optimization problems concerning the total pressure losses minimization in two 3D duct geometries are tackled; the results obtained by using the TN approach are compared to those computed by using Steepest Descent (SD), the Fletcher-Rives Conjugate Gradient (CG) and BFGS methods for updating the design variables.

9.1 Impact of the linear solver

To investigate the impact of the linear solver, an optimization problem with only 5 design variables was devised, making the computation of the Hessian matrix feasible. The shape optimization of a 2*D* U-bend duct is considered, targeting minimum total pressure losses. The upper part of the U-bend is parameterized using Bézier–Bernstein polynomials and the y coordinates of 5 control points are used as the design variables, fig. 1. The flow is laminar with $Re = 667$ based on the inlet length. The update of the design variables is driven by a number of different methods, among which the TN method coupled with the CG and GMRES solvers. Their convergence histories are presented in fig. 2. It can be seen that as the number of linear solver iterations M increases, the GMRES-based TN method greatly outperforms the CG-based one. If fact, when M is chosen to be equal to the Hessian matrix dimension, the GMRES-based TN method has exactly the same convergence with the pure Hessian method, as expected. On the contrary, the CG-based TN method requires approximately 4 times more EFS in order to reach the optimal solution. This can be attributed to the fact that CG is used to iteratively solve a slightly non-symmetric system. In detail, the symmetric, in theory, elements of the Hessian matrix computed using the AV-DD approach have a maximum difference of 0.8%, a mean difference of 0.1% and a standard deviation of 0.2%. Based on the above, for the remainder of this article, the GMRES solver is used in conjunction with the TN method.

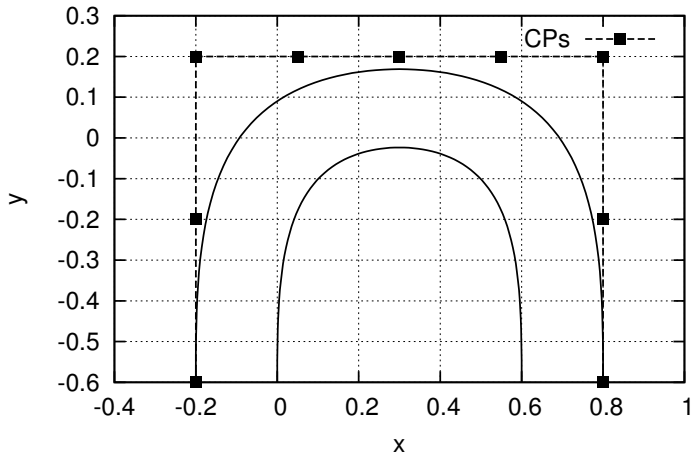
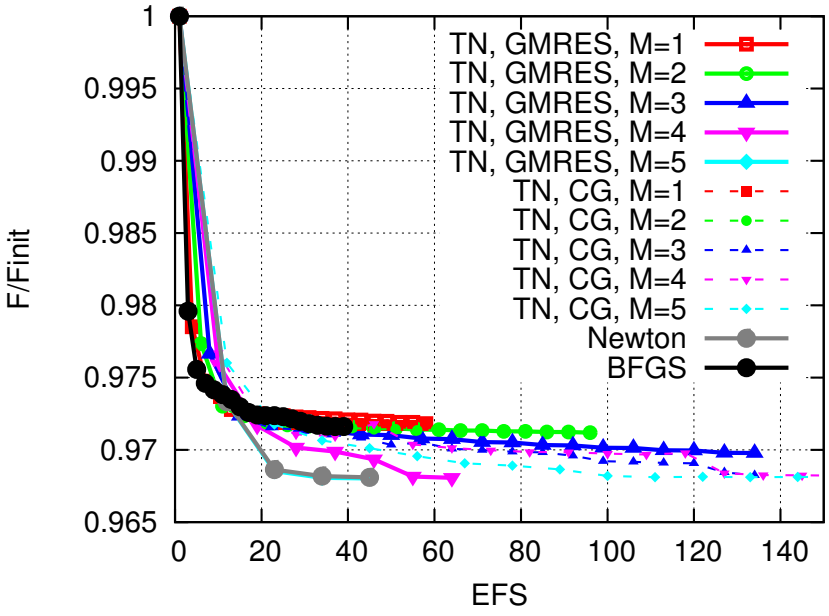


Figure 1: 2*D* U-bend duct optimization: dust shape and the Bézier–Bernstein control points parameterizing it. Only the y coordinates of the top 5 control points (CP) are allowed to vary during the optimization.



(a)

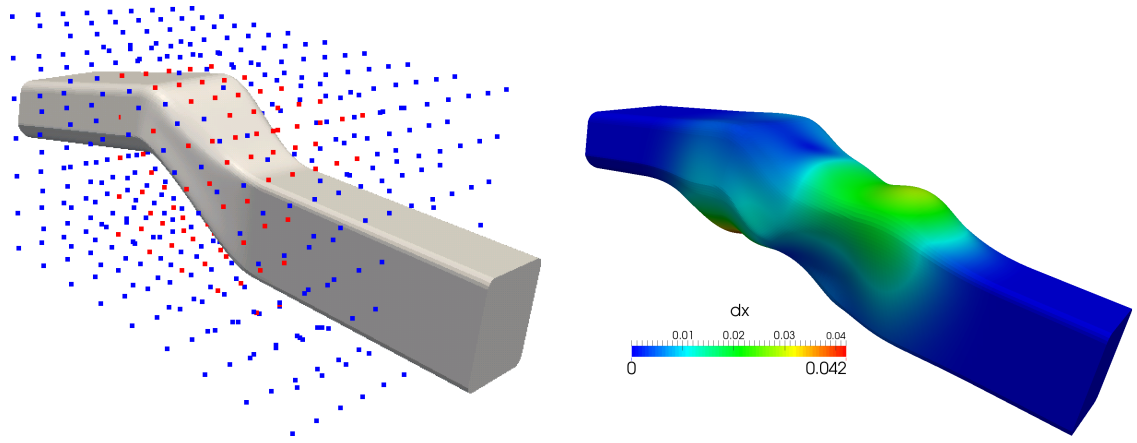
Figure 2: $2D$ U-bend duct optimization: Convergence of the BFGS, Newton and TN optimization algorithms. Convergence of the TN method is included with both CG and GMRES, with linear solver iterations in the range of $M \in [1,5]$. Results of GMRES-based TN are plotted with a continuous line while CG-based results with a dashed one. As expected, the convergence of the GMRES-based TN with $M=5$ coincides with the convergence of the Newton method (the two curves are hardly distinguishable) since the dimension of the problem is $N=5$.

9.2 3D shape optimization

In this section, two applications of the developed TN optimization algorithm are presented. The first one deals with the shape optimization of a $3D$ S-bend duct. The geometry and flow conditions are provided as one of the cases of the AboutFLOW ITN programme. The flow is laminar with a Reynolds number of $Re=400$ based on the inlet hydraulic diameter and the mesh is comprised of 474000 hexahedrals. A $9\times7\times9$ control grid is used to parameterize part of the duct which, after disregarding fixed control points, results to 375 design variables, fig. 3. In fig. 4, the convergence history of the developed TN algorithm is compared to those of the SD, CG and BFGS methods. Comparisons are presented twice, in terms of the cycles required to reach the minimum and the corresponding EFS. It can be observed that TN outperforms the other methods, since it computes the optimized duct shape using less optimization cycles and, especially, by requiring slightly less EFS. In fig. 5, the flow streamlines on the reference and optimized geometries are compared, indicating the significant reduction of the flow recirculation that leads to a total pressure losses reduction of $\sim 60\%$.

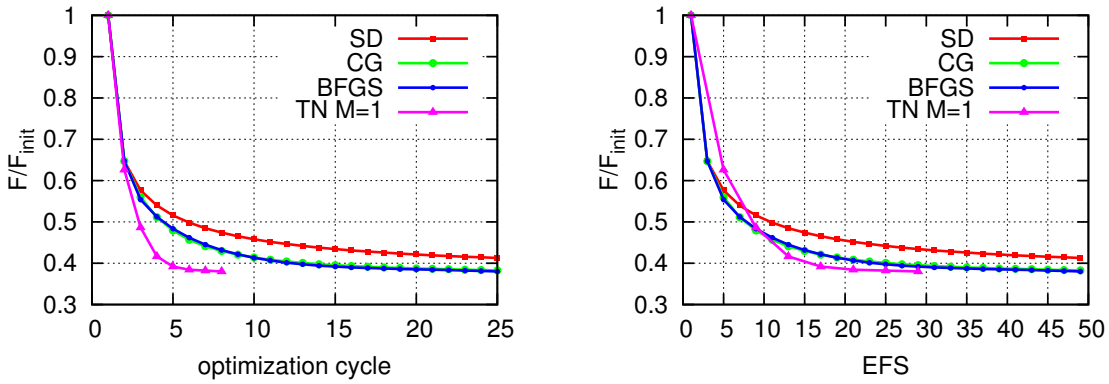
The second case is concerned with the optimization of a 3D U-bend duct. The flow Reynolds number is $Re=400$ and a mesh consisting of 7×10^5 hexahedrals is used. The reference geometry and the $4\times5\times2$ control grid parameterizing it are depicted in fig. 6. Since only two rows of control points are used in the z direction and the reference geometry was generated by stacking a $2D$ profile, the shape parameterization is practically $2D$. In fig. 7, the total pressure field along with velocity vectors are plotted for two slices, located at 10% and 50% of the duct height. It can be seen that the flow recirculation downstream of the U-shaped formation has been suppressed, leading to a reduction of about 18% in the objective function value. In fig. 8, the convergence histories of the TN, SD and CG methods are illustrated. For this case, TN and CG reach the

optimal solution at approximately the same CPU cost. The TN method outperforms CG during the initial phase of the optimization run, providing a better solution if the entire CPU cost of the optimization can not be afforded.



(a) (b)

Figure 3: S-bend duct optimization: (a) duct shape and the control grid parameterizing it. Control points in red are allowed to vary during the optimization while blue ones are kept fixed, (b) optimal shape coloured based on the cumulative displacement. Flow from right to left.



(a) (b)

Figure 4: S-bend duct optimization: comparison of the convergence of SD, CG, BFGS and TN w.r.t. (a) optimization cycles and (b) EFS. The TN method outperforms all other methods in both comparisons.

10 CONCLUSIONS

A Truncated Newton method for computing an approximation to the second-order correction of the design variables by iteratively solving the Newton equations using GMRES was presented. The method builds on previous work of the authors, by extending the mathematical formulation for a different grid displacement model and investigating the impact of the linear solver used to iteratively solve the Newton equations. It was observed that due to a slight non-symmetry of the Hessian matrix, caused by the lack of convergence of the adjoint equations to machine accuracy, CG may become inefficient when computing the solution of the Newton equations. Using GMRES as the linear solver within the TN loop significantly accelerated

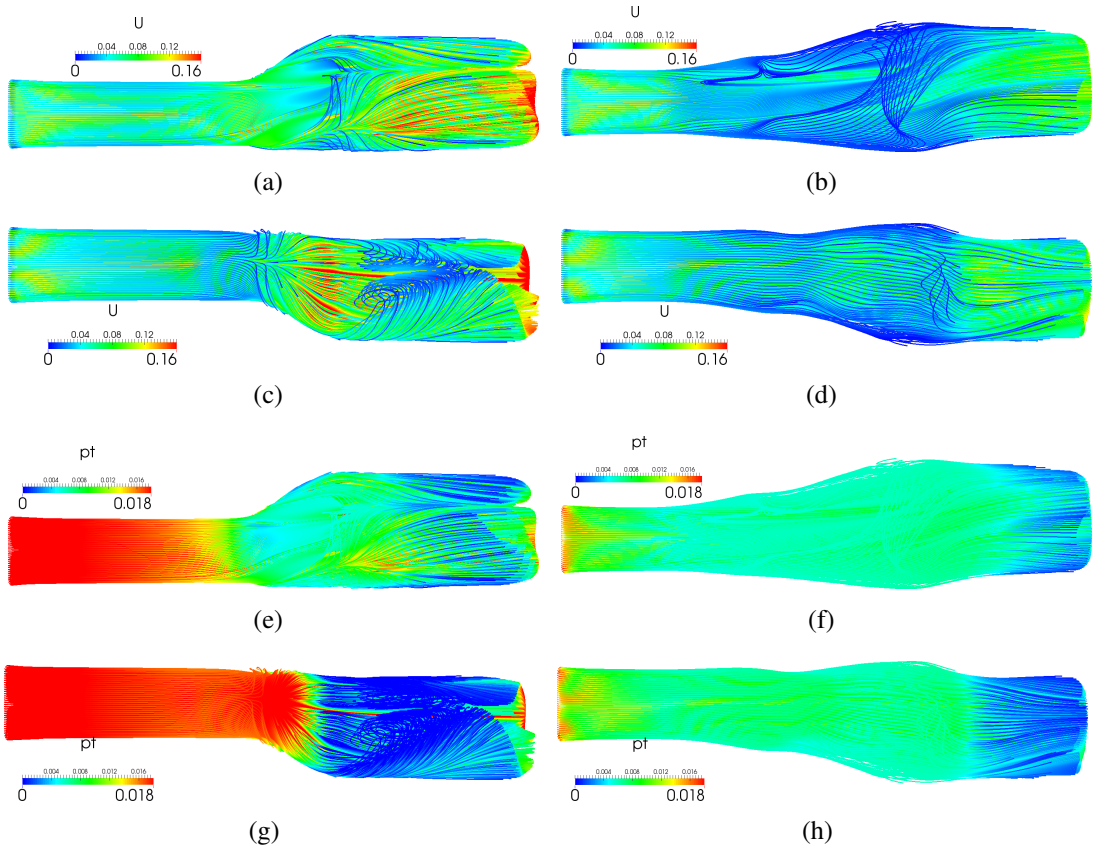


Figure 5: S-bend duct optimization: Velocity streamlines plotted for the reference (left column) and optimized (right column) geometries. In the top four figures, streamlines are coloured based on the flow velocity while, in the bottom four, on the total pressure values. The intense flow recirculation close to the bottom side of the wall (figs. c and g) has drastically been reduced (figs. d and h), leading to a reduction of about 60% in the objective function.

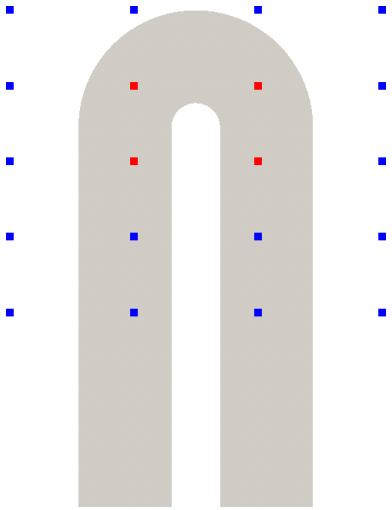
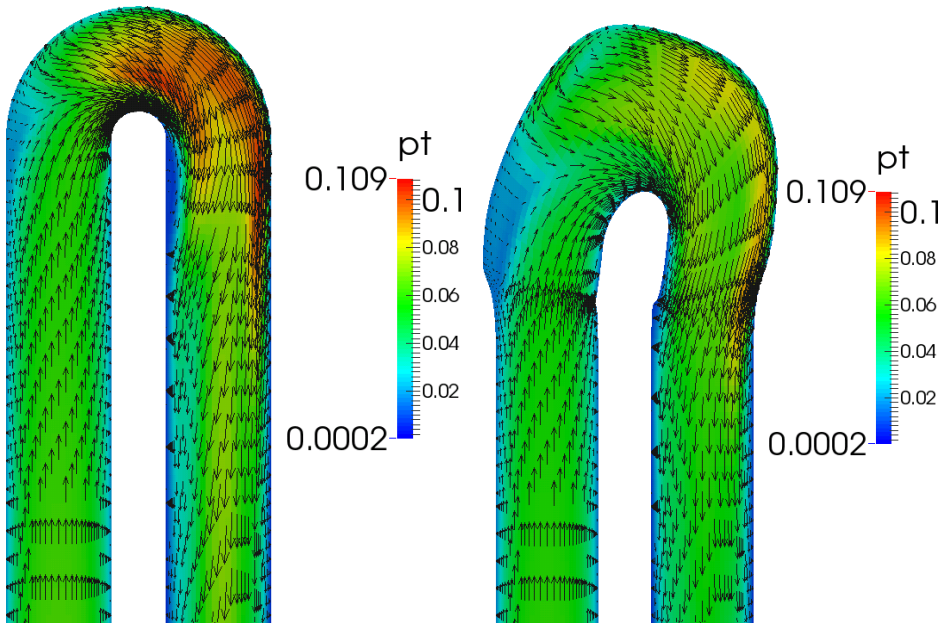
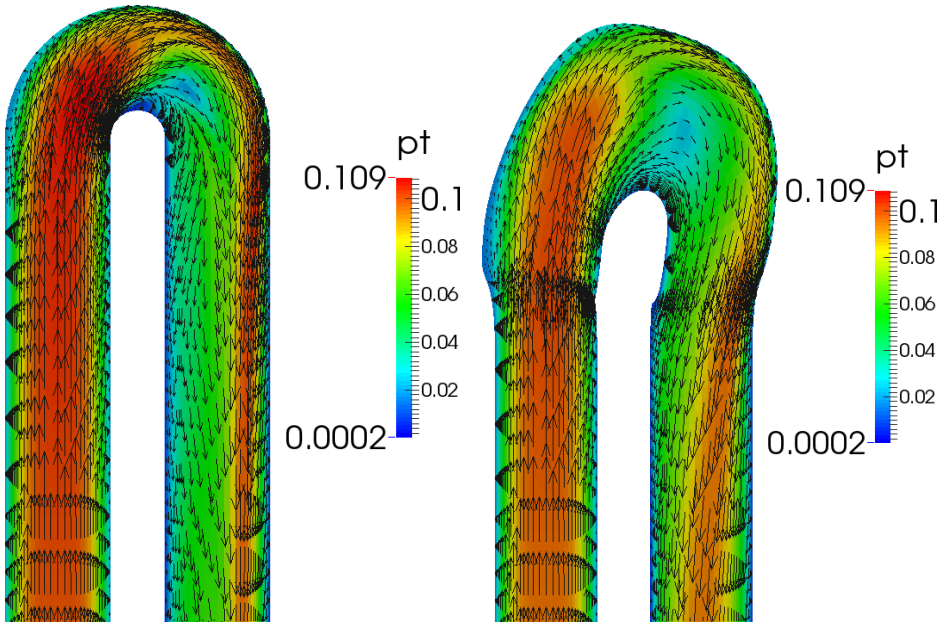


Figure 6: 3*D* U-bend optimization: Part of the duct shape along with one of two iso-z control point planes. Red control points are allowed to vary while blue ones are kept fixed during the optimization.

the convergence. The proposed TN method computes the required Hessian-vector products by utilizing a combination of (continuous) adjoint and direct differentiation. The cost per optimization cycle is approximately equal to $2+2M$ equivalent flow solutions, where M is the number of



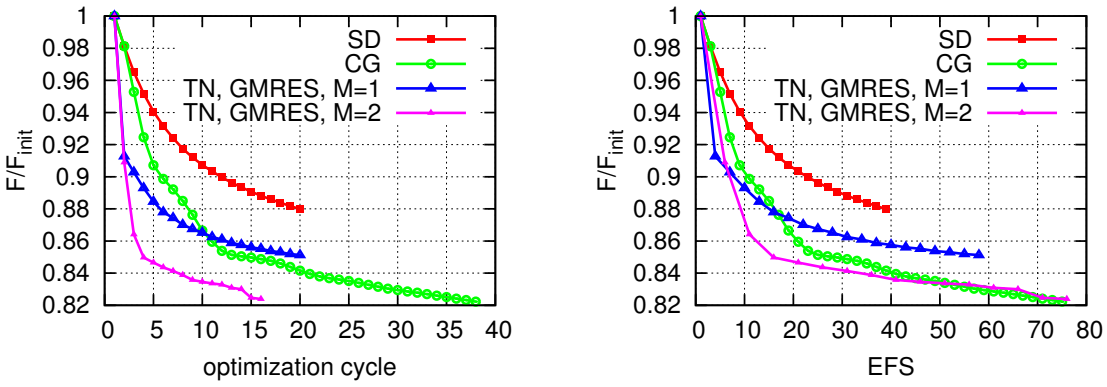
(a) reference, 10% thickness (b) optimized, 10% thickness



(c) reference, 50% thickness (d) optimized, 50% thickness

Figure 7: 3*D* U-bend optimization: Total pressure field plotted for the reference (left column) and optimized (right column) geometries, for a slice residing at 10% (top) and 50% of the duct height.

GMRES iterations used to approximate the solution of the Newton equations; this cost is practically independent of the design variables number. In the two applications presented, it was shown that TN outperforms other optimization methods in terms of optimization cycles and is, at least, as fast in terms of CPU cost. On going research, including the preconditioning of the Newton system and appropriate initialization, for further improving the TN method speed-up is performed.



(a) (b)

Figure 8: 3*D* U-bend optimization: Convergence of the SD, CG and TN methods wrt (a) optimization cycles and (b) EFS. For this case, $M=2$ has to be used for TN to outperform the CG method.

A ON THE SYMMETRY OF THE HESSIAN MATRIX

In this appendix, the symmetry of the Hessian matrix computed using the AV-DD approach (i.e. the approach used to also compute the Hessian-vector product in the TN approach) is examined. Since the continuous gradient and Hessian expressions are quite lengthy, the discrete approach is going to be used in this appendix. The conclusions, however, can be extended to the continuous formulation as well.

Let the objective function F and discretized residuals $\mathbf{R}$ be functions of the design, $\mathbf{b}$, and flow variables, $\mathbf{U}(\mathbf{b})$, i.e. $F=F(\mathbf{b},\mathbf{U}(\mathbf{b}))$ and $\mathbf{R}=\mathbf{R}(\mathbf{b},\mathbf{U}(\mathbf{b}))$. After introducing the augmented objective function as $F_{aug}=F+\Psi_k R_k$ and differentiating it w.r.t. to $\mathbf{b}$, we get

$$\frac{dF_{aug}}{db_i}=\frac{\partial F}{\partial b_i}+\Psi_k\frac{\partial R_k}{\partial b_i}+\left(\frac{\partial F}{\partial U_m}+\Psi_k\frac{\partial R_k}{\partial U_m}\right)\frac{dU_m}{db_i} \tag{26}$$

from which the adjoint equations and sensitivity derivatives are derived as

$$R_m^{\Psi}=\frac{\partial F}{\partial U_m}+\Psi_k\frac{\partial R_k}{\partial U_m}=0 \tag{27}$$

$$\frac{dF}{db_i}=\frac{\partial F}{\partial b_i}+\Psi_k\frac{\partial R_k}{\partial b_i} \tag{28}$$

Differentiating eq. 28 once more w.r.t. the components of $\mathbf{b}$ gives

$$\frac{d^2F}{db_i db_j}=\frac{\partial^2 F}{\partial b_i\partial b_j}+\Psi_k\frac{\partial^2 R_k}{\partial b_i\partial b_j}+\frac{\partial^2 F}{\partial b_i\partial U_k}\frac{dU_k}{db_j}+\Psi_k\frac{\partial^2 R_k}{\partial b_i\partial U_m}\frac{dU_m}{db_j}+\frac{\partial R_k}{\partial b_i}\frac{d\Psi_k}{db_j} \tag{29}$$

The Hessian expression given by eq. 29 is not symmetric, since permuting i and j does not yield $\frac{d^2F}{db_i db_j}=\frac{d^2F}{db_j db_i}$. This non-symmetric expression actually allows the computation of Hessian-vector products with a cost that does not depend on the design variables number in TN methods. However, since eq. 26 and eq. 28 are equivalent (upon the convergence of the residuals of the

adjoint equations to machine accuracy) for computing dF/db_i, differentiating eq. 26 yields

$$\begin{aligned}\frac{d^2F}{db_i db_j} &= \frac{\partial^2 F}{\partial b_i \partial b_j} + \Psi_k \frac{\partial^2 R_k}{\partial b_i \partial b_j} + \frac{\partial^2 F}{\partial b_i \partial U_k}\frac{dU_k}{db_j} + \frac{\partial^2 F}{\partial b_j \partial U_k}\frac{dU_k}{db_i} + \Psi_k \frac{\partial^2 R_k}{\partial b_i \partial U_m}\frac{dU_m}{db_j} \\ &+ \Psi_k \frac{\partial^2 R_k}{\partial b_j \partial U_m}\frac{dU_m}{db_i} + \frac{\partial^2 F}{\partial U_m \partial U_k}\frac{dU_m}{db_i}\frac{dU_k}{db_j} + \Psi_k \frac{\partial^2 R_k}{\partial U_m \partial U_l}\frac{dU_m}{db_i}\frac{dU_l}{db_j} \\ &+ \left(\frac{\partial R_k}{\partial b_i} + \frac{\partial R_k}{\partial U_m}\frac{\partial U_m}{\partial b_i}\right)\frac{d\Psi_k}{db_j} + \left(\frac{\partial F}{\partial U_m} + \Psi_k \frac{\partial R_k}{\partial U_m}\right)\frac{d^2 U_m}{db_i db_j}\end{aligned} \tag{30}$$

The sum of the first eight terms on the r.h.s. of eq. 30 is symmetric while the last two terms are zero since they include dR_k/db_i and R_k^{Ψ}, respectively. Hence, since the expression in eq. 30 is symmetric and eqs. 29 and 30 are equivalent, upon the convergence of the adjoint equations to machine accuracy, the Hessian matrix obtained through eq. 29 is symmetric as well.

However, if eq. 27 is not converged to machine accuracy, i.e. if

$$\tilde{R}_m^{\tilde{\Psi}} = \frac{\partial F}{\partial U_m} + \tilde{\Psi}_k \frac{\partial R_k}{\partial U_m} + r_m^a = 0, \ , \ r_m^a \neq 0 \tag{31}$$

where $\tilde{\Psi}_k$ is the slightly non-converged adjoint solution and r_k^a the adjoint residual, differentiation of the equivalent of eq. 26 would yield

$$\begin{aligned}\frac{d^2F}{db_i db_j} &= \frac{\partial^2 F}{\partial b_i \partial b_j} + \tilde{\Psi}_k \frac{\partial^2 R_k}{\partial b_i \partial b_j} + \frac{\partial^2 F}{\partial b_i \partial U_k}\frac{dU_k}{db_j} + \frac{\partial^2 F}{\partial b_j \partial U_k}\frac{dU_k}{db_i} + \tilde{\Psi}_k \frac{\partial^2 R_k}{\partial b_i \partial U_m}\frac{dU_m}{db_j} \\ &+ \tilde{\Psi}_k \frac{\partial^2 R_k}{\partial b_j \partial U_m}\frac{dU_m}{db_i} + \frac{\partial^2 F}{\partial U_m \partial U_k}\frac{dU_m}{db_i}\frac{dU_k}{db_j} + \Psi_k \frac{\partial^2 R_k}{\partial U_m \partial U_l}\frac{dU_m}{db_i}\frac{dU_l}{db_j} \\ &+ \frac{dR_k}{db_i}\frac{d\tilde{\Psi}_k}{db_j} + \tilde{R}_m^{\tilde{\Psi}}\frac{d^2 U_m}{db_i db_j} + \frac{dr_k^a}{db_i}\frac{d\tilde{\Psi}_k}{db_j}\end{aligned} \tag{32}$$

Eq. 32 states that if the adjoint equations are not converged to machine accuracy, eqs. 29 and 30 are no longer equivalent due to the last, non-symmetric term in eq. 32.

REFERENCES

[1] M. Ghavami Nejad, EM. Papoutsis-Kiachagias, and KC. Giannakoglou. Aerodynamic shape optimization using the adjoint-based truncated newton method. In *EUROGEN 2015, 11th International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems*, Glasgow, UK, September 14-16 2015.

[2] IS. Kavvadias, EM. Papoutsis-Kiachagias, and KC. Giannakoglou. On the proper treatment of grid sensitivities in continuous adjoint methods for shape optimization. *Journal of Computational Physics*, 301:1–18, 2015.

[3] MJ. Martin, E. Andres, C. Lozano, and E. Valero. Volumetric B-splines shape parametrization for aerodynamic shape design. *Aerospace Science and Technology*, 37:26–36, 2014.

[4] DI. Papadimitriou and KC. Giannakoglou. Aerodynamic shape optimization using first and second order adjoint and direct approaches. *Archives of Computational Methods in Engineering*, 15(4):447–488, 2008.

[5] DI. Papadimitriou and KC. Giannakoglou. Aerodynamic design using the truncated Newton algorithm and the continuous adjoint approach. *International Journal for Numerical Methods in Fluids*, 68:724–739, 2012.

[6] EM. Papoutsis-Kiachagias and KC. Giannakoglou. Continuous adjoint methods for turbulent flows, applied to shape and topology optimization: Industrial applications. *Archives of Computational Methods in Engineering, DOI 10.1007/s11831-014-9141-9*, 2014.

[7] EM. Papoutsis-Kiachagias, N. Magoulas, J. Mueller, C. Othmer, and KC. Giannakoglou. Noise reduction in car aerodynamics using a surrogate objective function and the continuous adjoint method with wall functions. *Computers & Fluids*, 122:223–232, 2015.

[8] L. Piegl and W. Tiller. *The NURBS book*. Springer, 1997.

[9] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

UNSTEADY ADJOINT TO THE CUT-CELL METHOD USING MESH ADAPTATION ON GPUS

K.D. Samouchos[1], S.G. Katsanoulis[1], and K.C. Giannakoglou[1]

[1]National Technical University of Athens, School of Mech. Eng., Parallel CFD & Optimization Unit, Athens , Greece
e-mail: ksamouchos@yahoo.com, skatsanoulis@gmail.com, kgianna@central.ntua.gr

Keywords: Cut-Cell Method, Moving Boundaries, Adjoint-Based Optimization, GPUs.

Abstract. *In its first part, this paper presents a cut-cell method for simulating 2D unsteady inviscid flows of compressible fluids in domains with moving boundaries. To solve shape optimization problems, the gradient of the aerodynamic shape optimization with respect to (w.r.t.) the design variables is computed via the continuous adjoint approach.*

An automatic grid adaptation method based on a quad-tree data structure allows low memory usage for storing geometric data. The no-penetration condition along the surface of the emerged bodies, is implemented using a second-order cut-cell approach. To avoid numerical instabilities during the solution of the flow and adjoint PDEs, very small cut-cells are merged with neighboring active cells to yield the finite-volume where the flow or adjoint equations are integrated.

In transonic flow simulations, the Cartesian grid is adapted not only to the solid boundaries but, also, to the evolving flow discontinuities. The refined grid follows the moving solid wall and discontinuities, by means of local refinement and derefinement processes. The paper focuses on the schemes used to interpolate the flow solution fields between successive time-steps, as the grid becomes adapted to the moving geometry. During the solution of the unsteady adjoint PDEs, since the adjoint solver marches backwards in time and the adapted Cartesian grids are continuously changing, care should be taken so as to have full access to the necessary geometric quantities at all cells, even if these did not exist at previous time-steps. Both the primal and adjoint solvers are programmed on GPUs (Graphics Processing Units), using CUDA-C, to reduce the optimization wall-clock time.

Results are presented for both the analysis and optimization problems.

1 Introduction

Flows with complex moving boundaries are of utmost importance in many engineering applications. The traditional approach involves the use of body-fitted grids and the corresponding discretization methods. Nevertheless, irrespective of the grid type (structured or unstructured), these methods require a highly delicate and often costly grid deformation process, especially when large boundary movements occur. A viable alternative encompasses the use of the Immersed Boundary Method, originally proposed by Peskin [1]. In the latter, the computation takes place on a Cartesian grid allowing for a simple and automatic grid generation.

Mittal and Iaccarino [2] discern two basic approaches stemming from the generalized term "Immersed Boundary Methods", namely the continuous and discrete forcing approaches. The continuous one makes use of an additional source term in the momentum equations to approximate the effect of solid boundaries. On the contrary, discrete forcing approaches rely on the discrete representation of the immersed boundary without altering the governing equations. Prevailing among them is the cut-cell finite volume method [3, 4] which guarantees both global and local conservation. This is achieved through the reshaping of cells that are intersected by the solid boundary, i.e., the so-called cut-cells, by making them conform to the boundary while the remaining fraction of the cut-cells as well as cells lying in the interior of solid bodies are totally discarded.

The present paper deals with the development of a cut-cell finite volume method for 2D inviscid flows of compressible fluids. In order to avoid numerical stability issues caused by the existence of small cells, a cell-merging technique is adopted. To maximize the accuracy of the flow simulation, the grid in the vicinity of both the solid boundary and flow discontinuities (e.g. shock waves) is refined. Extending this idea to moving boundaries, grid is dynamically adapted following their motion, which calls for a method to project the solution onto the grid in the next time-step. The software is developed in order to run on NVIDIA GPUs.

In addition, an optimization loop is implemented based on the continuous adjoint approach. That is, after the adjoint PDEs are derived from the flow equations, both for steady and unsteady flows, these are discretized and numerically solved utilizing the already developed method. In the course of this process, the correct computation of the derivatives of geometric quantities is crucial since the special nature of the stationary Cartesian grid must taken into account. Problems related to the maximization of lift of a stationary airfoil as well as the time-averaged lift of a pitching airfoil are showcased.

2 Grid Generation

The computational grid is dynamically generated based on criteria which make it compatible with the cut-cell approach for solving the flow and adjoint equations around stationary and moving bodies. Starting point is a uniform Cartesian grid within a rectangular domain; the initial cell volume is defined by the user (Ω_{max}). Then, cells intersected by the solid walls are recursively subdivided into four quadrants. The basic rule governing this iterative process is that cut–cells with area greater than Ω_{min} should further be subdivided; Ω_{min} is a threshold value defined by the user. This extra cell refinement may affect neighboring non–cut–cells which should also be subdivided in order to meet a second rule stating that a cell face cannot have more than two neighbors. This rule prevents the formation of tiny grid cells in contact with much greater ones, which makes the CFD solver prone to numerical inaccuracies. Two additional rules must be satisfied, see fig. 1. According to the first of them, cut-cells are not allowed to have any of their four edges intersected twice by the solid wall. The last rule does

not allow cut-cells to have their four edges intersected by solid walls. In either case, the cell must be subdivided further. In the case of a stationary isolated airfoil, the refined Cartesian grid for an inviscid flow simulation looks like that shown in fig. 2.

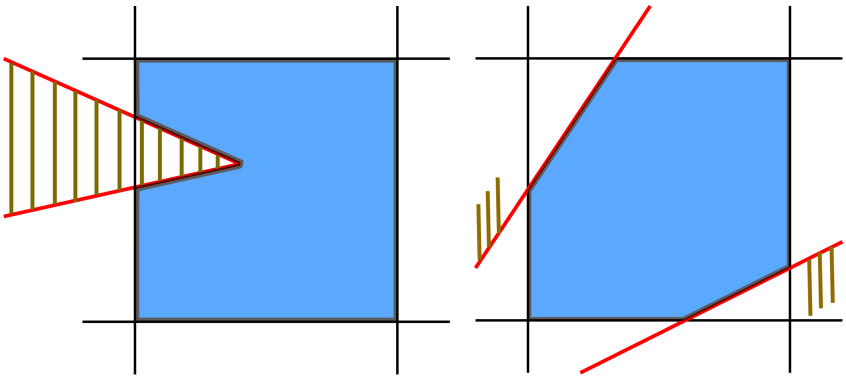
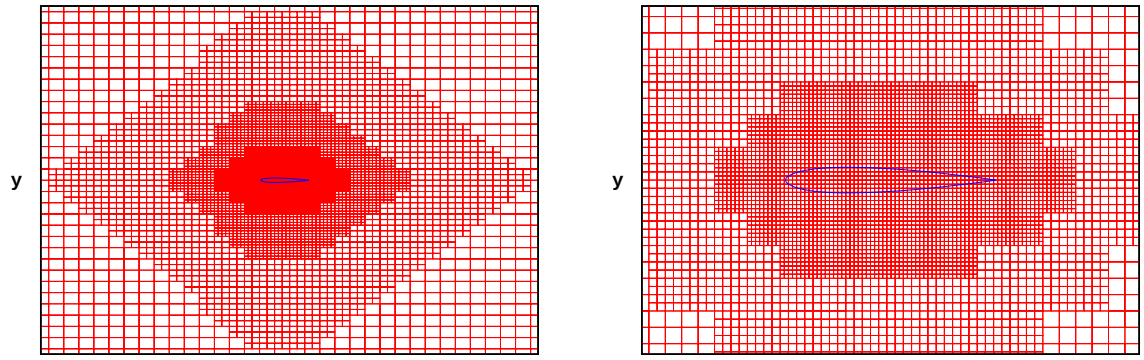


Figure 1: Visual representation of the third and fourth rule imposed during the Cartesian grid adaptation process, as described in the text. Both cases are not acceptable.



(a) (b)

Figure 2: Two views on the Cartesian grid in the vicinity of the boundary of an isolated airfoil, for an inviscid flow simulation. Further local refinement is possible if flow discontinuities appear.

Grid generation is based on a quad-tree data structure [5] which allows fast operations and ensures easy identification and access to neighboring cells, as required during the numerical solution of the flow and adjoint equations. Each and every cell is given a unique pair of integers (i, j), facilitating a lot the computation of geometric quantities such as cell volumes and barycentric coordinates. Among other, this two-index numbering system results to a noticeable reduction in memory footprint.

A special treatment of the cut–cells is needed in order to satisfy the conservation equations in the vicinity of solid boundaries with the required accuracy. For all cut–cells, their parts belonging to the non-fluid domain are discarded. This gives rise to fluid or active cells which can be shaped as triangles, quadrangles or pentagons; taking into account the aforementioned constraints, no other shape is possible in a 2D problem. By definition, the active cell area corresponds solely to the part of the cell which is within the fluid. In the course of this process, new geometric quantities such as the newly formed edges, the active area and the corresponding barycenter position appear.

A common problem in the cut-cell method is the creation of very small cells, depending on the instantaneous locations of the solid walls within the grid. It is known [7, 8] that the presence of these cells might cause convergence issues. To tackle this problem, a cell-merging approach is used, as shown in fig. 3, according to which the flow variables are stored at the barycenter of the merged cell.

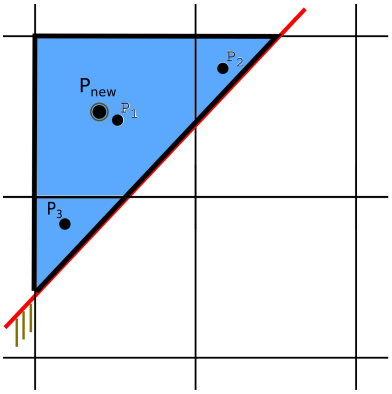


Figure 3: Three cut–cells with barycenters P_1, P_2 and P_3. Since P_2 and P_3 are considered as very small cut-cells, these are merged into a single cell with barycenter P_{new}. Flow variables for the newly formed cell are all stored at P_{new}.

3 Governing Equations & Discretization

The Euler equations for unsteady compressible flows in an inertial reference frame are written as

$$\frac{\partial \overrightarrow{U}}{\partial t} + \frac{\partial \overrightarrow{f}_x}{\partial x} + \frac{\partial \overrightarrow{f}_y}{\partial y} = \overrightarrow{0} \tag{1}$$

where $\overrightarrow{U} = \begin{bmatrix} \rho & \rho u & \rho v & E \end{bmatrix}^T$ is the vector of conservative variables and $\overrightarrow{f}_x = \begin{bmatrix} \rho u & \rho u^2 + p & \rho uv & u(E+p) \end{bmatrix}^T$, $\overrightarrow{f}_y = \begin{bmatrix} \rho v & \rho uv & \rho v^2 + p & v(E+p) \end{bmatrix}^T$ are the inviscid fluxes in the x and y direction, respectively, ρ the fluid density, u and v the Cartesian velocity components, p the static pressure and E the total energy per unit volume.

The integration of equations (1) is based on the cell-centered finite volume scheme. The integration volume consists of the part of the cell that lies in the fluid. For the merged cells, this volume is defined as the fluid part of each of the constituent cells (e.g. the blue area in fig. 4). According to this, if Ω_P^{k+1} is the control volume of a cell with centroid P (to be referred to as cell P) at time-step $k+1$, S_P^{k+1} is its boundary and $\overrightarrow{n}$ is its normal unit vector, the corresponding residual of the flow equations is

$$\overrightarrow{R}_P := \int\limits_{\Omega_P^{k+1}} \frac{\partial \overrightarrow{U}}{\partial t} d\Omega + \int\limits_{S_P^{k+1}} \overrightarrow{f}_i \cdot n_i \, dS = \overrightarrow{0} \tag{2}$$

The discretization of eqs. (2) is based on the Roe [9] scheme. The no-penetration condition is imposed along the solid walls, that is, $(\overrightarrow{u} - \overrightarrow{u}_w) \cdot \overrightarrow{n} = \overrightarrow{0}$, where $\overrightarrow{u}_w$ is the wall velocity. Based on the latter, at all cut-cells' boundary edges that correspond to solid walls the flux vector becomes $\overrightarrow{f}_w = \begin{bmatrix} \rho u_{w,n} & \rho u_{w,n} u + p n_x & \rho u_{w,n} v + p n_y & u_{w,n}(p+E) \end{bmatrix}^T$ where $u_{w,n} = \overrightarrow{u}_w \cdot$

$\overrightarrow{n}$. This flux corresponds to the center B of the solid boundary segment of the cut–cell as shown in fig. 4(a). Variables appearing in $\overrightarrow{f}_w$ are computed at point B using the Taylor expansion formula $\overrightarrow{U}_B = \overrightarrow{U}_P + \overrightarrow{PB} \cdot \left(\nabla \overrightarrow{U}\right)_P$ where $\left(\nabla \overrightarrow{U}\right)_P$ is computed using a weighted least–squares method. The same process can be extended in the case of merged cells, as shown in fig. 4(b)

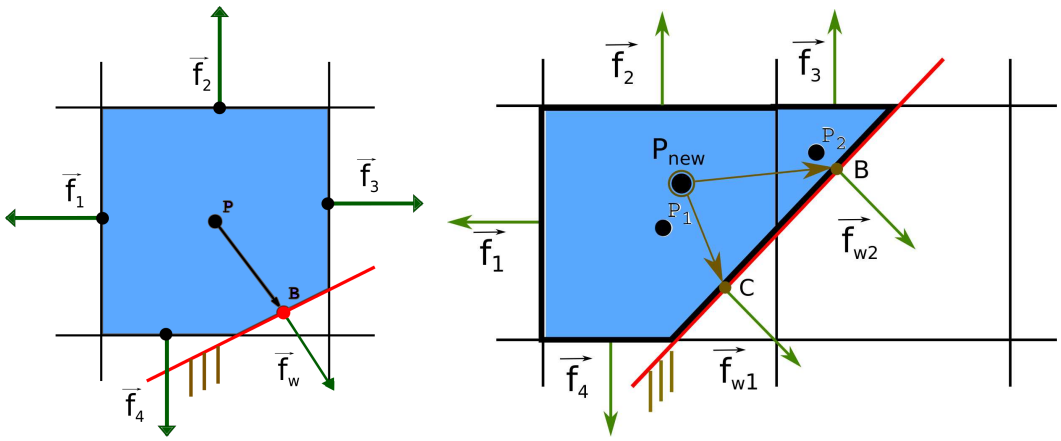


Figure 4: (a): The no-penetration condition is imposed via $\overrightarrow{f}_w$ at point B using the extrapolated variables from barycenter P. (b): Two cut–cells with barycenters P_1 and P_2 are merged into a single cell with barycenter P_{new}. Values at P_{new} are extrapolated to B and C in order to compute $\overrightarrow{f}_{w1}$ and $\overrightarrow{f}_{w2}$.

Regarding the discretization of the temporal term in eq. (2) when the immersed body moves inside the Cartesian grid, four different cases might occur. The simplest case is that of fluid cells at time-step t^k which remain intact at t^{k+1}. Thus, the integration volume is unchanged and a second-order backward differentiation scheme is readily applied. The second case concerns all cells being intersected by the solid boundary at any of the three time-steps (t^{k-1}, t^k, t^{k+1}). To account for changes in Ω_P, the Reynolds transport theorem [10] is applied,

$$\int\limits_{\Omega_P^{k+1}} \frac{\partial \overrightarrow{U}}{\partial t} d\Omega = \frac{\partial}{\partial t} \int\limits_{\Omega_P^{k+1}} \overrightarrow{U}\, d\Omega - \int\limits_{S_P^{k+1}} \overrightarrow{U}\, u_{w,n}\, dS \tag{3}$$

For the first term on the r.h.s., a second-order backward scheme is used

$$\frac{\partial}{\partial t} \int\limits_{\Omega_P^{k+1}} \overrightarrow{U}\, d\Omega \simeq \frac{1}{2\Delta t} \left(3\overrightarrow{U}_P^{k+1} \Omega_P^{k+1} - 4\overrightarrow{U}_P^{k} \Omega_P^{k} + \overrightarrow{U}_P^{k-1} \Omega_P^{k-1} \right) \tag{4}$$

while, for the surface term, the assumption

$$\begin{aligned} \int\limits_{S_P^{k+1}} \overrightarrow{U}\, u_{w,n}\, dS &\simeq \overrightarrow{U}_w^{k+1} \int\limits_{S_P^{k+1}} u_{w,n}\, dS \\ &= \overrightarrow{U}_w^{k+1} \left(\frac{\partial \Omega_P}{\partial t} \right)^{k+1} \simeq \frac{\overrightarrow{U}_w^{k+1}}{2\Delta t} \left(3\Omega_P^{k+1} - 4\Omega_P^{k} + \Omega_P^{k-1} \right) \end{aligned} \tag{5}$$

is made, where $\overrightarrow{U}_w^{k+1}$ is the vector of flow variables computed at the wall face.

According to the third case, a cell lying entirely in the solid boundary at t^k, becomes fluid cell at t^{k+1}. In such a case, Ω_P^k and Ω_P^{k-1} are equal to zero and eqs. (4) and (5) are properly transformed. Finally, cells that are solidified should transfer their conservative values to their neighboring cells in the fluid region via a merging process.

4 Dynamic Grid Adaptation

In case of moving solid boundaries, the grid must continuously be adapted to the new boundary shape. During the dynamic adaptation, which embodies both coarsening and refinement tasks, the resulting grid at each new time-step must comply with the four rules mentioned in section (2). The grid adaptation from one time-step (t^k) to the next (t^{k+1}) takes place as follows:

Initially, the grid undergoes a coarsening process, according to which every cell with volume less than Ω_{max} is merged with its neighbors. Then, the solid boundary is moved to its new position and, starting from the just coarsened grid, cells split anew in the vicinity of the displaced wall. Through this procedure, grid refinement in the vicinity of the new position of the solid boundary is achieved while the no longer necessary refined grid around the previous wall position is canceled. This process is exemplified in fig. 5. Along with the grid adaptation, interpolation should be employed to transfer the flow solution from the previous time-step (t^k) to the new one (t^{k+1}).

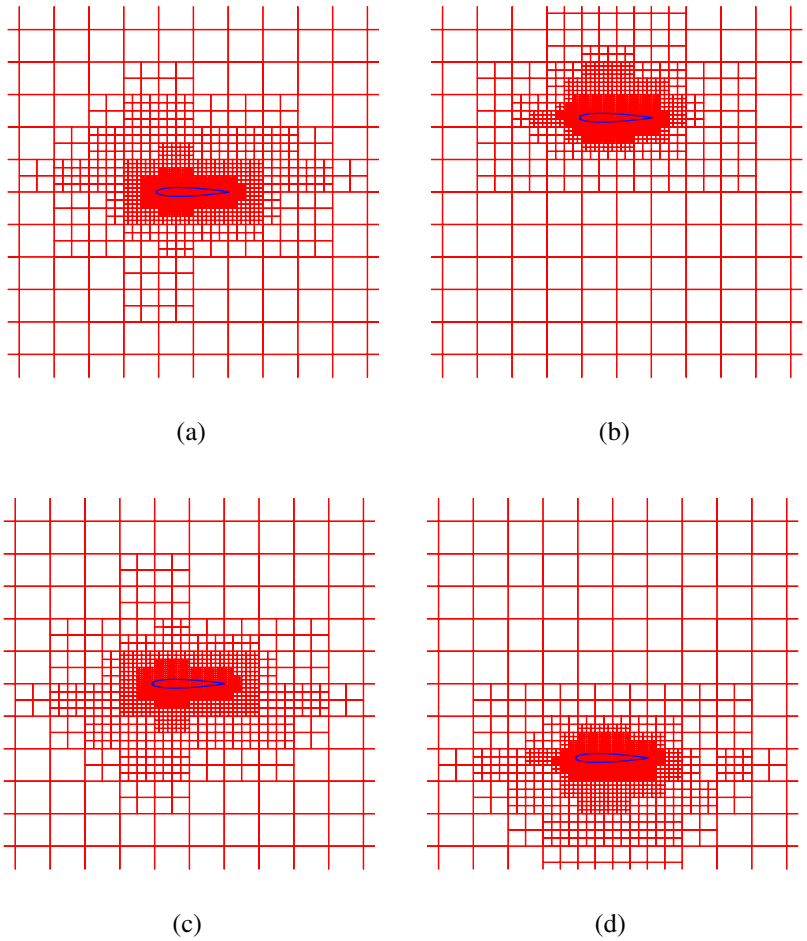


Figure 5: Translational periodic, in the vertical direction, motion of an airfoil. Adapted grid at four time-steps within the same period.

In order to attribute values to the conservative variables at the barycenter of the newly formed cells at t^{k+1}, an intermediate step (t^*) at which the solid boundary is considered to its previous position (t^k) and, then, immersed in the new grid (t^{k+1}), is used. The flow variables $\overrightarrow{U}^*$ at the intermediate step represent the time history of the newly created cells at t^{k+1}. The values transferred from t^k to t^* must ensure the satisfaction of the conservation laws.

For the sake of clarity, two simple cases will be demonstrated. According to the first one, due to the movement of the solid boundary, a cell at t^k is decomposed into four smaller cells at t^{k+1}, fig. 6. Considering that each of the newly formed cells P_i should share the same $\overrightarrow{U}$ values

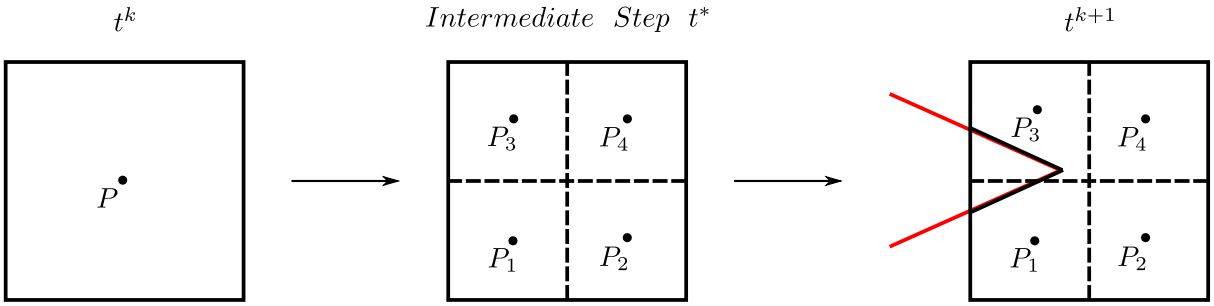


Figure 6: Cell P at step t^k, decomposed into four smaller cells P_i at step t^{k+1}, makes use of an intermediate step (t^*) in order to attribute values to the flow variables $\overrightarrow{U}$ to each of the newly formed cells at step t^{k+1}.

at t^k, it is

$$\overrightarrow{U}^*_{P_i} = \overrightarrow{U}^k_P \,, \quad i = 1, \ldots, 4 \tag{6}$$

where $\overrightarrow{U}^k_P$ is the flow variables array of cell P, at t^k, and $\overrightarrow{U}^*_{P_i}$ is the unknown vector of conservative variables for each newly formed cell P_i at t^*.

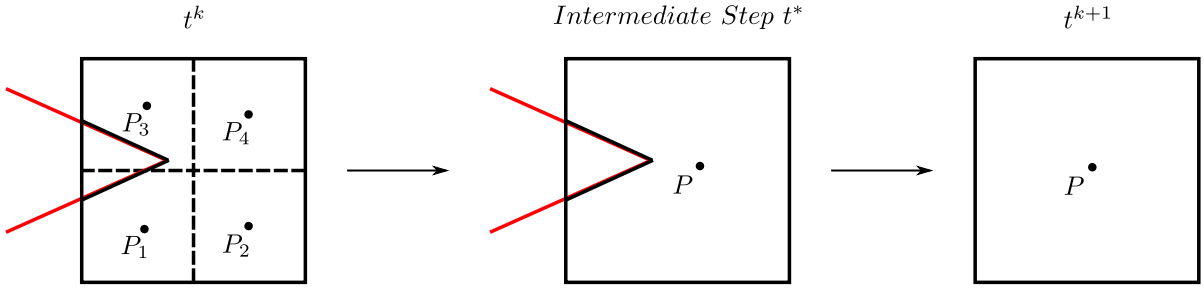


Figure 7: The process of associating values of the flow variables $\overrightarrow{U}$ to the newly created cell barycenter P at t^{k+1}.

In the second case, four cells P_i at t^k are merged to form a single new cell, fig. 7. In this case,

$$\overrightarrow{U}^*_P = \frac{\sum_{i=1}^{4} \overrightarrow{U}^k_{P_i} \, \Omega^k_{P_i}}{\Omega^*_P} \tag{7}$$

The values $\overrightarrow{U}^*_P$, computed by eqs. (6) and (7) and Ω^*_P, are considered as the flow fields and volumes, respectively, at t^k and are referred to as $\overrightarrow{U}^k_P$ and Ω^k_P in eq. (4). The same process must be followed in order to compute $\overrightarrow{U}^{k-1}_P$ and Ω^{k-1}_P.

5 Validation of the flow solver

In this section, the use of the flow solver in steady and unsteady flows is demonstrated and validated against other results. The steady flow case encompasses grid refinement due to the formation of a shock wave over the suction-side of an airfoil with non-zero incidence, while the unsteady case showcases the grid refinement according to both the airfoil motion and the formation of the shock wave.

The software was programmed in CUDA-C and runs on NVIDIA GPUs by taking advantage of the superior hardware characteristics they exhibit compared to CPUs. In order to exploit the advantages of GPUs, special care must be taken in memory handling. Moreover, an appropriate solution method should be chosen, which can be efficiently parallelized. Here, the discretized equations are solved in each pseudo-time step using the Jacobi method. The simulation was carried out on a single computational node with two NVIDIA Tesla K40 GPUs with 12 GB of GPU memory each.

The NACA0012 airfoil is investigated at transonic flow conditions ($M_\infty = 0.8$ and $a_\infty = 1.25\ deg$). The Mach number field around the airfoil with the adapted grid in the vicinity of the shock wave is presented in fig. 8. The same case was studied using several Euler-flow solvers on body-fitted grids in [11] and, as shown in Table 1, the results of the present cut–cell solver are in good agreement with those solvers.

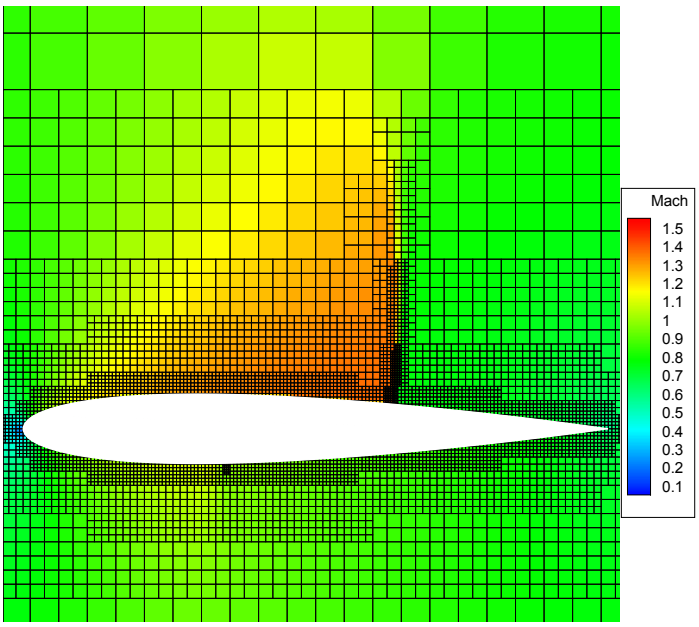


Figure 8: Transonic flow around NACA0012 airfoil-Steady flow: Computed Mach number field.

	C_L
Present cut–cell method	0.3492
Body–fitted Min	0.3482
Body–fitted Max	0.3562

Table 1: C_L comparison between the present cut-cell method and other solvers using body-fitted grids where Min and Max denote the spread of C_L values found in [11].

As far as the unsteady flow solver is concerned, its validation is carried out using the same airfoil undergoing a periodic pitching motion about $c/4$ (c is the chord) with 2.51^o amplitude and 0.016^o mean angle of attack. The Mach number is 0.755 and the reduced frequency $k := \frac{\omega}{2V_\infty}$ is 0.0814, where ω is the angular velocity of the airfoil. The computed Mach number fields at the two extreme positions of airfoil motion are shown in fig. 9.

This case has been investigated experimentally, by Landon [12]; over and above, an in-house Euler code [13] running on a body-fitted grid was used to get CFD results to compare with. Comparisons are presented in fig. 10.

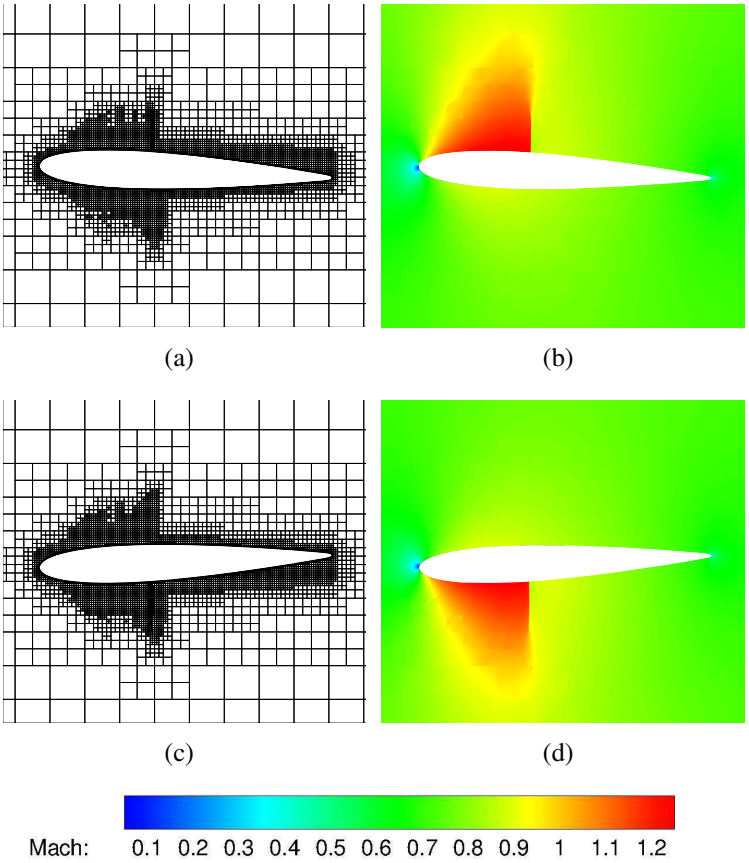


Figure 9: Transonic flow around the pitching NACA0012 airfoil: (a),(c): Dynamically adapted grid with further refinement in the vicinity of the shock wave and (b),(d): Mach number fields at the two extreme positions of airfoil motion.

6 Formulation of the Steady Adjoint Method

The objective function to be maximized is the lift of an airfoil, namely

$$F = \int_{S_w} p n_k r_k dS \tag{8}$$

where r_k are the components of the unit vector normal to the freestream velocity. The airfoil is parameterized using Bézier curves, the coordinates of the control points of which are the design variables (b_q) of the optimization process. The necessary derivative of F is computed by

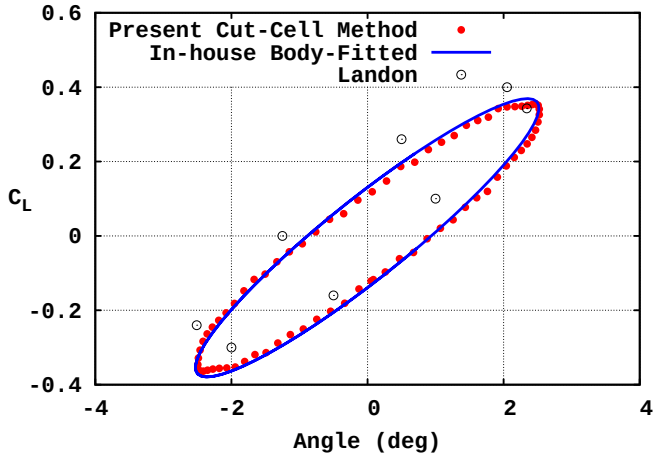


Figure 10: Time-evolution of the lift coefficient, in terms of the angle formed between the airfoil chord and the horizontal axis. Present results are contrasted with CFD results obtained using an in-house Euler solver for body-fitted grids [13] and measurements by Landon [12].

expanding it as an augmented objective function $F_{aug} = F + \int_{\Omega} \Psi_n R_n d\Omega$, where $\Psi_n,\ n = 1,4$ are the adjoint variable fields. By differentiating F_{aug} w.r.t. to the design variables and setting the multipliers of the variations in all flow variables equal to zero, the steady adjoint equations

$$-A_{nmk}\frac{\partial \Psi_n}{\partial x_k} = 0 \tag{9}$$

arise [6]. Boundary conditions to be imposed along the solid wall are $\Psi_{k+1} n_k + n_k r_k = 0$ and, in the farfield, $\Psi_n = 0,\ n = 1,4$. Then, the gradient of F w.r.t. the design variables b_q becomes

$$\begin{aligned}\frac{\delta F}{\delta b_q} &= \int_{S_w} p\frac{\delta (n_k r_k dS)}{\delta b_q} + \int_{S_w} (\Psi_{k+1} p - \Psi_i f_{ik})\frac{\delta n_k}{\delta b_q} dS \\ &- \int_{S_w} \Psi_i \frac{\partial f_{ik}}{\partial x_l} n_k \frac{\delta x_l}{\delta b_q} dS + \int_{S_w} \Psi_i \frac{\partial f_{il}}{\partial x_l}\frac{\delta x_k}{\delta b_q} n_k dS\end{aligned} \tag{10}$$

Eq. (10) requires the computation of $\frac{\delta x_k}{\delta b_q}$ which stands for the variation in the airfoil's surface w.r.t. the design variables. In contrast to body-fitted grids, the Cartesian grid used in the cut-cell method is stationary even if the geometry changes during the optimization process. This must be taken into consideration and, thus, the above term is computed at the intersection points of the geometry with the Cartesian grid. Each intersection point can move only along the corresponding cut-cell's edge. Consequently, it is the projection of $\frac{\delta x_k}{\delta b_q}$ to the edge direction that must only be taken into account in order to compute the derivatives at the mid-point of each cut-cell's solid edge, as shown in fig. 11.

7 Steady Adjoint Results

Below, the shape optimization of a symmetric airfoil is examined. The airfoil ($M_\infty = 0.44$ and $a_\infty = 0\ deg$) pressure and suction sides are both parameterized using Bézier curves. The design variables are the coordinates of six control points marked with empty circles, fig. 12.

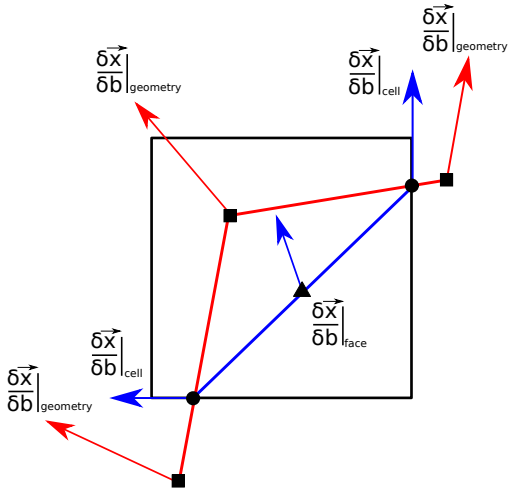


Figure 11: The red polyline denotes the surface of the geometry which cuts the cell in two points. The blue line segment is part of the boundary of the cut-cell. The derivatives of the boundary w.r.t. the design variable b (red vectors) are initially computed at the points marked with squares by differentiating the geometry's parameterization. These derivatives are used to compute $\frac{\delta\vec{x}}{\delta b}$ at the intersection points marked with circles which are taken into account to compute the derivative at the mid-point of the line segment.

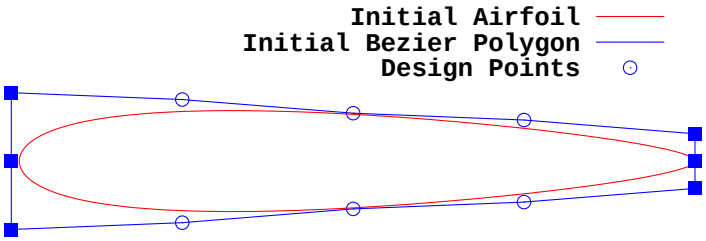


Figure 12: Bézier control polygon generating the initial shape of the airfoil. Points marked as empty circles signify control points, the coordinates of which are used as design variables.

The maximization of lift, defined by eq. (8), is chosen as the target of the optimization process.

After ten optimization cycles, the suction side is cambered enough, causing an increase in the lift coefficient from zero to approximately 0.36, as shown in figs. 13 and 14.

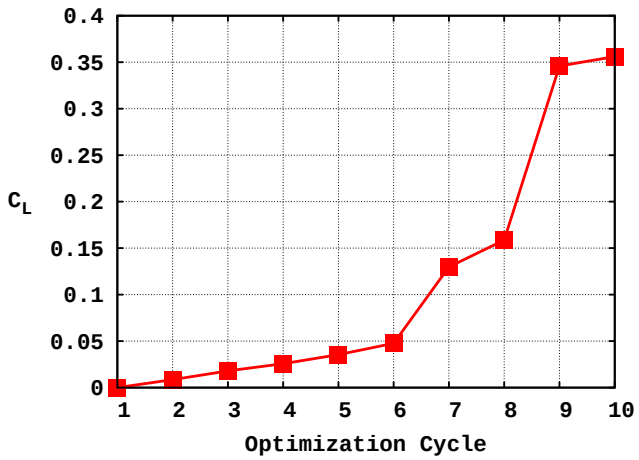


Figure 13: Subsonic flow around the airfoil parameterized as in fig. 12. Adjoint-based optimization for lift maximization. Lift coefficient evolution during the optimization cycles.

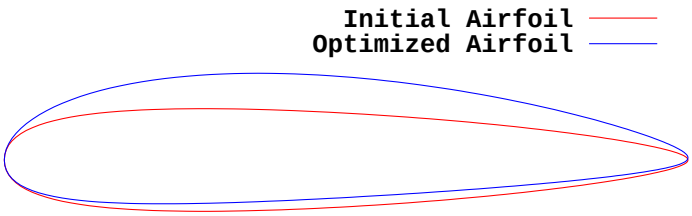
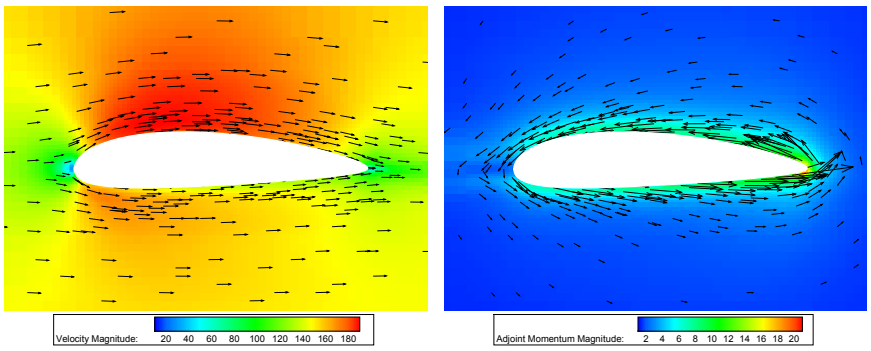


Figure 14: Subsonic flow around the airfoil parameterized as in fig. 12. Adjoint-based optimization for lift maximization. Initial and optimized airfoils.

For the objective function defined in eq. (8) with r_k pointing along the y-axis, the imposed adjoint conditions at the wall can be simplified as $\Psi_n = -n_y$, where Ψ_n is the normal adjoint momentum and n_y is the ordinate of the normal unit vector over the airfoil, pointing towards the solid. Along the pressure side, n_y is positive enforcing the adjoint momentum flux to exit the airfoil. In contrast, along the suction side, n_y is negative so the adjoint momentum flux points to the opposite direction. The aforementioned conditions result in the adjoint momentum magnitude field presented in fig. 15(b) while the corresponding flow velocity field is shown in fig. 15(a).



(a) (b)

Figure 15: Subsonic flow around the airfoil parameterized as in fig. 12. Adjoint-based optimization for lift maximization. (a): Velocity magnitude field and velocity vectors. (b): Adjoint momentum magnitude field and vectors. Both figures correspond to the optimized geometry.

8 Formulation of the Unsteady Adjoint Method

The unsteady adjoint method is tailored to the flow around of pitching airfoil, with period T. The objective function to be maximized is defined as

$$F = \frac{1}{T}\int_0^T \int_{S_w} p n_k r_k dS dt \tag{11}$$

The derivative of F is computed by differentiating the augmented objective function $F_{aug} = F + \int_0^T \int_\Omega \Psi_n R_n d\Omega dt$, where Ψ_n, $n = 1,4$ are the unsteady adjoint variable fields. The resulting unsteady adjoint equations are

$$-\frac{\partial \Psi_m}{\partial t} - A_{nmk}\frac{\partial \Psi_n}{\partial x_k} = 0 \tag{12}$$

The imposed boundary conditions along the moving solid wall are $\Psi_{k+1}n_k+\Psi_4 u_{n,w}+\frac{1}{T}n_k r_k=0$ and, at the farfield, $\Psi_n=0,\ n=1,4$. Then, the gradient of F w.r.t. the design variables b_q is

$$\begin{aligned}\frac{\delta F}{\delta b_q} &= \frac{1}{T}\int_0^T\int_{S_w} p\frac{\delta(n_k r_k dS)}{\delta b_q}dt+\int_0^T\int_{S_w}\left(\Psi_{k+1}p-\Psi_i f_{ik}\right)\frac{\delta n_k}{\delta b_q}dSdt \\ &- \int_0^T\int_{S_w}\Psi_i\frac{\partial f_{ik}}{\partial x_l}n_k\frac{\delta x_l}{\delta b_q}dSdt+\int_0^T\int_{S_w}\Psi_i\frac{\partial f_{il}}{\partial x_l}\frac{\delta x_k}{\delta b_q}n_k dSdt \\ &+ \int_0^T\int_{S_w}\Psi_i\frac{\partial U_i}{\partial x_l}\frac{\delta x_l}{\delta b_q}u_{n,w}dSdt+\int_0^T\int_{S_w}\left(\Psi_i U_i+p\Psi_4\right)\frac{\delta u_{n,w}}{\delta b_q}dSdt \end{aligned} \tag{13}$$

As described in section 4, the grid is dynamically adapted to the motion of the solid boundary. Due to the fact that the unsteady adjoint equations are solved backwards in time, each and every adapted grid must be stored. The quad-tree data structure allows for the minimum amount of data to be kept in memory, i.e. the pair of integers (i,j) (see section 2) and information needed to transfer values from one grid to the next one, in order to retrieve the proper grid at every time-step during the unsteady solution. Regarding the adjoint variables' projection to the grid at the next time-step, eqs. (6),(7) are used.

9 Unsteady Adjoint Results

An optimization loop is implemented for the pitching airfoil with amplitude 2.5 *deg*, at fixed infinite flow conditions of $M_\infty=0.44$ and $a_\infty=0$ *deg*. The optimization aims at maximizing the lift integral given by eq. (11). Starting from a symmetric airfoil, (fig. 12), where $F=0$ and after ten optimization cycles, the airfoil became cambered and the objective function increased to almost 0.32 (figs. 16, 17).

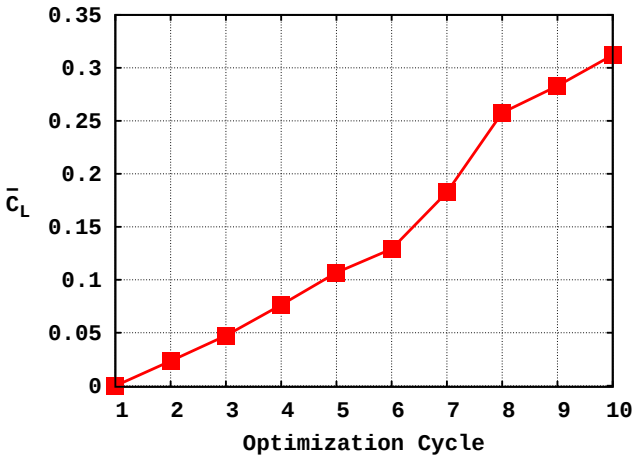


Figure 16: Unsteady subsonic flow around the pitching airfoil parameterized as in fig. 12. Adjoint-based maximization of the time-averaged lift. Variation in the objective function in terms of the optimization cycles.

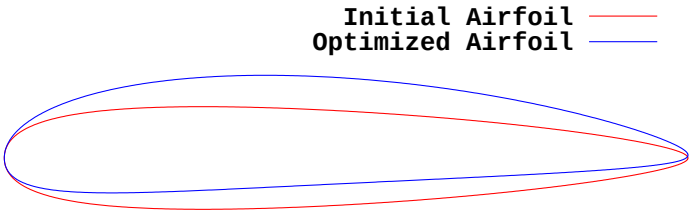


Figure 17: Unsteady subsonic flow around the pitching airfoil parameterized as in fig. 12. Adjoint-based maximization of the time-averaged lift. Initial and optimized airfoils.

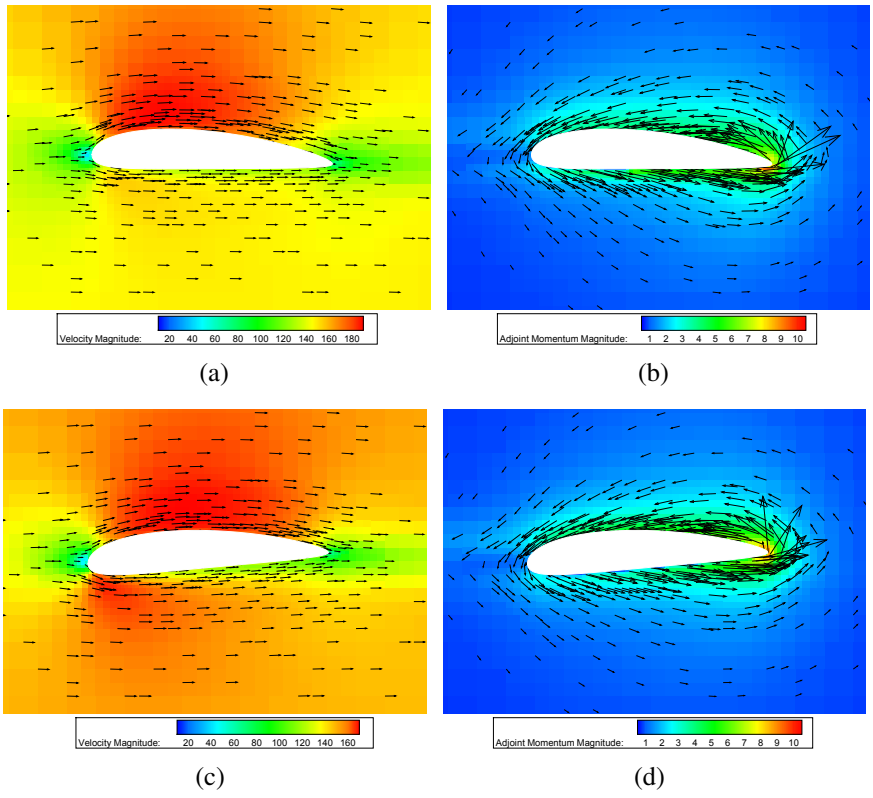


Figure 18: Unsteady subsonic flow around the pitching airfoil parameterized as in fig. 12. Adjoint-based maximization of the time-averaged lift. Flow and adjoint momentum magnitude fields at the two extreme positions of the optimized airfoil's oscillation. Vectors represent the velocity and adjoint momentum for flow and adjoint fields, respectively.

10 Conclusions

The cut-cell method for inviscid flows of compressible fluids and the corresponding continuous adjoint method were developed and programmed on GPUs. These were used in steady and unsteady analysis and optimization problems. Both steady and unsteady analysis results were validated against measurements and CFD results obtained using body-fitted grids. A technique for merging very small cut-cells is employed to avoid numerical instabilities. The Cartesian grid is adapted to the stationary solid wall boundaries and the emerging flow discontinuities. This conforms to the quad-tree data structure, ensuring a significant reduction in memory requirements. The proposed scheme for transferring the flow and adjoint variables between successive time steps, in case refinement or coarsening locally applies, proved to be accurate and capable of supporting unsteady problems. The developed software was used to solve steady and unsteady lift maximization problems in external aerodynamics but this could be extended to other

objective functions with minor modifications.

REFERENCES

[1] C.S. Peskin. Flow patterns around heart valves: a numerical method. *J. Comp. Phys.*, **10** (2): 252-271, 1972.

[2] R. Mittal, G. Iaccarino. Immersed boundary methods. *Annu. Rev. Fluid Mech.*, **37** (1): 239-261, 2005.

[3] D.K. Clarke, H. Hassan, M. Salas. Euler calculations for multielement airfoils using Cartesian grids. *AIAA J.*, **24** (3): 353-358, 1986.

[4] R. Gaffney Jr., H.A. Hassan. Euler calculations for wings using Cartesian grids. *25th AIAA Aerospace Sciences Meeting, American Institute of Aeronautics and Astronautics*, Reston, Virginia, 1987.

[5] H. Ji, F. Lien, E. Yee. A new adaptive mesh refinement data structure with an application to detonation. *J. Comp. Phys.*, **229**: 8981-8993, 2010.

[6] D.I. Papadimitriou, K.C. Giannakoglou. Aerodynamic shape optimization using first and second order adjoint and direct approaches. *Arch. Comput. Methods Eng.*, **15**: 447-488, 2008.

[7] J. Hee Seo, R. Mittal. A sharp-interface immersed boundary method with improved mass conservation and reduced spurious pressure oscillations. *J. Comp. Phys.*, **230** (19): 7347-7363, 2011.

[8] H. Ji, F. Lien, E. Yee. Numerical simulation of detonation using an adaptive Cartesian cut-cell method combined with a cell-merging technique. *Computers & Fluids*, **39**: 10411057, 2010.

[9] P. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comp. Phys.*, **43** (2): 357-372, 1981.

[10] O. Reynolds. Papers on mechanical and physical subjects. *Cambridge University Press*, Vol. 3, 1903.

[11] J. Vassberg, A. Jameson. In pursuit of grid convergence for two-dimensional Euler solutions. *AIAA J. Aircraft*, **47** (4): 1152-1166, 2010.

[12] R. H. Landon. NACA 0012 oscillatory and transient pitching. *AGARD Report 702*, Dataset 3, January 1982.

[13] I.C. Kampolis, X.S. Trompoukis, V.G. Asouti, K.C. Giannakoglou. CFD-based analysis and two-level aerodynamic optimization on graphics processing units. *Comput. Methods Appl. Mech. Engrg.*, **199**: 712-722, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SHAPE OPTIMIZATION OF TURBOMACHINERY ROWS USING A PARAMETRIC BLADE MODELLER AND THE CONTINUOUS ADJOINT METHOD RUNNING ON GPUS

K. T. Tsiakas[1], **F. Gagliardi**[1], **X. S. Trompoukis**[1], **and K. C. Giannakoglou**[1]

[1]National Technical University of Athens, School of Mech. Eng., Lab. of Thermal Turbomachines,
Parallel CFD & Optimization Unit, Athens, Greece
e-mail: {tsiakost,fl.gagliardi,xeftro}@gmail.com ,kgianna@central.ntua.gr

Keywords: Turbomachines, Blade–Row Parameterization, Adjoint Optimization, GPUs.

Abstract. *This paper presents a gradient-based shape optimization method for turbomachinery rows. The developed continuous adjoint method for incompressible flows, with a fully differentiated turbulence model, is coupled with an in-house 3D blade parameterization software, which is differentiated to support the gradient-based optimization process. The in-house flow and adjoint solvers are implemented on a cluster of NVDIA Graphics Processing Units (GPUs). The parameterization software creates the blade by superimposing thickness on both sides of the mean-camber surface. The design variables are NURBS coefficients which ensure smooth shape changes during the optimization, despite the great number of degrees of freedom. In addition, NURBS surfaces are used to describe the final shape. Geometric sensitivities, which stand for the ratio of boundary displacements over the corresponding variation in any of the CAD parameters, are computed by differentiating the parameterization software. Based on the chain rule these are combined with the gradient of the objective function with respect to (w.r.t.) the displacements of the blade or casing nodes, as computed by the adjoint method, and used to update the design variables. During the optimization, the grid is deformed according to the updated shape of the flow domain; regenerating the grid is avoided, by making use of the NURBS surface point inversion technique to retrieve the new surface grid and, then, Radial Basis Functions (RBFs) to propagate the displacement of the boundary nodes to the interior of the computational grid.*

1 INTRODUCTION

In recent years, progress in the development of analysis and optimization tools has made numerical optimization an indispensable component of industrial design processes. In turbomachinery, optimization tools must fully be automated, explore a wide design space and permit the imposition of geometric constraints. Moreover, bringing the geometry back to CAD format and keeping the computational cost as low as possible are both essential.

Recently, gradient–based methods, supported by the continuous or discrete adjoint for computing the objective function gradient, have become very attractive, mainly because the gradient can be computed at a cost independent of the number of design variables. The latter makes the adjoint–based methods suitable for the shape optimization of turbomachinery rows. In CAD–free methods, the objective function sensitivity derivatives w.r.t. the displacement of the nodes lying over the wall boundaries (to be referred to as flow sensitivities), computed by the adjoint solver, are used for modifying the shape of blades and/or casing. In these methods (a) special care must be taken in order to obtain smooth deformed shapes, and (b) the optimized shape must be converted into a CAD format for further processing in the industrial workflow. These issues can be overcome through the use of a parameterization software which is compatible with standard CAD formats. The objective function ($\mathcal{F}$) gradient w.r.t. the design variables $\mathbf{b}$ is computed as

$$\frac{\delta \mathcal{F}}{\delta \mathbf{b}} = \frac{\delta \mathcal{F}}{\delta \mathbf{x}} \frac{\delta \mathbf{x}}{\delta \mathbf{b}} \tag{1}$$

where $\delta\mathcal{F}/\delta\mathbf{x}$ stands for the flow sensitivities computed by the adjoint solver and $\delta\mathbf{x}/\delta\mathbf{b}$ represents the sensitivities of the boundary node positions w.r.t. changes in the design variables (also referred to as geometric sensitivities). In [4], the latter are computed for faceted surfaces by projecting the perturbed geometry onto the original one and applying finite differences. Other approaches rely on the deformation of the surface grids using RBFs followed by a finite difference scheme applied to the resulting and initial grids, [16]. Automatic differentiation of CAD kernels is under investigation by other researchers but only a few basic functions are supported yet.

The adaptation of the initial grid to the updated CAD parameterized boundaries is a crucial component of an automatic optimization loop. This process is often referred as grid morphing. Two main strategies can be employed. The first one exploits the connectivity of the internal grid nodes, applying spring analogy [8] or pseudo–elastic solid techniques[9] while, in the second, each grid node is moved individually by interpolating the surface grid displacement. RBFs [6],[7] are well established tools used in the second approach, due to their robustness and compatibility with any kind of grid connectivity, while preserving elements quality.

In this paper, the optimization of a compressor stator for minimizing viscous losses while penalizing designs which decrease the static pressure rise, is carried out. The blade parameterization software and the computation of geometric sensitivities are presented in Section 2. The continuous adjoint formulation as well as key features of the GPU–enabled flow and adjoint solver are presented in Section 3. Grid morphing is incorporated in the optimization loop (fig. 1) as also described in Section 3. Finally, results of the stator row optimization are discussed in Section 4.

2 BLADE PARAMETERIZATION METHOD

A parameterization/design software for turbomachinery blades is used. The method creates the row geometry by first parameterizing the shape of the blade's mean camber surface and,

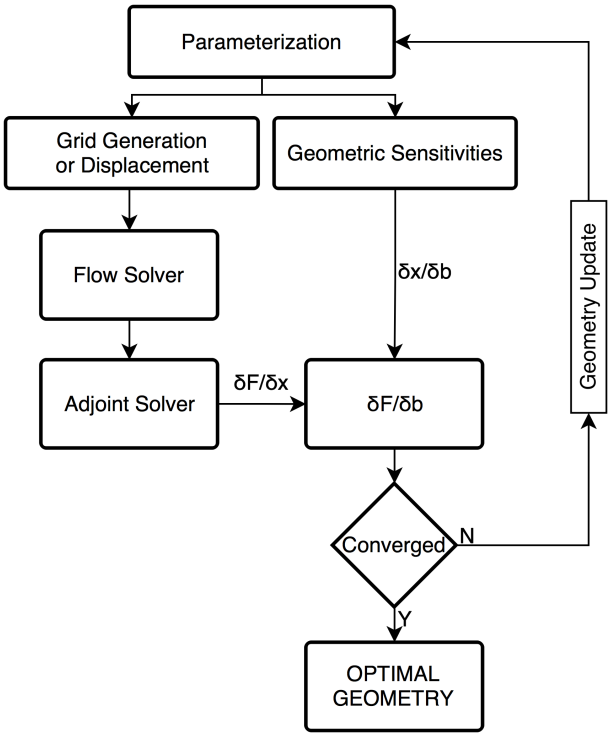


Figure 1: Structure of the overall optimization loop

then, adding thickness.

The majority of data given as input to the parameterization software are Non–Uniform Rational B–Spline (NURBS) curves defining the meridional shape of the row as well as other geometric distributions in the spanwise direction. The NURBS curves provide flexibility to the parameterization method, offering also a compact description for a wide range of blade geometries. In order to parameterize axisymmetric geometries, a conformal mapping [3] of a surface of revolution on the $(m$–meridional, θ–peripheral$)$ plane is employed.

2.1 Outline of the method

The parameterization process and computation of geometric sensitivities $\frac{\delta x}{\delta b_i}$ consists of five basic steps, as follows:

Step 1: Parameterization of the row generatrices and the meridional shape of the LE and TE, Fig. 2.

Step 2: Parameterization of the mean–camber surface of the blade.

Step 3: Superposition of thickness distribution on the mean–camber surface to form the blade shape.

Step 4: Export of the geometry in standard CAD format.

Step 5: Computation of geometric sensitivities.

The first step is the definition of the hub and shroud generatrices via NURBS. In case of a tip clearance between the blade and the hub or shroud, the generatrix representing the meridional shape of the blade tip must also be defined. The meridional shape of the blade LE and TE are

also defined using NURBS curves. Next step is to generate a number of meridional curves, equidistant in the spanwise direction, by interpolating the already defined generatrices for the hub, shroud and blade tip (fig. 2). These curves are used to create surfaces of revolution on which blade sections are defined.

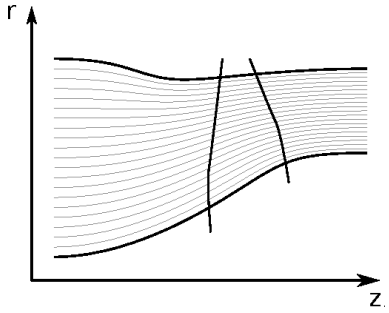


Figure 2: Meridional shape parameterization. Apart from the LE and TE edge meridional projections, the hub and shroud generatrices are shown along with the generatrices of a number of intermediate surfaces of revolution, on which blade sections are defined. Thick lines are defined as NURBS curves, while the thinner ones result from the interpolation of the hub and shroud curves.

After defining the surfaces of revolution for all blade sections, the mean–camber surface is constructed by defining the mean camber line at each section. The mean camber lines are computed on the (m, θ) plane. Since the mapping $(x, y, z) \mapsto (m, \theta)$ is conformal, the blade angles defined on the (m, θ) plane are preserved. The LE and TE of each blade section are defined by their peripheral position (θ_{LE} and θ_{TE}, respectively). Then, the blade metal angles are defined (β_{LE} and β_{TE}, respectively). As shown in fig. 3, the starting and ending–point of the mean camber line ($\mathbf{P}_0$ and $\mathbf{P}_3$, respectively) of each blade section and the point $\tilde{\mathbf{P}}$ where the two tangents intersect are defined. The remaining two control points are located using two non-dimensional weights ζ_{LE} and ζ_{TE}, as follows:

$$\begin{aligned} \mathbf{P_1} &= \zeta_{LE}\mathbf{P_0} + (1-\zeta_{LE})\tilde{\mathbf{P}} \\ \mathbf{P_2} &= \zeta_{TE}\mathbf{P_3} + (1-\zeta_{TE})\tilde{\mathbf{P}} \end{aligned} \tag{2}$$

In order to compactly define the θ, β and ζ for the leading and trailing edge and also achieve a smooth geometry, their distributions in the spanwise direction are specified as NURBS curves.

Thicknesses t_{PS} and t_{SS} for the pressure and the suction side, respectively, might be different. To add flexibility they are defined in two steps. First, the normalized thickness ($\hat{t}$) profiles w.r.t. the normalized arc–length of the mean camber line (s) are defined at some spanwise positions. For any other blade section, ($\widetilde{\hat{t}}$) is interpolated. Then, a spanwise distribution for the thickness factor (t_f), that scales the pre-computed profile, is specified. By doing so, the thickness corresponding to a point at distance s along the mean camber line positioned at a normalized spanwise position (η) is given as $t(\eta, s) = \widetilde{\hat{t}}(s, \eta)\ t_f(\eta)$. Having drawn the mean camber line for each blade section on the (m, θ) plane, the normal vector is computed. The two sides of each airfoil are then formed by defining points in the normal direction at a distance specified by the computed thickness distributions scaled as follows

$$\mathbf{c}_{m\theta}(\eta, s) = \boldsymbol{\mu}_{m\theta}(\eta, s) \pm \mathbf{n}_{m\theta}(\eta, s)\frac{t(\eta, s)}{r^2(m(\eta, s))} \tag{3}$$

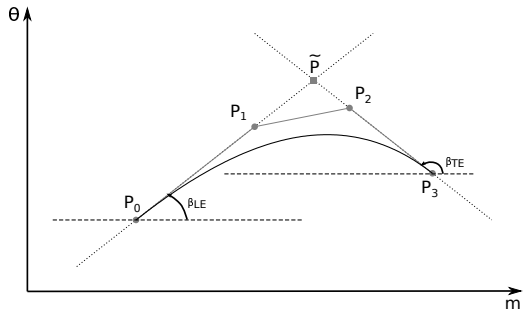


Figure 3: Definition of the mean camber line on the (m, θ) plane. The β_{LE} and β_{TE} angles as well as the curvature of the mean camber line is preserved when mapped back on a surface of revolution, in the 3D Cartesian space.

where $\mathbf{c}$ represents the point on the airfoil in the (m, θ) plane, μ the corresponding point on the mean–camber line and $\mathbf{n}$ the vector normal to the mean–camber line. The resulting blade airfoils on the (m, θ) plane are mapped back onto the 3D Cartesian coordinates to yield the final blade shape.

The parameterization software can export the geometry in neutral CAD format (IGES) for visualization and grid generation purposes. IGES format supports many different entities, among which free–form surfaces in the form of NURBS surfaces that can be used to describe the skin of 3D models.

In the 3D Cartesian space, airfoils are spanwise interpolated through NURBS curves and a skinning algorithm is used to construct the NURBS surface describing the blade; skinning is the process of passing a smooth surface through a set of cross-sectional curves (airfoils). This method requires all cross-sectional NURBS curves to be compatible (equal knot-vector and degree) and can, thus, cause data explosion [5]. Controlling the parameterization of the blade airfoils to have natively compatible cross-sectional curves overcomes this problem.

Hub and shroud surfaces are easily generated as surfaces of revolution based on generatrices and the pitch angle. Similar to the blades themselves, periodic surfaces are generated through skinning. The inlet and outlet boundaries are generated as Coons patches through bilinear blending [15].

Finally, hub and shroud surfaces must be trimmed: in order to avoid numerical instabilities the blade sides are extended. An extension algorithm is applied to the NURBS blade surface [10], as shown in Fig. 4.

Flow sensitivities ($\frac{\delta \mathcal{F}}{\delta \mathbf{x}}$) are computed by the continuous adjoint method for the nodal coordinates of the CFD surface grid. To close the chain rule of eq. 1, geometric sensitivities are needed. Making use of the NURBS representation of the geometry and finite differences, these sensitivities are computed directly on the CFD grid.

Starting from an initial parameterization, the software module perturbs each appropriate design variable and calls the parameterization process, computing the corresponding NURBS surfaces. Surface mesh nodes are inverted onto the NURBS parametric space and, for each surface node, parametric coordinates become available. These parametric coordinates can be used to compute each surface node in the original or any perturbed geometry. Finite differences are applied to the computed nodes. In order to avoid the propagation of numerical errors, intro-

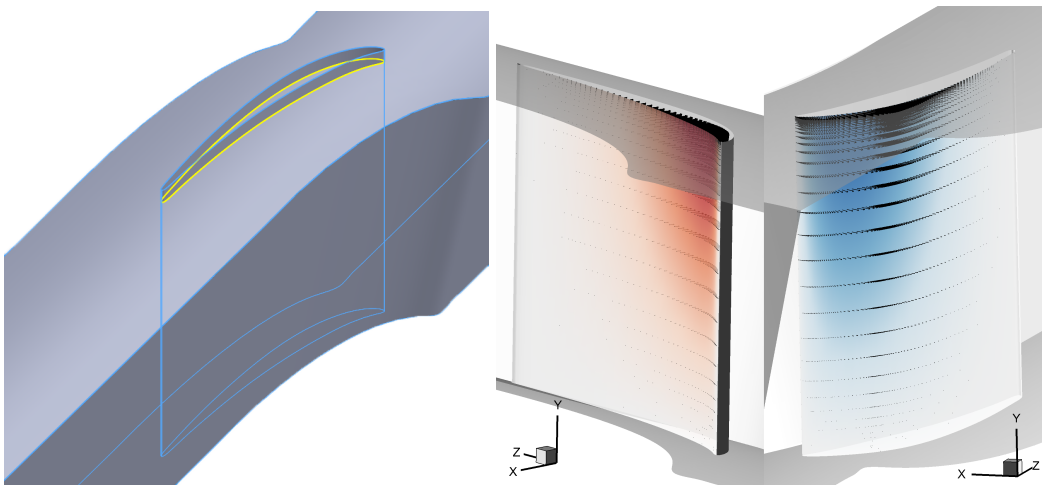


Figure 4: 3D view of the blade extension and the trimming of the blade and shroud surfaces (left). Geometric sensitivities of the blade w.r.t. β_{LE} (fig. 3) at shroud. Contours represent the geometric sensitivities in the normal direction $\frac{\delta \mathbf{x}}{\delta \mathbf{b}}\mathbf{n}$: red denotes inwards (towards the solid) displacement, whereas blue outwards. Arrows represent the actual geometric sensitivities $\frac{\delta \mathbf{x}}{\delta \mathbf{b}}$ (right).

duced by the point inversion algorithm, to the computation of the geometric sensitivities, nodes in the original geometry are recomputed through their parametric coordinates. An application is shown in fig. 4.

3 ADJOINT–BASED SHAPE OPTIMIZATION AND GRID DISPLACEMENT

3.1 Flow Model

The flow model is based on the Navier–Stokes equations for incompressible flows, using the pseudo–compressibility approach introduced by Chorin [1]. By introducing the pseudo–time t the flow equations are expressed as

$$R_n = \frac{\partial U_n}{\partial t} + \frac{\partial f_{nk}^{inv}}{\partial x_k} - \frac{\partial f_{nk}^{vis}}{\partial x_k} = 0 \tag{4}$$

where $n = 1, \ldots, 4$ and $U_n = [p \;\; v_1 \;\; v_2 \;\; v_3]$, with p denoting the static pressure divided by the constant density of the fluid and v_k ($k = 1, \ldots, 3$) the components of the fluid velocity. The inviscid (f_{nk}^{inv}) and viscous (f_{nk}^{vis}) fluxes, in the Cartesian frame, are expressed as

$$f_k^{inv} = \begin{bmatrix} \beta v_k \\ v_k v_1 + p\delta_{1k} \\ v_k v_2 + p\delta_{2k} \\ v_k v_3 + p\delta_{3k} \end{bmatrix}, \quad f_k^{vis} = \begin{bmatrix} 0 \\ \tau_{1k} \\ \tau_{2k} \\ \tau_{3k} \end{bmatrix}, \quad \tau_{mk} = (\nu + \nu_t)\left(\frac{\partial v_k}{\partial x_m} + \frac{\partial v_m}{\partial x_k}\right) \tag{5}$$

where β stands for the pseudo–compressibility parameter, ν and ν_t the kinematic and turbulent viscosity respectively, δ_{ij} is the Kronecker symbol and τ_{km} denotes stresses. The turbulent viscosity is computed using the Spalart–Allmaras turbulence model by solving one additional PDE ($R_{\tilde{\nu}}$) for the turbulence variable $\tilde{\nu}$, as in [2].

3.2 Continuous Adjoint for Aerodynamic Shape Optimization

The first objective is to minimize the volume–averaged total pressure (p_t) losses of the flow through a stationary row. Without any loss in generality, this paper is dealing only with stationary bladings, though the programmed software may also handle rotating bladings by switching to a rotating reference frame. The first function to be minimized is expressed as

$$f_1 = -\int_{S_I} p_t v_k n_k dS - \int_{S_O} p_t v_k n_k dS \tag{6}$$

where S_I is the stator inlet and S_O the stator outlet. The second objective, practically, is used to penalize designs which decrease the static pressure rise. This gives the second objective function (being a negative quantity for compressor blading)

$$f_2 = -\int_{S_I} p v_k n_k dS - \int_{S_O} p v_k n_k dS \tag{7}$$

The same method could also be used for the optimization of rotating rows, with the first objective being the minimization of the relative total pressure p_{tR} between inlet and outlet.

The problem, given by eqs. 6 and 7 can be reformulated as a single one by minimizing ,

$$\mathcal{F} = \omega_1 f_1 + \omega_2 f_2 \tag{8}$$

where ω_1 and ω_2 are appropriate positive user–defined weights. To make the cost of computing $\frac{\delta \mathcal{F}}{\delta b_i}$ independent of the number of design variables, the adjoint method is employed. For a function $\mathcal{I}$ (which stands for either f_1 or f_2), the adjoint formulation starts by defining the augmented function

$$\mathcal{I}_{aug} = \mathcal{I} + \int_\Omega \Psi_n R_n d\Omega + \int_\Omega \tilde{\nu}_a R_{\tilde{\nu}} d\Omega \tag{9}$$

where Ψ_n are the adjoint mean flow variable fields and $\tilde{\nu}_a$ the adjoint to turbulence variable $\tilde{\nu}$. By differentiating eq.9 and applying the Leibniz theorem, we have

$$\frac{\delta \mathcal{I}_{aug}}{\delta b_i} = \frac{\delta \mathcal{I}}{\delta b_i} + \int_\Omega \Psi_n \frac{\partial R_n}{\partial b_i} d\Omega + \int_\Omega \tilde{\nu}_a \frac{\partial R_{\tilde{\nu}}}{\partial b_i} d\Omega + \int_S \Psi_n R_n \frac{\delta x_k}{\delta b_i} n_k dS + \int_S \tilde{\nu}_a R_{\tilde{\nu}} \frac{\delta x_k}{\delta b_i} n_k dS \tag{10}$$

Using the Green–Gauss theorem, the volume integrals of eq. 10 are transformed into (a) volume integrals containing $\frac{\partial U_n}{\partial b_i}$ and $\frac{\partial \tilde{\nu}}{\partial b_i}$, ($b$) surface integrals containing $\frac{\delta U_n}{\delta b_i}$ and $\frac{\delta \tilde{\nu}}{\delta b_i}$ and (c) surface integrals containing $\frac{\delta x_k}{\delta b_i}$. Similarly, $\frac{\delta \mathcal{I}}{\delta b_i}$ results in surface integrals containing $\frac{\delta U_n}{\delta b_i}$ and $\frac{\delta \tilde{\nu}}{\delta b_i}$.

By eliminating all integrals containing $\frac{\partial U_n}{\partial b_i}$, [18], the adjoint to eqs. 4, are derived. These are

$$\begin{aligned} &\frac{\partial \Psi_1}{\partial t} \quad -\frac{\partial \Psi_{k+1}}{\partial x_k} = 0 \\ &\frac{\partial \Psi_{m+1}}{\partial t} \quad -v_k \left(\frac{\partial \Psi_{k+1}}{\partial x_m} + \frac{\partial \Psi_{m+1}}{\partial xk} \right) - \frac{\partial \tau^{adj}_{mk}}{\partial xk} - \tilde{\nu} \frac{\partial \tilde{\nu}_a}{\partial x_m} + \mathcal{T}^{TM}_m = 0, \quad m = 1, \ldots, 3 \end{aligned} \tag{11}$$

where the adjoint stresses are given as $\tau^{adj}_{km} = (\nu + \nu_t) \left(\frac{\partial \Psi_{m+1}}{\partial xk} + \frac{\partial \Psi_{k+1}}{\partial x_m} \right)$ and $\mathcal{T}^{TM}$ stands for terms resulting from the differentiation of the turbulence model. Similarly, eliminating volume integrals containing $\frac{\partial \tilde{\nu}}{\partial b_i}$ results to the adjoint to the turbulence model equation, [17]. Eliminating the corresponding surface integrals gives rise to the adjoint boundary conditions: the remaining

terms give the final expression of the gradient. The expression of the gradient of $\mathcal{I}$ w.r.t. the coordinates of all surface nodes is

$$\frac{\delta \mathcal{I}}{\delta x_k} = -\int_{S_w} \beta \Psi_1 n_m \frac{\partial v_m}{\partial x_\ell} \frac{\delta x_\ell}{\delta x_k} dS + \int_{S_w} \Psi_{m+1} n_q \frac{\partial \tau_{qm}}{\partial x_\ell} \frac{\delta x_\ell}{\delta x_k} dS - \int_{S_w} \tau_{mq}^{adj} n_q \frac{\partial v_m}{\partial x_\ell} \frac{\delta x_\ell}{\delta x_k} dS \qquad (12)$$

Since the inlet and outlet boundaries remain unchanged during the optimization, $\frac{\delta x_k}{\delta b_i} = 0$ along S_I and S_O.

3.3 Numerical Solution of the Flow and Adjoint Equations on GPUs

The primal and adjoint solvers use the vertex-centered variant of the finite volume method on unstructured/hybrid meshes. Both solvers are implemented on NVIDIA GPUs using CUDA-C, taking advantage of the great parallel speed–up they offer. Some comments on the discretization and numerical solution techniques used as well as their implementation on GPUs follow.

The primal inviscid fluxes are computed using the Roe's approximate Riemann solver [11] adapted to incompressible flows. Adjoint fluxes are computed in a similar way, by considering the non-conservative form of the field adjoint equations. The spatial discretization of the inviscid fluxes is second–order accurate, with appropriate flux limiters, if necessary. Both primal and adjoint viscous fluxes are computed using an edge–based central difference scheme. The discretized equations are solved in each pseudo-time step using a point–implicit Jacobi iterative scheme.

Modern GPUs are massively parallel co-processors to CPUs offering at least an order of magnitude more FLOPS and higher memory bandwidth than CPUs. However, due to the different architecture, software directly ported from CPUs to GPUs cannot efficiently exploit the available GPU computational resources. The authors group has developed a flow and adjoint solver that is capable of running on many GPUs belonging to the same or different computational nodes with a parallel speed–up of up to 45x compared to the CPU implementation of the same code. This parallel speed–up results from *a*) the optimization of memory access patterns for the used GPU architecture [14], *b*) the use of GPU specific scatter–adding techniques [13] *c*) the minimization of the communication overhead when using many GPUs, by simultaneous data transfers and computations.

Another important feature of the flow and adjoint solver is the use of Mixed–Precision Arithmetics(MPA); the interested reader could find more about MPA in [12]. In order to reduce memory footprint and number of memory accesses, in MPA, the LHS terms are computed in Double–Precision Arithmetics (DPA) and stored using Single–Precision Arithmetics (SPA). Of course, the RHS terms are still stored using double–precision, so as to maintain the accuracy of a purely DPA scheme.

3.4 Grid Displacement using RBF

With the RBF model the deformation is treated as a scattered data interpolation where surface node displacements are smoothly interpolated at the internal nodes. The new surface grid, which corresponds to the new geometry, is obtained by inverting [15] and displacing nodes in the NURBS parametric space (u, v). Special care is taken in case of trimmed surfaces. An application is shown in fig. 5.

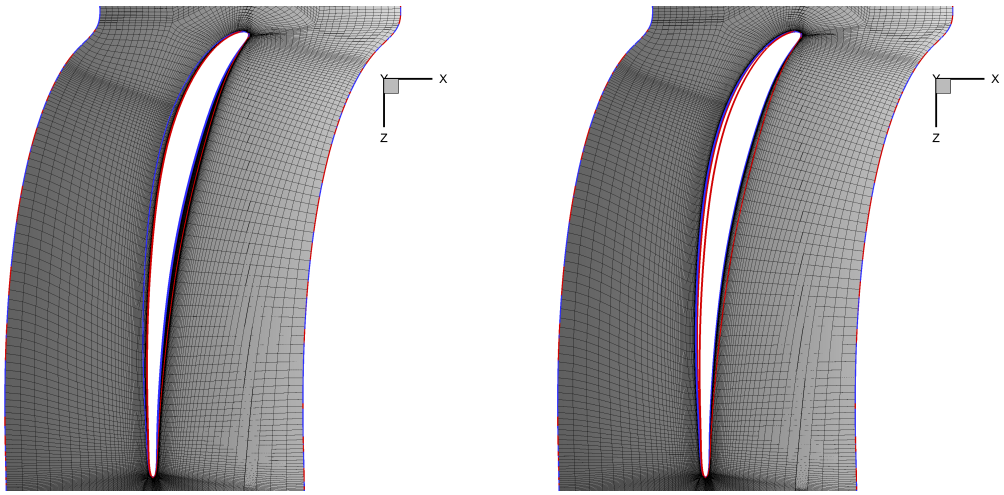


Figure 5: Example of the use of the RBF-based grid displacement method showing the initial (left) and displaced at cycle 4 (right) shroud surface mesh. All boundary nodes are displaced into the NURBS parametric space and, then, transformed back to the Cartesian space.

4 APPLICATION TO A LOW–SPEED COMPRESSOR STATOR

The test case concerns the optimization of the stationary row of a low–speed compressor. The objective is to minimize the volume averaged total pressure losses between the inlet and outlet of the CFD domain penalizing designs which decrease the volume averaged static pressure rise. The weights used in the objective function 8 are $\omega_1 = 0.9$ and $\omega_2 = 0.1$. The inlet flow angle is 46° w.r.t. the axial direction and $Re = 6.5{\times}10^5$ based on the blade axial chord–length (c_{ax}). The initial blade is formed by the spanwise extrusion of the same airfoil and trimming with the cylindrical hub and shroud.

The total number of design variables used to carry out the optimization is 40; note that 32 out of the 40 variables are used to control the blade thickness distribution. In fig. 6, the initial distribution of the thickness is reported along with the NURBS control polygon; control points are allowed to vary only in one direction. In order to maintain $C0$ continuity, at the trailing and leading edge, the first and last control point are kept fixed.

The metal angles β_{LE} and β_{TE} and parameters ζ_{LE} and ζ_{TE} are used to control the blade camber at the hub and shroud positions. Only two span positions are used to describe the distribution of these geometric characteristics. A graphical representation is shown in fig. 6, together with a 3D view of the stator to be optimized. The hub and shroud generatrices remain frozen.

The optimization run on a single computation node with two 6-core Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz processors and two NVIDIA Tesla K40 GPUs with 12GB of GPU memory. The computational grid consists of ~$1.9{\times}10^6$ nodes and is carefully stretched close to the solid walls, where the non-dimensional distance of the first grid nodes off the wall is $y^+{\approx}1$. The solution of the flow field runs for ~$17min$ while ~$12min$ are needed for the solution of the adjoint equations per optimization cycle.

The optimization history as well as both the initial and optimized shapes, are shown in fig. 7.

To compare the flow fields around the initial and optimized geometries 4 stations/positions in the axial direction are chosen as shown in fig. 8. In figs. 9 to 10, the flow fields around the initial and the optimized blade geometries are compared.

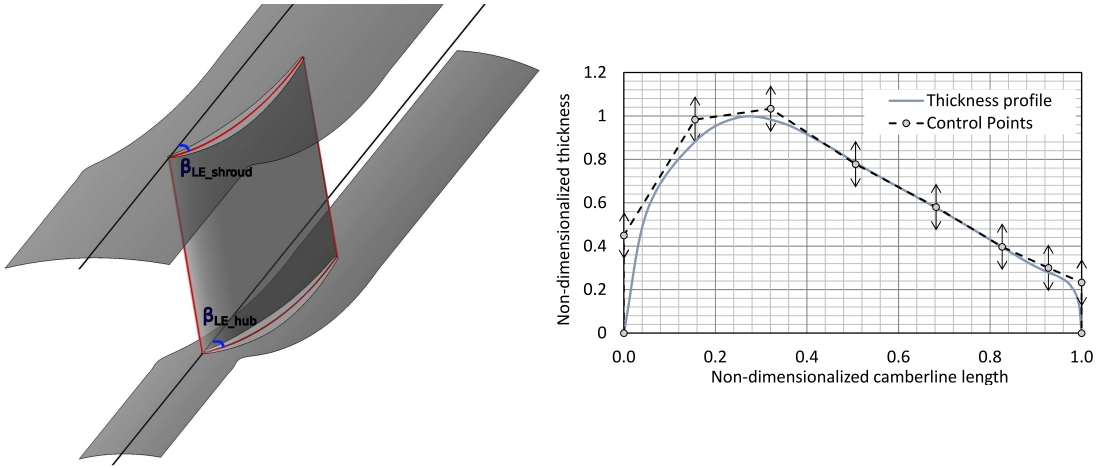


Figure 6: Parameterization of the β angles at LE and TE for hub and shroud position (left). The normalized thickness profile distributions are defined as NURBS curves (right). The control points of the NURBS curves are used as design variables: displacements in the y direction (defining the non–dimensionalized thickness) are safely allowed.

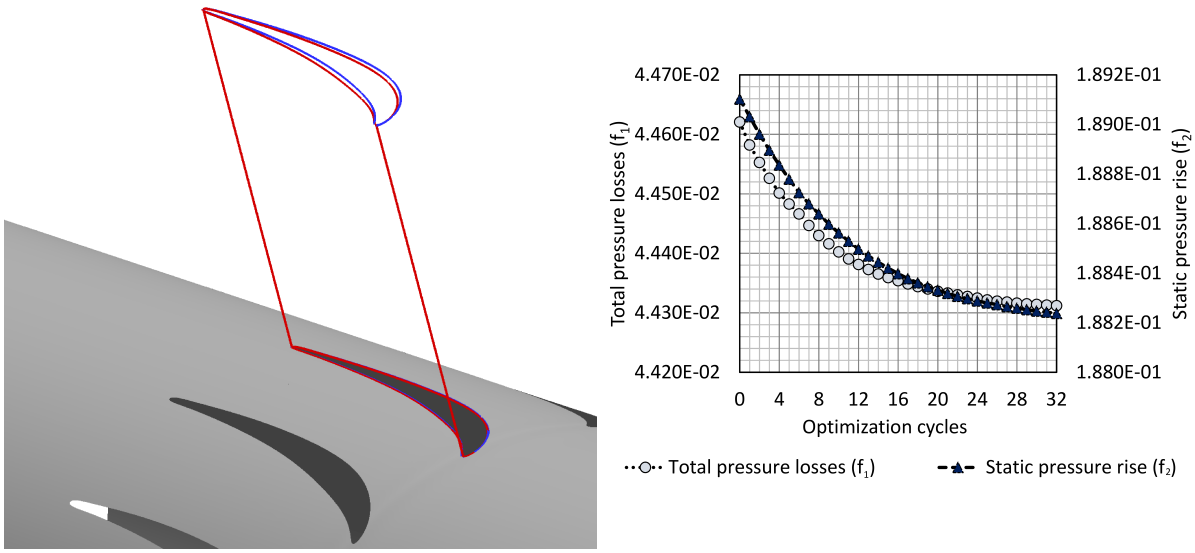


Figure 7: Comparison of the initial (blue) and the optimized (red) blade shape (left). Convergence history for f_1 and f_2 during the optimization (right). Both absolute convergence history (right–top) and relative convergence history (right–bottom) are reported. Both p_t and p are non–dimensionalized by $(V^{in})^2$.

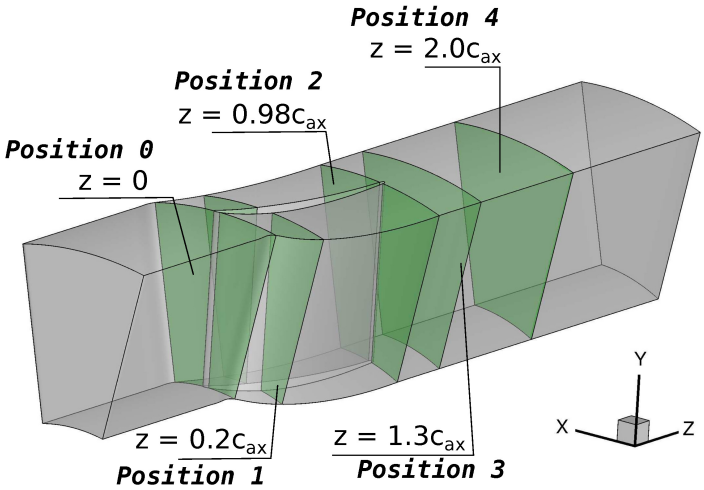


Figure 8: Axial positions chosen to compare the flow fields around the initial and optimized geometry.

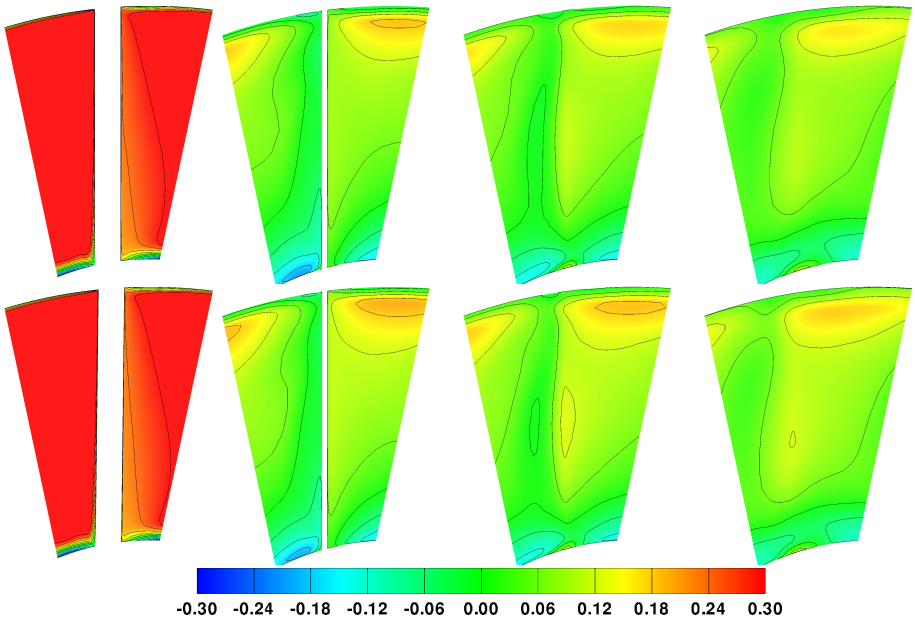


Figure 9: Peripheral velocity component field at axial positions from 1 (left) to 4 (right) for the initial (top) and the optimized (bottom) geometry. The peripheral velocity component is non-dimensionalized by V^{in}.

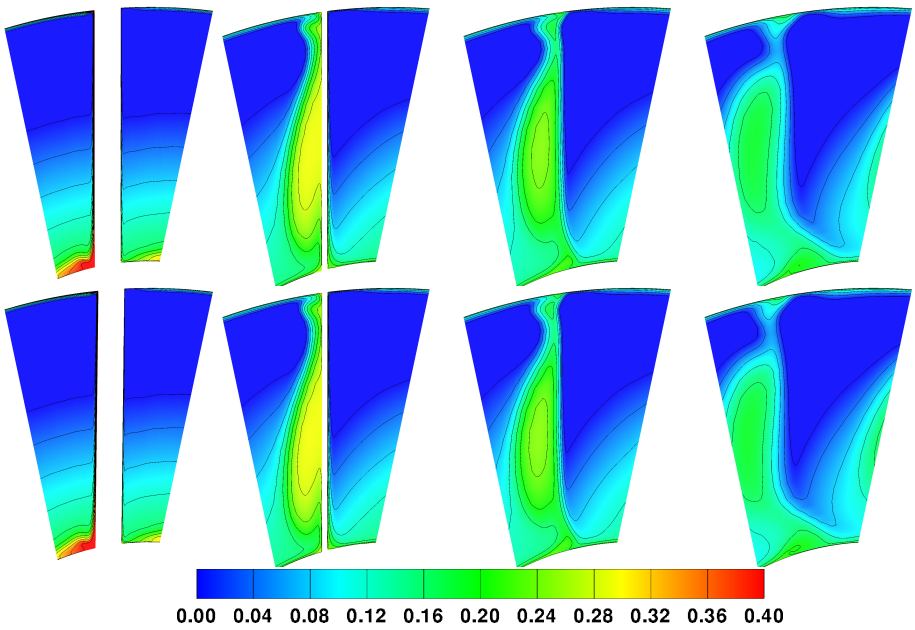


Figure 10: Loss coefficient ($p_t^{in} - p_t$) field at axial positions from 1 (left) to 4 (right) for the initial (top) and the optimized (bottom) geometry. The total pressure is non-dimensionalized by $(V^{in})^2$.

5 CONCLUSIONS

A method for the optimization of turbomachinery rows, by coupling the adjoint–based optimization process with a parametric blade modeller was presented. This enables the exploration of the design space while maintaining a CAD representation. The use of the (continuous) adjoint method makes the cost of computing the objective function gradient independent of the number of design variables. Moreover, the optimization wall–clock time is further reduced by implementing the flow and adjoint solvers on GPUs, speeding–up the corresponding processes by ~45x, compared to the use of CPUs. Regenerating the CFD grid for each new geometry was overcome by using a grid displacement technique based on RBFs. The reduced cost, combined with the capability of linking to CAD, makes the proposed method appropriate for the industrial design workflow.

6 ACKNOWLEDGEMENT

This research was funded from the People Programme (ITN Marie Curie Actions) of the European Union's H2020 Framework Programme (MSCA-ITN-2014-ETN) under REA Grant Agreement no. 642959 (IODA project). The second author is an IODA Early Stage Researcher.

REFERENCES

[1] A. Chorin, A numerical method for solving incompressible viscous flow problems. *Journal of Computational Physics,* ***2-1***, 12–26, 1967.

[2] P. Spalart, S. Allmaras, A one–equation turbulence model for aerodynamic flows. *La Recherche Aérospatiale*, **1**, 5–21, 1994.

[3] M. Rossgatterer, B. Jüttler, M. Kapl, G. Della Vecchia, Medial design of blades for hydroelectric turbines and ship propellers. *Computers & Graphics*, **36**, 434–444, 2012.

[4] T.T. Robinson, C. G. Armstrong, H. S. Chua, Carsten Othmer, T. Grahs, Optimizing parameterized CAD geometries using sensitivities based on adjoint functions. *Computer-Aided Design & Applications*, **9(3)**, 363–268, 2012.

[5] L. Piegl, W. Tiller, Algorithm for approximate NURBS skinning. *Computer-Aided Design*, **9(28)**, 699–706, 1996.

[6] M. Buhmann, Radial basis functions: theory and implementations. *Cambridge University Press, Cambridge*, 2003.

[7] D. Sieger, S. Menzel, M. Botsch, High quality mesh morphing using triharmonic radial basis functions. *Proceedings of the 21st International Meshing Roundtable*, 1–15, 2013.

[8] S. Zhu, A semi-torsional spring analogy model for updating unstructured meshes in 3D moving domains. *Finite Elements in Analysis and Design*, **41**, 1118–1139, 2005.

[9] Z. Xu, M. Accorsi, Finite element mesh update methods for fluid structure interaction simulations. *Finite Elements in Analysis and Design*, **40**, 1259–1269, 2004.

[10] S.M. Hu, C.L. Tai, S.H. Zhang, An extension algorithm for B-splines by curve unclamping. *Computer Aided Design*, **34**, 415–419, 2002.

[11] P. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, **43(2)**, 357–372, 1981.

[12] I. C. Kampolis, X. S. Trompoukis, V. G. Asouti, K. C. Giannakoglou, CFD–based analysis and two–level aerodynamic optimization on Graphics Processing Units. *Computer Methods in Applied Mechanics and Engineering*, **199(9-12)**, 712–722, 2010.

[13] V. G. Asouti, X. S. Trompoukis, I. C. Kampolis, K. C. Giannakoglou, Unsteady CFD computations using vertex–centered finite volumes for unstructured grids on Graphics Processing Units. *International Journal for Numerical Methods in Fluids*, **67(2)**, 232–246, 2011.

[14] X. S. Trompoukis, V. G. Asouti, I. C. Kampolis, K. C. Giannakoglou, CUDA implementation of vertex–centered, finite volume CFD methods on unstructured grids with flow control applications. *GPU Computing Gems, Jade Edition*, 2011.

[15] L. Piegl, W. Tiller, The NURBS book. *Springer*, 2013.

[16] A. Ronzheimer, Aircraft geometry parameterization with high-end CAD-software for design optimization. *Proceedings ECCOMAS 2012, Vienna, Austria*, 2012.

[17] E. M. Papoutsis-Kiachagias, K. C. Giannakoglou, Continuous adjoint methods for turbulent flows, applied to shape and topology optimization: Industrial applications. *Archives of Computational Methods in Engineering*. *Springer*, 2016.

[18] K. T. Tsiakas, X. S. Trompoukis, V. G. Asouti, K. C. Giannakoglou, Shape optimization of wind turbine blades using the continous adjoint method and volumetric NURBS on a GPU cluster. *Proceedings of EUROGEN 2015, Glasgow, UK*, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

GEOMETRIC CONTINUITY CONSTRAINTS FOR ADJACENT NURBS PATCHES IN SHAPE OPTIMISATION

Xingchen Zhang, Yang Wang, Mateusz Gugala and Jens-Dominik Müller

Queen Mary University of London
Mile End Road, E1 4NS, UK
e-mail: {xingchen.zhang, yang.wang, m.gugala, j.mueller}@qmul.ac.uk

Keywords: shape optimisation, adjoint methods, automatic differentiation, geometrical continuity, constraints, NURBS, CAD.

Abstract. *The automatic shape parametrisation method 'NURBS-based parametrisation with complex constraints (NSPCC)' which uses the control points of multiple B-spline patches as design variables, is extended here to include Non-Uniform Rational B-Splines (NURBS) patches. The necessary modifications to the continuity constraints across patch interfaces are presented and the number of required test points to evaluate the constraints is estimated. Numerical tests to demonstrate the developments are presented.*

1 INTRODUCTION

In shape optimisation problems, several important issues should be considered carefully. A crucial aspect is the choice of parametrisation method as it determines the design space. A wide range of parametrisations have been presented in the literature [1].

Node-based methods use the displacement of the nodes of the surface grid provide the richest design space the CFD surface mesh can express. This space is richer than the CFD discretisation can typically see as high-frequency oscillations can be expressed which are not adequately resolved by the CFD. This is addressed by regularisation or smoothing of the gradients or displacements [2, 3]. Lattice-based methods where suitable shape modes are defined on skeleton grids such as Hicks-Henne bumps for aerofoils [4] or stacks of B-spline curves to define turbomachinery blade shapes [5] are popular in aeronautical design. While they can give appropriate design freedom, they are typically limited to topologically rectangular planforms. Free-form deformations (FFD) with volume splines [6] or Radial Basis Function (RBF) [7] are often used for more general design approaches but are cumbersome to define if interfaces between design (deformable) and fixed surfaces need to be respected.

All of the aforementioned methods suffer from the fact that the optimal shape is produced as a deformed mesh, further processing or manufacturing will then require a manual step to capture the relevant modes in a CAD model. This step will typically incur approximations and impair the optimal shape.

At the other end of the spectrum one could consider using parametrisation defined in CAD to define the design space. While this produces the optimal design in CAD, it is typically very time-consuming and complex to define a design space that is rich enough to contain the relevant shape modes. Furthermore, commercial CAD systems are black-box and the relevant shape derivatives for gradient-based optimisation need to be approximated using finite-differences [8], resulting in high run-times for rich design spaces and issues with robustness under topological changes. The companion paper [9] presents work of the group to overcome some of these issues by applying exact algorithmic differentiation to an open-source CAD system.

As an alternative, we here pursue a vendor-neutral CAD approach which does not rely on a parametrisation defined in a CAD system. Using an initial net of surface patches from a CAD Boundary Representation (BRep), the NSPCC approach [10] derives a rich parametrisation in an automatic fashion by considering all control points of these patches as design variables. The unique feature of NSPCC is that it is not limited to single patches or clamped interfaces between patches, but formulates constraints across patch interfaces to guarantee a user-selected level of geometric continuity such as water-tightness ($G0$), tangency ($G1$) and curvature ($G2$). The approach exploits the polynomial nature of B-Splines by evaluating the value and derivative of the constraint function at a small number of test points. Using a projected gradient method, the design space is then projected into the null space of the constraint matrix, while the range space is used to recover a non-linear constraint using a normal step. The method has been successfully demonstrated on a range of configurations and is currently extended to include shape constraints such as thickness and radius [11].

This paper further develops the approach of Xu et al. [10] in two aspects, a) the approach will be extended from B-Splines to NURBS b) an approximation for the number of required test points is proposed. Finally, a S-bend air duct test case using NURBS is presented.

The remaining parts of this paper are structured as follows: Sec. 2 provides the background knowledge of this study, Sec. 3 describes the methods used to describe and maintain the geometrical continuity. Sec. 4 presents number the estimation of the number required test points.

Conclusions and future research directions are presented in Sec. 5.

2 Geometry parametrisation for CFD shape optimisation

Optimisation of shapes immersed in fluids using CFD typically uses gradient-based algorithms as they converge to the optimum in many fewer design iterations compared to stochastic or surrogate modelling approaches. The gradients $\frac{\partial J}{\partial \alpha}$ of the objective function J with respect to a design variable α induced in each control volume by a design change is best computed using the adjoint method [12, 13, 14]. Writing the flow equations as

$$R(U(x_V), x_V) = 0,$$

where R is the conservative residual of the flow equations, U is the state and x_V are the volume grid coordinates, the adjoint equations can be written as

$$\mathbf{A}^T v = g$$

where $\mathbf{A}^T$ is the transpose of the system Jacobian $\mathbf{A} = \frac{\partial R}{\partial U}$, $v = \frac{\partial J}{\partial R}$ is the adjoint solution and $g = \frac{\partial J}{\partial U}$ only depends on the objective function J, but not on the design variables α.

The adjoint approach allows to compute the sensitivities of J with respect to all N design variables α in a single computation of comparable cost to the flow,

$$\frac{\mathrm{d}J}{\mathrm{d}\alpha} = \frac{\partial J}{\partial \alpha} + \frac{\partial J}{\partial R}\frac{\partial R}{\partial \alpha} = \frac{\partial J}{\partial \alpha} + v^T f$$

The direct dependence of J on α, e.g. if the J is defined on the design surface, is straightforward to compute and omitted for simplicity in the following.

Using the chain rule of calculus, the differentiation of the remaining term $\frac{\partial J}{\partial R}\frac{\partial R}{\partial \alpha}$ can be separated into contributions from the perturbations in volume grid coordinates x_V, which in turn are linked to perturbations of the surface grid coordinates x_B through a mesh deformation algorithm such as Laplacian smoothing.

$$\frac{\partial J}{\partial R}\frac{\partial R}{\partial \alpha} = \frac{\partial J}{\partial R}\frac{\partial R}{\partial x_V}\frac{\partial x_V}{\partial x_B}\frac{\partial x_B}{\partial \alpha} = v^T \frac{\partial R}{\partial x_V}\frac{\partial x_V}{\partial x_B}\frac{\partial x_B}{\partial \alpha}.$$

The first term on the right hand side $\partial R/\partial x_V$ is computed by differentiating the flux and metrics computation of the flow solver w.r.t. the mesh coordinates. The second term $\partial x_V/\partial x_B$ arises from differentiating the volume mesh smoothing. The final term in the chain rule is the sensitivity of the surface coordinates w.r.t. the changes in design variables, $\partial x_B/\partial \alpha$, which requires a differentiation of the parametrisation.

The NSPCC approach uses a source-code implementation of NURBS patches which can then be differentiated using AD, in our work we employ Automatic Differentiation Software tools (AD) [15]. While the implementation in Xu et al.'s work [10] used the coordinates of the control points $\mathbf{P}_{ij}$ as design variables, with each control point having three degrees of freedom (DoF) to move, here we extend the approach to also include the weights ω_{ij} as design variable, leading to 4 DoF per control point.

2.1 NURBS surface patches

Non-Uniform Rational B-splines (NURBS) are widely used to describe geometries. A NURBS patch is a 3D surface defined as [16]:

$$S(u,v) = \frac{\sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(u)N_{j,q}(v)\omega_{i,j}P_{i,j}}{\sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(u)N_{j,q}(v)\omega_{i,j}} \quad 0 \le u,v \le 1 \tag{1}$$

where $P_{i,j}$ are control points, $\omega_{i,j}$ are weights, $N_{i,p}(u)$ and $N_{j,q}(v)$ are p-th and q-th degree bspline basis functions defined in the following knot vectors:

$$\{\underbrace{0,\ldots,0}_{p+1}, u_{p+1},\ldots,u_i,\ldots,u_{r-p-1}, \underbrace{1,\ldots,1}_{p+1}\}$$
$$\{\underbrace{0,\ldots,0}_{q+1}, v_{q+1},\ldots,v_j,\ldots,v_{s-q-1}, \underbrace{1,\ldots,1}_{q+1}\}$$

where $r = n+p+1$ and $s = m+q+1$. $N_{i,p}(u)$ and $N_{j,q}(v)$ are given by the following expression:

$$N_{i,0}(u) = \begin{cases} 1 & if \quad u_i \le u < u_{i+1} \\ 0 & otherwise \end{cases}$$

$$N_{i,k}(u) = \frac{(u-u_i)}{u_{i+k}-u_i}N_{i,k-1}(u) + \frac{(u_{i+k+1}-u)}{u_{i+k+1}-u_{i+1}}N_{i+1,k-1}(u) \tag{2}$$

Written in homogeneous form, a NURBS surface can be expressed as

$$S^{\omega}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(u)N_{j,q}(v)P^{\omega}_{i,j} \qquad 0 \le u,v \le 1, \tag{3}$$

where $P^{\omega}_{i,j} = (\omega_{i,j}x_{i,j}, \omega_{i,j}y_{i,j}, \omega_{i,j}z_{i,j}, \omega_{i,j})$. Written in this form, most of algorithms formulated for B-spline surfaces can straightforwardly be applied to NURBS. We shall hence omit the superscript ω and use P for the homogeneous control point coordinate in the following.

3 NSPCC algorithm using NURBS surfaces

3.1 Geometric continuity

In shape optimisation problems, the congruence between patches usually requires at least $G0$ (adjacent surfaces touch or intersect to be watertight). Smoothness of the geometry is typically a very important aspect in the CFD design, the changes in tangency and curvature often have a very strong influence on the pressure field. $G2$ continuity (continuous curvature) would be required to not induce pressure spikes, however this seems difficult to implement with standard methods and is often not imposed. While our methodology could also be used to impose $G2$, in this study $G0$ and $G1$ continuity are considered.

In the NSPCC approach the constraint between patches is evaluated numerically at the same position in all touching or intersecting patches. Pairs of test points are evenly distributed along the common edges as shown in Fig. 1. $G0$ continuity can then be written as:

$$G0 = X_{s,L} - X_{s,R} = 0 \tag{4}$$

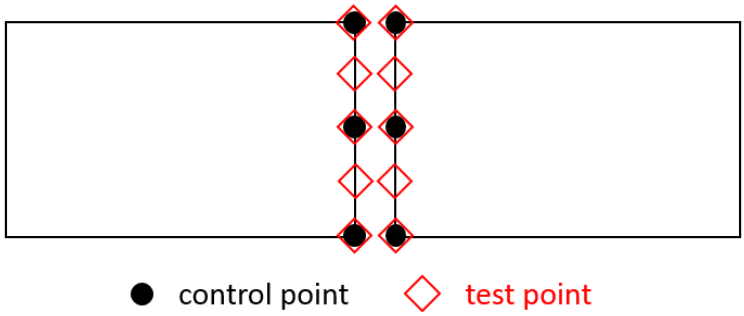


Figure 1: Two patches sharing one common edge.

where $X_{s,L}$ is the test point on the left patch, and $X_{s,R}$ is the test point on the right patch sharing an edge. $G1$ continuity can be written as:

$$G1 = \frac{\mathbf{n}_L \cdot \mathbf{n}_R}{||\mathbf{n}_L||||\mathbf{n}_R||} = 1 \tag{5}$$

where $\mathbf{n}_L$ and $\mathbf{n}_R$ are the unit normal vectors of the tangent plane at the test points either side of the patch interface. The unit normal vector is expressed as:

$$\mathbf{n} = \frac{(\frac{\partial X_s}{\partial u} \times \frac{\partial X_s}{\partial v})}{\left\| \frac{\partial X_s}{\partial u} \times \frac{\partial X_s}{\partial v} \right\|} \tag{6}$$

where u and v are the parametric coordinates of the surface as introduced in equation (1).

3.2 Computation of the constraint matrix using singular value decomposition (SVD)

Linearising the difference in continuity constraint values G between two design iterations n and $n+1$ one obtains

$$G^{n+1} = G^n + \sum_{i=1}^{N} \frac{\partial G}{\partial P_i} \delta P_i, \tag{7}$$

where G comprises the constraints of all required levels of continuity, δP_i is the displacement of the homogeneous coordinate of control point P_i. Requiring $G^{n+1} = G^n$ one obtains a linearised condition for zero change in constraints, hence

$$\sum_{i=1}^{N} \frac{\partial G}{\partial P_i} \delta P_i = \mathbf{C}\delta P = 0. \tag{8}$$

The matrix $\mathbf{C}$ has M_C rows and $4 \times N$ columns in total, corresponding to the total number of M_C constraint equations in the M test points and the N NURBS control points. The elements of the matrix $\mathbf{C}$ are calculated using Automatic Differentiation (AD) of the STEP file interpreter. In the typical situation of all patches having a much higher number of control points in each direction compared to the order of the spline, the majority of the control points do not affect the continuity of the patch borders and the matrix $\mathbf{C}$ is sparse.

It can be seen from equation (8) that the displacements of the control points δP have to lie in the null space of $\mathbf{C}$ to satisfy the continuity constraints in a linearised sense. To obtain the

space of displacement vectors of control points, we need to compute the null space of $\mathbf{C}$. This is performed by means of a SVD from the LAPACK library [17],

$$\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{9}$$

where $\mathbf{U}$ is an $M_C \times M_C$ unitary matrix, Σ is an $M_C \times 4N$ diagonal matrix with non-negative real numbers on the diagonal, and the $4N \times 4N$ unitary matrix $\mathbf{V}^T$ denotes the transpose of $\mathbf{V}$. The rank of the matrix, r, is the number of the non-zero diagonal entries in Σ. The last $(4N - r)$ columns of the matrix $\mathbf{V}$ span the nullspace of $\mathbf{C}$, denoted by $\mathrm{Ker}(\mathbf{C})$.

3.3 Required number of test points

The polynomial of order N_q on the parametric coordinate of the patch edge, along a section of a B-spline curve can be matched uniquely if there are N_q+1 distinct test points within the knot vector interval that supports this section of the curve. In practice it is algorithmically cumbersome to determine the appropriate number as the support intervals overlap.

The number of required test point pairs for a B-spline curve can be determined a-priori by considering each non-zero knot-interval. To match the polynomial of order N_q exactly we need $N_q + 1$ test point pairs in the interval. The relationship between the number of knots N_k, the number of control points N_p and the order of the spline N_q is

$$N_k = N_p + N_q \tag{10}$$

The number of non-zero knot intervals is

$$N_i = (N_k - 1) - 2(N_q - 1) - N_M \tag{11}$$

where N_M is the number of zero knot intervals because of internal multiplicities. From Eq. (10) and Eq. (11), the number of non-zero knot intervals becomes

$$N_i = N_p - N_q - N_M + 1 \tag{12}$$

In each interval we then need N_q+1 test points to fit the polynomial exactly, for the left side of a patch edge with N_p control points, we hence need M_L test points

$$M_L \geq (N_q+1)(N_p - N_q - N_M + 1),$$

similarly for the right side M_R. Assuming a regular spacing of knots, the number of required test points $M_{T,E}$ along edge E then becomes

$$M_{T,E} \geq \max(M_L, M_R). \tag{13}$$

However, imposing too many test points does not pose a problem as the SVD filters out linearly dependent constraint equations. To allow for non-regular knot-intervals we use

$$M_{T,E} \geq f_T \max(M_L, M_R)$$

with an inflation factor f_T chosen around $1.2 \leq f_T \leq 1.5$. In the typical case of equal polynomial orders $N_q|_L = N_q|_R$ this becomes

$$M_{T,E} \geq f_T(N_q+1)(\max(N_p|_L, N_p|_R) - N_q - N_M + 1). \tag{14}$$

The exactness argument employed for B-Splines does not carry over straightforwardly to the non-rational functions in NURBS. This could be accommodated by increasing the factor f_T, however in practice we have not found this to be necessary.

Due to finite-precision arithmetic and ill-conditioning of the constraints, the singular values do not drop off to zero sharply after all non-singular modes, but show a rather gradual decrease. In practice the authors have found the NSPCC method applied to B-Splines to be very insensitive against setting the cut-off threshold. The constraint equations G are non-linear and for finite-size steps the design iteration introduces a constraint violation, requiring a recovery step in the range of $\mathbf{C}$. An exact representation of the feasible space is hence not essential.

4 Results

4.1 Discrete incompressible flow and adjoint solver

In this work, our in-house code GPDE [18] is used. GPDE is an incompressible, viscous, steady flow solver on unstructured grids. The spatial discretisation of the flow solver (primal) is based on a finite volume approach with 2nd order schemes. The SIMPLE-type time-discretisation [19] is used to decouple velocity and pressure computation.

A discrete adjoint solver is used to obtain the sensitivity of the objective function w.r.t. changes in surface node coordinates, $\frac{\partial J}{\partial x_S}$. The adjoint solver is derived from the flow solver via automatic differentiation with the AD tool Tapenade [15]. AD is used to produce the relevant derivative source code for fluxes and source terms. The differentiated routines are then assembled in a hand-written driver code to improve performance, exploiting the fixed-point nature of the primal. In the fixed point iteration of GPDE, stabilisation strategies such as CFL-ramping, skewness correction, semi-coupled methods are implemented to stabilise both primal and discrete adjoint solvers [20].

4.2 The number of test points

Two test cases are utilised in this study to investigate the number of test points required. The first one is a half cylinder case as shown in Fig. 2 and the other one is the ONERA M6 case shown in Fig. 3.

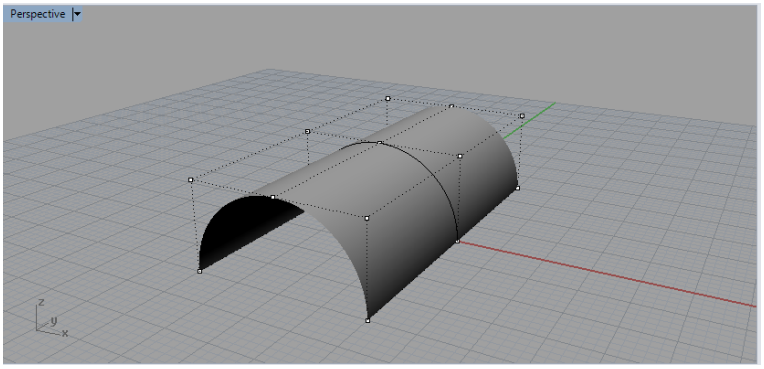


Figure 2: The 3D model of a half cylinder and its control points.

For both cases, all of parameters about the geometry such as the control points, the knot vectors are read from the STEP file. STEP file is a standard file to exchange data among different systems [21]. In addition, a random perturbation to the geometry is given in these two cases.

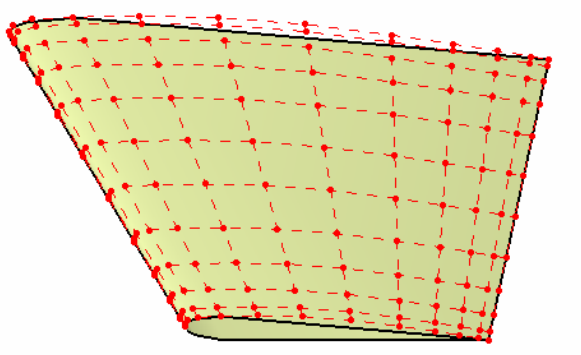


Figure 3: The 3D model of ONERA M6 wing which consists of two patches, and the control points of one patch.

4.2.1 Half cylinder

The half cylinder case used in this study is created using Rhino 5.0 [1] This half cylinder consists of two NURBS surfaces sharing one common edge. The number of control points on each surface along the common edge is 5, and the degree of surface along the common edge is 2, hence order 3. The knot vector along the common edge is

$$\{0,0,0,0.5,0.5,1.0,1.0,1.0\},$$

hence the number of non-zero knot vector interval is 2, matching the estimate of $N_p - N_q - N_M + 1$ in Eq. (12). According to Eq. (14), the number of test points should be:

$$M \geq (3+1)2 = 8.$$

Fig. 4 shows the number of singular values found vs. the number of test points. To useful numerical precision, the number of non-singular values no longer increases beyond 10 test points., which means that 10 test points are enough to ensure the continuity constraints. This matches well with the result of equation (14).

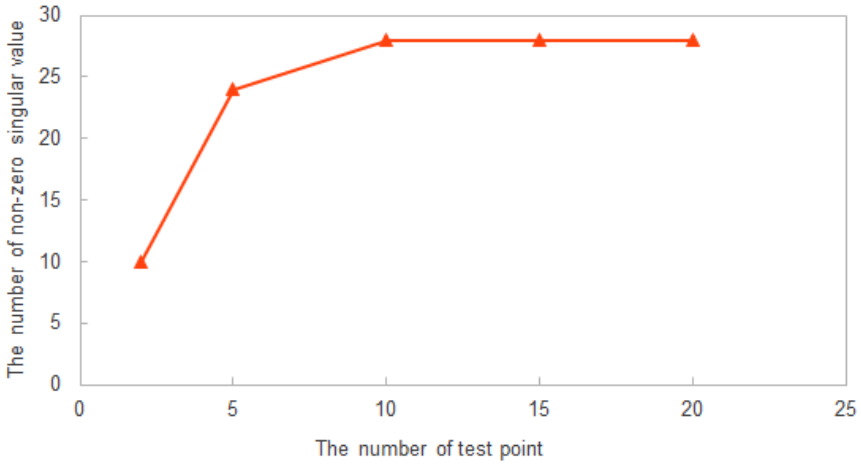


Figure 4: The number of non-zero singular values when the number of test points changes for half cylinder case.

[1] https://www.rhino3d.com/

4.2.2 ONERA M6 wing

The M6 wing case used in this study is created using CATIA V5. It consists of two B-spline patches and there are two common edges as shown in Fig. 3. In this study, B-splines are treated as NURBS by setting weight 1.0 to every control point. There are 12 control points along each common edge and the degree of surface along the common edge is 5, hence order 6. The knot vector along the common edge is

$$\{0,0,0,0,0,0,0.12,0.27,0.41,0.56,0.70,0.85,1.0,1.0,1.0,1.0,1.0,1.0\}\ ,$$

hence the number of non-zero knot vector intervals is 7, matching the estimate of $N_p - N_q - N_M + 1$ in Eq. (12).

According to equation (14), the number of test points should be:

$$M \geq (6+1)7 = 49.$$

In this study, the relationship between the number of non-zero singular value and the number of test points is shown in Fig. 5. It indicates that the number of singular values keep constant after the number of test points reaches 60 which means that 60 test points are enough.

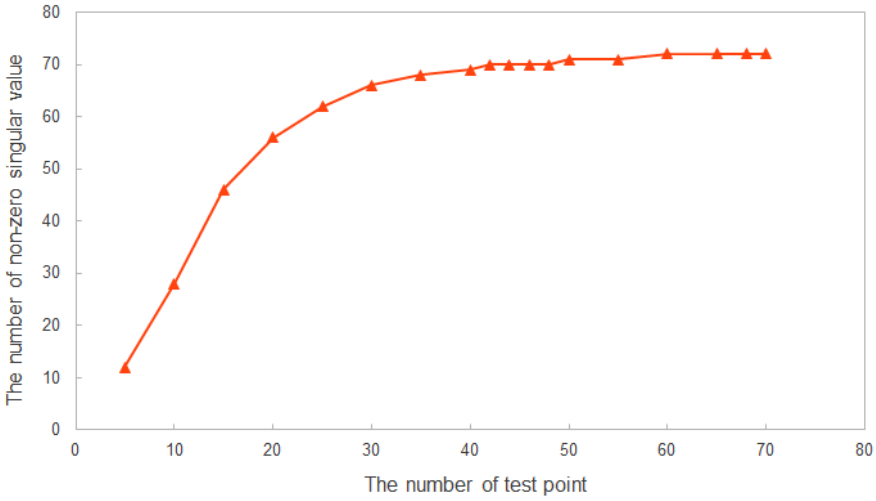


Figure 5: The number of non-zero singular values when the number of test points changes for M6 wing case.

4.3 Optimisation results

In this paper, we apply the NURBS-based parametrisation method to a S-bend air duct case. The original shape of this S-bend is shown in Fig. 6. There are 30 surfaces in this geometry, only the 8 surfaces in the middle section are design surfaces. 400 control points on these surfaces are allowed to move and change weights, thus there are 1600 DoF in total. $G1$ continuity is imposed across the interfaces of these 8 surfaces.

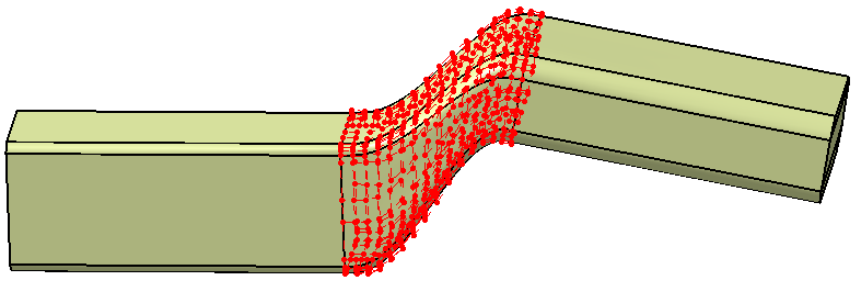


Figure 6: The S-bend air duct and the control points of middle section.

The cost function is the mass-averaged total pressure loss between inlet and outlet defined as

$$J = \frac{\int_{inlet} p_{total}(\mathbf{u}\cdot\mathbf{n})d\mathbf{S} - \int_{outlet} p_{total}(\mathbf{u}\cdot\mathbf{n})d\mathbf{S}}{\int_{inlet}(\mathbf{u}\cdot\mathbf{n})d\mathbf{S}} \tag{15}$$

The flow solver GPDE is used with steepest descent with fixed step width as optimiser. The cost function and gradient history during the 62 iterations are shown in Fig. 7 and Fig. 8, respectively. From Fig. 7 it can be seen that, the cost function is reduced by about 9.4% during the optimisation.

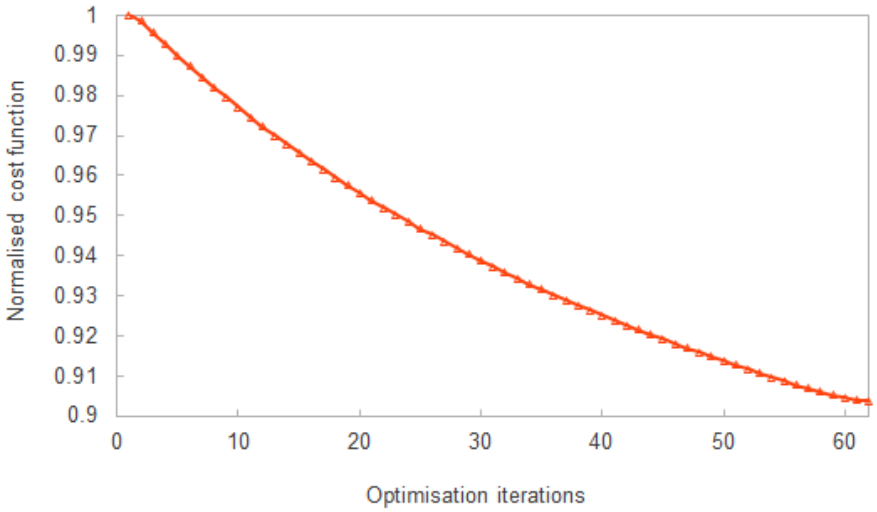


Figure 7: The cost function history.

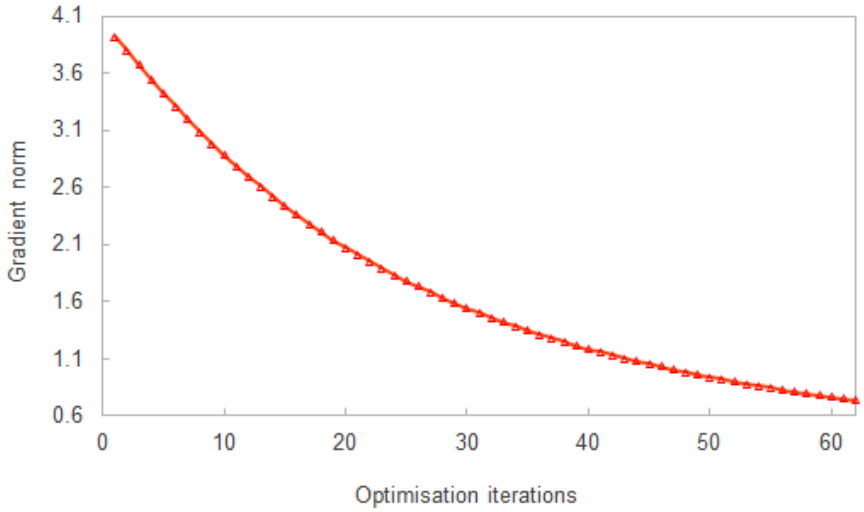


Figure 8: The gradient history.

The optimised shape is shown in Fig. 9. Checks with the CAD software, confirm that the $G1$ continuity across the interfaces in the middle section is satisfied. In addition, the maximum $G1$ deviation value during the optimisation is 2.3×10^{-10}. This means after the extension to NURBS, the geometric continuity can be ensured across different patches.

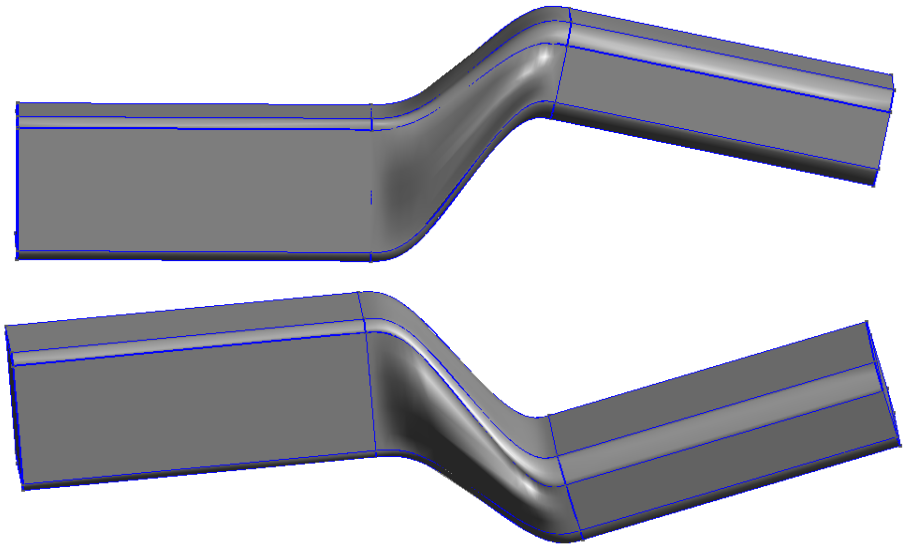


Figure 9: The optimised S-bend air duct.

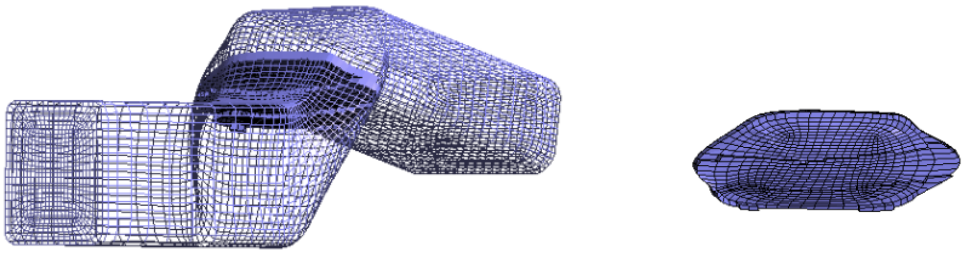


Figure 10: The cross section of optimised S-bend air duct.

Fig. 10 shows the cross section of the cranked part of this S-bend. It can be seen that the side surfaces are pushed out after optimisation offering more cross-sectional area to the flow which through reduction the velocity gradients at the wall reduces skin friction and ultimately total pressure loss. A more significant effect is the reduction in secondary flow motion as visible in Fig. 11, showing the streamlines for the initial and deformed shape after 62 optimisation iterations. The optimised shape shows diminished secondary flow, which results in lower total pressure loss.

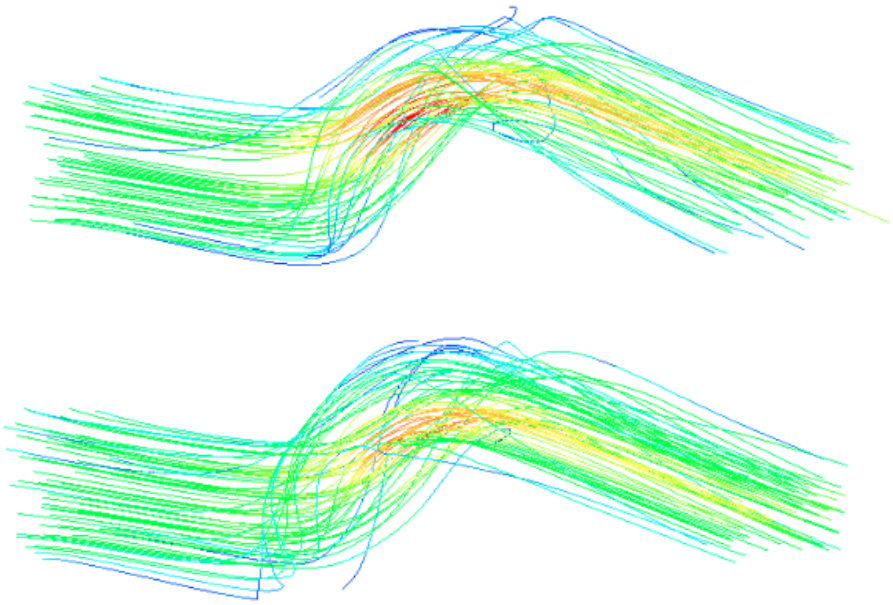


Figure 11: Streamline plots for initial (top) and optimised ducts (bottom).

5 CONCLUSIONS

The NSPCC method of Xu et al. [10] has been extended to include NURBS patches. The test-point approach projecting the design space of control point movements has been extended to include weight changes. An estimate of the required number of test points has been presented, numerical tests confirm the validity of the estimate.

An S-bend air duct optimisation case is presented to demonstrate the proposed method works for shape optimisation problem. The NURBS-based parametrisation using a typical density

of control points is able to develop a very rich but smooth geometry perturbation with $G1$ geometric continuity. Using a simple optimiser the objective function was reduced by over 9%.

6 ACKNOWLEDGEMENTS

The authors would like to thank the funding from China Scholarship Council and Queen Mary, University of London for this research (No. 201306230097). The discussion with colleagues from EC projects About Flow (No. 317006) and IODA (No. 642959) are also very appreciated.

REFERENCES

[1] J.A. Samareh, A survey of shape parameterization techniques. *NASA Conference Publication*, Citeseer, 1999.

[2] A. Jameson, A. Vassberg, Studies of alternative numerical optimization methods applied to the brachistochrone problem. *CFDJ*, **9**(3), 2000.

[3] A. Jaworski and J.-D. Müller, Toward modular multigrid design optimisation. *Lecture Notes in Computational Science and Engineering, Vol. 64*. Springer, 2008.

[4] R.M. Hicks, Wing Design by Numerical Optimization. *Journal of Aircraft*, **15**(7), 407-412, 1978.

[5] S. Shahpar, Challenges to overcome for routine usage of automatic optimisation in the propulsion industry. *The Aeronautical Journal*, **115**(1172), 612-25, 2011.

[6] W. Song, X. Yang, Free-form deformation with weighted T-spline. *The Visual Computer*, **21**(3), 139-151, 2005.

[7] S. Jakobsson, O. Amoignon, Mesh deformation using radial basis functions for gradient-based aerodynamic shape optimization. *Computers & Fluids*, **36**(6), 1119-1136, 2007.

[8] T.T. Robinson, C.G. Armstrong, H.S. Chua, C. Othmer, T. Grahs, Sensitivity-based optimization of parameterised CAD geometries. *8th World Congress on Structural and Multidisciplinary Optimisation*, Lisbon, 2009.

[9] S. Auriemma, M. Banovic, O. Mykhaskiv, H. Legrand, J.D. Müller, A. Walther, Optimisation of a u-bend using CAD-based adjoint method with differentiated CAD kernel. *ECCOMAS Congress 2016*, Crete Island, Greece, 5-10 June 2016.

[10] S. Xu, W. Jahn, J.-D. Müller, CAD-based shape optimisation with CFD using a discrete adjoint. *IJNMF*, **74**(3), 153-68, 2013.

[11] S. Xu, D. Radford, M. Meyer, J.-D. Müller, CAD-based adjoint shape optimisation of a one-stage turbine with geometric constraints. *ASME Turbo Expo 2015*, June, 2015. GT2015-42237.

[12] O. Pironneau, On Optimum Design in Fluid Mechanics. *JFM*, **64**, 97-100, 1974.

[13] A. Jameson, Aerodynamic Design via Control Theory. *JSC*, **3**, 233-260, 1988.

[14] M.B. Giles, N.A. Pierce, An introduction to the adjoint approach to design. *Flow, turbulence and combustion*, **65**(3-4), 393-415, 2000.

[15] L. Hascoët, V. Pascual, TAPENADE 2.1 user's guide. *Technical Report*, INRIA, 2004.

[16] L. Piegl, W. Tiller, *The NURBS Book. 2nd*, New York: Springer, 1997.

[17] E. Anderson, Z. Bai, C. Bischof, et al. LAPACK Users' guide[M], Siam, 1999.

[18] D. Jones, J.D. Müller, F. Christakopoulos, Preparation and assembly of discrete adjoint CFD codes. *Computers & Fluids*, **46**(1). 282-286, 2011.

[19] J.H. Ferziger, M. Peric, Computational methods for fluid dynamics[M], Springer Science & Business Media, 2012.

[20] Y. Wang, S. Akbarzadeh, J.D. Müller, Stabilisation of Discrete Adjoint Solvers for Incompressible Flow. *22nd AIAA Computational Fluid Dynamics Conference*, 2749, 2015

[21] M.J. Pratt, Introduction to ISO 10303 - the STEP Standard for Product Data Exchange. *Journal of Computing and Information Science in Engineering*, **1**(1), 102-103, 2001.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOWARDS MULTIDISCIPLINARY ADJOINT OPTIMIZATION OF TURBOMACHINERY COMPONENTS

Marc Schwalbach[1] **and Tom Verstraete**[2]

[1]von Karman Institute for Fluid Dynamics
72, Chaussee de Waterloo, B1640 Rhode Saint Genese, Belgium
e-mail: marc.schwalbach@vki.ac.be

[2] Queen Mary University Of London
E1 4NS Mile End Road, London, UK
e-mail: t.verstraete@qmul.ac.uk

Keywords: adjoints, algorithmic differentiation, turbomachinery, CAD, multidisciplinary optimization, fluid-structure interaction

Abstract. *The current state-of-the-art adjoint design optimizations for turbomachinery components focus solely on aerodynamic cost functions and constraints, yet disregard structural feasibility during the optimization procedure. This paper presents the first steps taken towards including structural constraints in a multidisiplinary adjoint optimization design chain for turbomachinery components. Particularly in turbomachinery, deformations arise due to centrifugal and pressure load during running conditions, which lead to a coupled fluid-structure interaction problem. While most optimization methods treat the fluid and structure domains separately in a single-disciplinary fashion, we seek to directly include the coupled fluid-structure interaction within the adjoint optimization. To this end, a cold-to-hot transformation tool that deforms a CAD geometry based on FEM displacements is implemented and differentiated using adjoint algorithmic differentiation to compute the required transformation sensitivities.*

1 INTRODUCTION

Current adjoint optimization methods for turbomachinery components focus mainly on aerodynamic cost functions and constraints , e.g. [3], [10], [18]. The structural feasibility of the resulting optimized shape is usually tested a posteriori by conducting a stress analysis. As a result, an aerodynamically optimized shape may exceed defined stress tolerance levels and a new shape will have to be designed. This can lead to several costly design iterations. Multidisciplinary adjoint optimization seeks to take structural constraints into consideration during the optimization process by coupling the fluid and structural disciplines. This will avoid unnecessary design iterations and directly compute an aerodynamically optimized shape, which is also structurally feasible.

Turbomachinery components pose an additional challenge due to the deformations which occur at running conditions. These are caused by rotation-induced centrifugal forces, gas pressure loads, as well as thermal expansions due to temperature changes. The deformation process is known as the *cold-to-hot* transformation [5], which can be seen as a fluid-structure interaction problem. The *cold* state defines the at-rest state at which a component is manufactured. The *hot* state defines the deformed state of a component during running conditions. Typically, these deformations are not taken into account during the optimization process itself, but rather a shape is optimized in its hot state and transformed a posteriori to its cold shape for manufacturing. In this case, the parametrized CAD geometry used in the optimization, also named the master CAD geometry, is defined in the hot state [6]. However, this method has the downside that the master CAD geometry used in the optimization was generated for one specific design point. Different design points, each with their own centrifugal and pressure loads, would result in different deformed hot geometries. Nevertheless, often the same master CAD geometry is used to compute other design points. Additionally, the cold geometry that is generated after the optimization may no longer fulfill manufacturing constraints, e.g. flank milling of radial machines requires ruled surfaces.

As opposed to a hot state optimization and an a posteriori hot-to-cold transformation, we propose setting the cold state CAD parameters as the design parameters for the optimization with the cold-to-hot deformation being computed during the simulation. As a result, the optimization would directly compute a manufacturable cold state geometry of a component optimized for a chosen design point. Since the final goal is an adjoint multidisiplinary optimization, the differentiation of an entire cold-to-hot chain is required. In this paper, we present first steps towards this alternative by implementing an adjoint version of a cold-to-hot transformation tool to compute the required sensitivities.

First, we will motivate our goal of adjoint multidisciplinary optimization in section 2. Sections 3 and 4 outline the cold-to-hot algorithm and its differentiation with AD, respectively. The results are presented in section 5 with concluding remarks in section 6.

2 ADJOINT MULTIDISCIPLINARY OPTIMIZATION

This section introduces the motivation of adjoint multidisciplinary optimization in turbomachinery. With the goal of optimizing the shape of a turbomachinery component, a metric referred to as the *cost function* $J(\boldsymbol{x}) \in \mathbb{R}^c$, e.g. efficiency, has to be optimized, i.e. maxi-

mized. The parameters which can be varied to minimize the cost function are referred to as the *design parameters* $\boldsymbol{x} \in \mathbb{R}^d$. The cost function is minimized using an optimization method, which mainly falls into one of the two categories of *gradient-free* or *gradient-based* optimization methods.

Gradient-free optimization methods can be used to compute improved designs using only cost function evaluations, but finding the true optimal design is not guaranteed. Additionally, converging towards an optimum can require a large number of evaluations and the cost of the optimization is dependent on the number of design parameters d. Gradient-based optimization methods, on the other hand, use sensitivity information to compute a minimum with less iterations than gradient-free methods. However, these methods compute a local minimum and cannot guarantee that the computed minimum is indeed the global minimum, unless the optimization problem is convex. Furthermore, the sensitivities of the cost function with respect to the design variables

$$\frac{\partial J}{\partial \boldsymbol{x}} \in \mathbb{R}^{c \times d} \tag{1}$$

have to be computed. Typical approximation methods, such as finite differences (FD), compute the gradient (1) at a cost proportional to the number of design parameters d, which can lead to high computational costs.

Using the *adjoint* approach, first introduced by Pironneau [15] and later in the application of aerodynamic design optimization by Jameson [8], [16], the gradient (1) can be computed at a cost proportional to the size c of the cost function J. In the context of aerospace engineering, one is typically interested in few parameters such as efficiency, mass flow rate, pressure ratio, and maximum stresses. Thus, the size of the cost function is usually much smaller than the size of the design space, such that $c \ll d$. Thus, the adjoint method offers a significant performance advantage for gradient-based optimization methods over gradient-free methods, given a smooth design space. Notably, the cost of computing the gradient is independent of the number of design variables d, allowing a much greater design space. The adjoint approach requires the evaluation of the *adjoint model*

$$\boldsymbol{x}_{(1)} = \frac{\partial J(\boldsymbol{x})}{\partial \boldsymbol{x}}^T J_{(1)} \tag{2}$$

of the cost function. First deriving a continuous form of the adjoint model (2), then discretizing it, is referred to as the *continuous adjoint* [9] approach. However, deriving a continuous adjoint model of a complex system of equations can be extremely difficult and error prone. Alternatively, the *discrete adjoint* [11] approach discretizes the system of equations first, then derives an adjoint model of the discretized system of equations. By using the code transformation method *algorithmic differentiation* (AD) [7], [13], also known as *automatic differentiation*, an adjoint model of a computer program code can be easily generated. In total, this makes the discrete adjoint optimization method an efficient and precise alternative to gradient-free optimization methods.

Gradient-free optimization methods have been applied in the context of CAD-based multidisciplinary optimization of turbomachinery components e.g. by [12], [17]. In this project, we seek to apply the discrete adjoint approach. We especially want to focus on the fluid-structure interaction (FSI) problem portrayed in the introduction (section 1). An essential aspect of FSI within a CAD-based optimization is the interface between the fluid, solid, and CAD disciplines. The

cold-to-hot transformation (section 3) provides one such interface. Within an adjoint optimization, a discrete adjoint model of this interface is required, which is achieved by differentiating the cold-to-hot transformation using AD (section 4).

3 COLD-TO-HOT TRANSFORMATION

The cold-to-hot transformation implemented in this project is based on the method detailed in [5]. Given an FEM mesh that is generated from a cold state CAD geometry, an FEM linear elastic computation is carried out to compute the displacements generated by given centrifugal forces. For now, we are focusing on deformations caused only by centrifugal forces. By perturbing the B-spline surface control points, the computed FEM displacements are matched and the CAD geometry is transformed to its hot state. In this section, the key steps of the transformation will be outlined. The interested reader is referred to [5] for a more detailed explanation of the algorithm and to [4], [14] for an in-depth discussion of computational geometry.

A B-spline surface

$$S(u,v) = \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} N_i^p(u) N_j^q(v) \boldsymbol{p}_{i,j}^0, \tag{3}$$

is given with basis functions $N_i^p(u)$ and $N_j^q(v)$ of order p and q, respectively, foot points $0 \leq u \leq 1$ and $0 \leq v \leq 1$, and control points $\boldsymbol{p}_{i,j}^0 \in \mathbb{R}^3$. The position of FEM nodes $\boldsymbol{d}_k^0 \in \mathbb{R}^3$ and their respective displacements $\Delta \boldsymbol{d}_k \in \mathbb{R}^3$ are given as well. $k = 0, ..., s-1$, where s defines the number of FEM nodes on the surface.

3.1 FEM node projection

The FEM nodes $\boldsymbol{d}_k^0$ are projected onto surface foot points (u_k, v_k) by solving a point inversion problem, which can be solved using a suitable algorithm from [14]. As depicted in figure 1, the displacements $\Delta \boldsymbol{d}_k$ are imposed onto the foot points (u_k, v_k), to compute the displaced node points

$$\boldsymbol{d}_k = S(u_k, v_k) + \Delta \boldsymbol{d}_k. \tag{4}$$

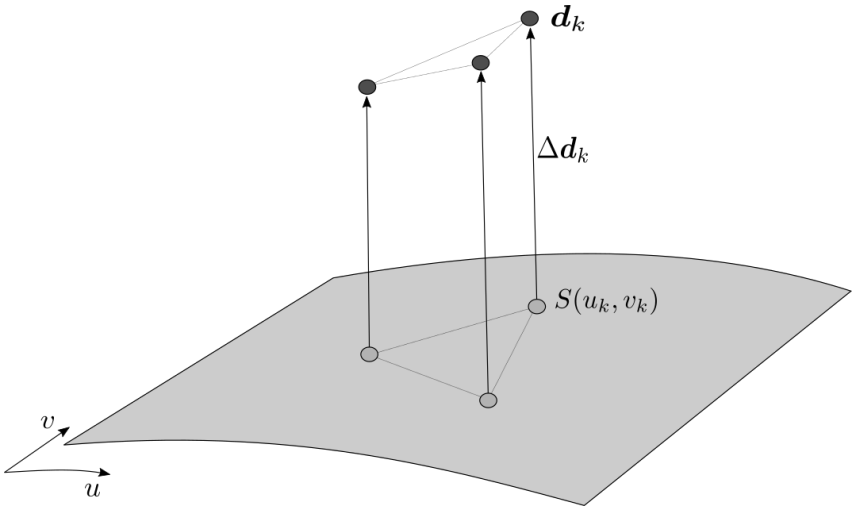


Figure 1: Displacing surface point $S(u_k, v_k)$ with FEM displacement $\Delta \boldsymbol{d}_k$

Using the assumption that the displaced node points $\boldsymbol{d}_k$ must pass through the deformed B-spline surface $\widehat{S}(u_k, v_k)$, the following system of equations can be derived:

$$\sum_{i=0}^{n-1}\sum_{j=0}^{m-1} N_i^p(u_k)N_j^q(v_k)\Delta\boldsymbol{p}_{i,j} = \Delta\boldsymbol{d}_k, \quad k = 0, ..., s-1, \tag{5}$$

where $\Delta\boldsymbol{p}_{i,j}$ represents the change to control point $\boldsymbol{p}_{i,j}$. Typically, this system is overdetermined since the number of FEM nodes s is usually greater than the number of control points $n \cdot m$. This system can be solved using a least-squares method.

3.2 Solve for outer control point displacements

While the solution of the system (5) can deliver the required control point displacements $\Delta\boldsymbol{p}_{i,j}$ for a single surface $\widehat{S}(u, v)$, multiple surfaces with adjacent edges are mostly involved in a complex geometry. An explicit treatment of the surface edges is required to avoid different solutions of $\Delta\boldsymbol{p}_{i,j}$ along adjacent edges, which can lead to undesirable kinks. The control points along the surface edges are thus solved first by treating the edges as B-spline curves. This results in the overdetermined system of equations

$$\sum_{i=0}^{n-1} N_i^p(u_k)\Delta\boldsymbol{p}_i = \Delta\boldsymbol{d}_k, \quad k \in \mathbf{K}_{edge}, \tag{6}$$

with $\mathbf{K}_{edge}$ defining the set of FEM nodes along the edge.

3.3 Solve for inner control point displacements

Once the control point displacements along the edges $\Delta\boldsymbol{p}_{i,0}, \Delta\boldsymbol{p}_{i,m-1}, \Delta\boldsymbol{p}_{0,j}, \Delta\boldsymbol{p}_{n-1,j}$ are determined, the known values of (5) can be moved to the right-hand side to solve for the remaining inner control point displacements

$$\begin{aligned}\sum_{i=1}^{n-2}\sum_{j=1}^{m-2} N_i^p(u_k)N_j^q(v_k)\Delta\boldsymbol{p}_{i,j} = \Delta\boldsymbol{d}_k &- \sum_{i=0}^{n-1} N_i^p(u_k)[N_0^q(v_k)\Delta\boldsymbol{p}_{i,0} + N_{m-1}^q(v_k)\Delta\boldsymbol{p}_{i,m-1}] \\ &- \sum_{j=1}^{m-2} N_j^q(v_k)[N_0^q(u_k)\Delta\boldsymbol{p}_{0,j} + N_{n-1}^p(u_k)\Delta\boldsymbol{p}_{n-1,j}], \\ k \in \mathbf{K}_{inner},\end{aligned} \tag{7}$$

with $\mathbf{K}_{inner}$ defining the set of remaining inner FEM nodes. Afterwards, the control points can be updated by

$$\boldsymbol{p} = \boldsymbol{p}^0 + \Delta\boldsymbol{p} \tag{8}$$

and continuity correction methods can be used to ensure continuity at the common edges of deformed surfaces. Results of the implemented algorithm are presented in section 5.1.

4 SENSITIVITY COMPUTATION

For the computation of the sensitivities of the structural quantity of interest, e.g. the maximum von Mises stress $\sigma_{max} \in \mathbb{R}$, with respect to the design parameters $\boldsymbol{\alpha} \in \mathbb{R}^d$, the sensitivities of an entire chain of operations have to be considered. While the inclusion of pressure loads from the CFD solution is required to fully couple the fluid and structural disciplines, for now

we regard only centrifugal loads. Consider figure 2 for the chain of computations required to compute the structural quantity σ_{max}.

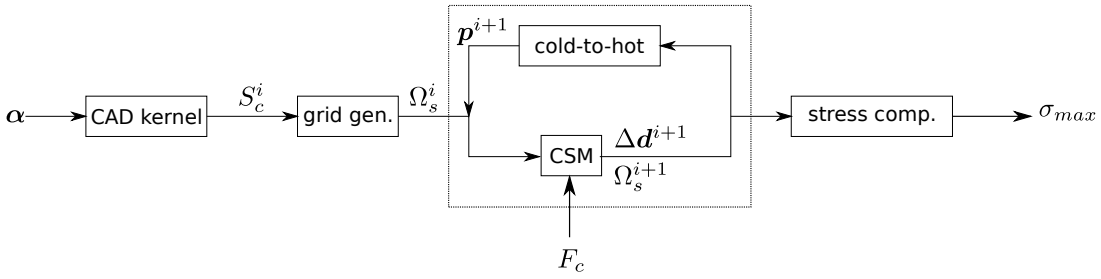


Figure 2: cold-to-hot flow chart

One starts off with design parameters $\boldsymbol{\alpha}$, which can define initial control points of the cold geometry, as well as application specific parameters such as inlet- and outlet radius. $\boldsymbol{\alpha}$ is then plugged into the CAD kernel which generates a cold B-spline surface S_c^i, which is used to generate the initial solid mesh Ω_s^i. The solid mesh, along with the known centrifugal forces F_c, is then used in the computational structural mechanics (CSM) solver to compute the FEM displacements $\Delta \boldsymbol{d}^{i+1}$. The solution of the CSM solver automatically gives us an updated solid mesh Ω_s^{i+1}. The FEM displacements $\Delta \boldsymbol{d}^{i+1}$ are then plugged into the cold-to-hot transformation algorithm to compute the updated control points $\boldsymbol{p}^{i+1}$, which can be used to generate an updated geometry S^{i+1}. Once this deformation loop, which is boxed within a dashed line in figure 2, has converged, the maximum von Mises stress σ_{max} can be computed.

Performing a structurally constrained gradient-based optimization would require differentiating the entire described chain to compute the gradient

$$\frac{\partial \sigma_{max}}{\partial \boldsymbol{\alpha}} \in \mathbb{R}^{1 \times d}. \tag{9}$$

Typically, it would hold that $d \gg 1$, which gives the opportunity of a cheap gradient computation using adjoints. However, this would involve computing the adjoints of the entire cold-to-hot chain. A first step towards computing the adjoints of this entire chain is made by computing the adjoints of the cold-to-hot transformation

$$\boldsymbol{p}(\Delta \boldsymbol{d}) = \boldsymbol{p}^0 + \Delta \boldsymbol{p}(\Delta \boldsymbol{d}). \tag{10}$$

The corresponding 1st-order adjoint model is given by

$$\Delta \boldsymbol{d}_{(1)} = \frac{\partial \Delta \boldsymbol{p}}{\partial \Delta \boldsymbol{d}}^T \boldsymbol{p}_{(1)}, \tag{11}$$

which will yield the gradient

$$\frac{\partial \Delta \boldsymbol{p}(\Delta \boldsymbol{d})}{\partial \Delta \boldsymbol{d}} = \frac{\partial \boldsymbol{p}(\Delta \boldsymbol{d})}{\partial \Delta \boldsymbol{d}} \in \mathbb{R}^{n \cdot m \times s} \tag{12}$$

by performing $n \cdot m$ runs of the adjoint model (11). Each adjoint run i is seeded with the i-th unit vector $\vec{\boldsymbol{i}} \in \mathbb{R}^{n \cdot m}$, where $\vec{\boldsymbol{i}}$ is a vector of 0s, except for the i-th variable, which is a 1 :

$$\boldsymbol{p}_{(1)} = \vec{\boldsymbol{i}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{13}$$

Analogously, the 1st order tanget model is given by

$$\boldsymbol{p}^{(1)} = \frac{\partial \Delta \boldsymbol{p}}{\partial \Delta \boldsymbol{d}} \Delta \boldsymbol{d}^{(1)}, \tag{14}$$

which allows the computation of the gradient (12) by performing s runs with each run j seeded with a unit vector $\vec{\boldsymbol{j}} \in \mathbb{R}^s$:

$$\Delta \boldsymbol{d}^{(1)} = \vec{\boldsymbol{j}}. \tag{15}$$

For the differentiation with AD, the cold-to-hot algorithm is not treated as a black box. Notably, the first step of the cold-to-hot transformation, the projection of FEM nodes onto the surface (section 3.1), does not need to be differentiated. This is because the projection is performed only to determine the foot points (u_k, v_k) of a corresponding FEM node $\boldsymbol{d}_k$. Differentiating equations (6) and (7) is sufficient to compute the gradient (12). This is achieved using the open-source AD tool CoDiPack [1] developed by the Chair for Scientific Computing at TU Kaiserslautern, Germany. The results of the adjoint sensitivity computation compared against finite difference (FD) approximations are presented in section 5.2.

5 RESULTS

In this section, test results of the implemented cold-to-hot transformation and its sensitivity computation are presented. The selected test case is an axial fan blade geometry, which has been used as a baseline geometry for an optimization in [2].

5.1 Cold-to-hot transformation results

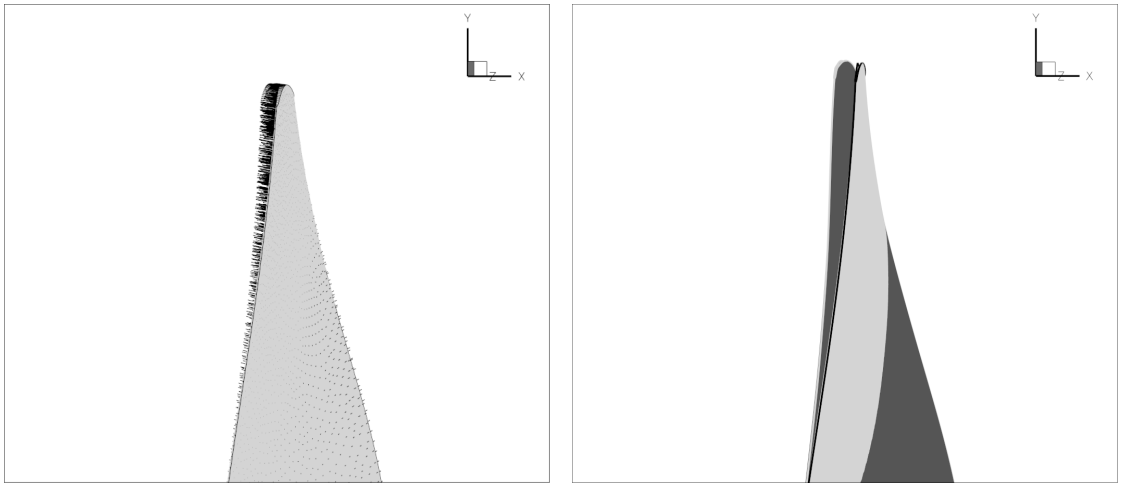
The results of the cold-to-hot transformation of the axial fan are quantified by computing the distance error between the deformed B-spline surface $\widehat{S}(u, v)$ and the displaced points $\boldsymbol{d}$. The error is computed in the same manner as in [5]:

$$\Delta_k = \min_{u,v \in [0,1]} \|\boldsymbol{d}_k - \widehat{S}(u, v)\|^2, \quad k = 0, ..., s-1 \tag{16}$$

The maximum and mean distance errors are computed as

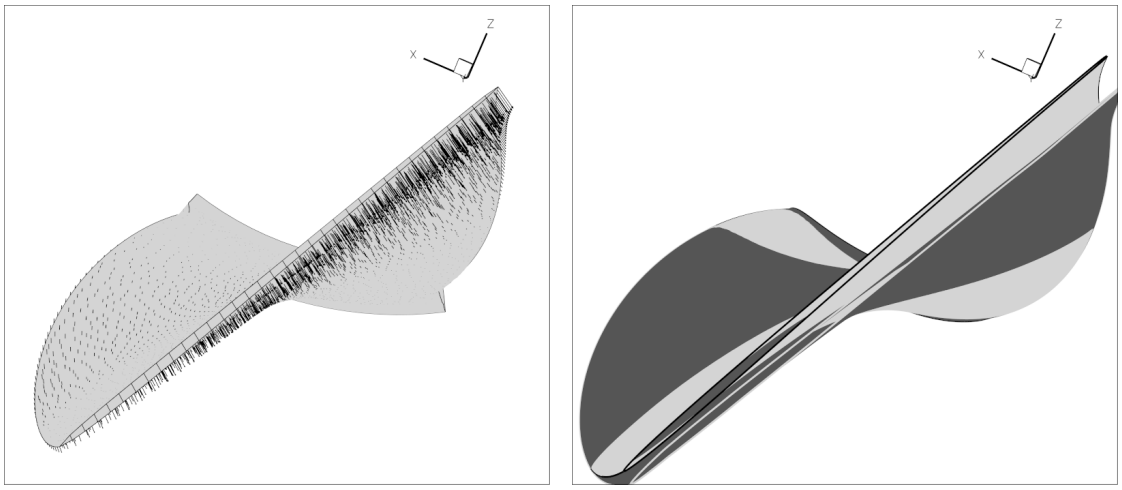
$$\Delta_{max} = \max_k \Delta_k, \quad \Delta_{mean} = \frac{1}{s} \sum_{k=0}^{s-1} \Delta_k. \tag{17}$$

The deformations are visualized in figures 3 and 4, which show that the deformed blade geometry in figures 3(b) and 4(b) follows the displacements given by the vectors in figures 3(a) and 4(a). Distance error results are summarized in table 1.



(a) displacements vectors of FEM results (b) deformed blade: dark. original blade: light

Figure 3: axial fan cold-to-hot transformation



(a) displacements vectors of FEM results (b) deformed blade: dark. original blade: light

Figure 4: axial fan cold-to-hot transformation

	Suction Side	Pressure Side
Δ_{max}[m]	2.83×10^{-3}	2.49×10^{-3}
Δ_{mean}[m]	7.36×10^{-5}	7.53×10^{-5}

Table 1: Maximum distance error Δ_{max} and mean distance error Δ_{mean} of the axial fan blade

5.2 Sensitivity computation results

The cold-to-hot transformation was differentiated using CoDiPack in both forward and reverse mode to compute the 1st order sensitivities (12). The input variables are the FEM node

displacements $\Delta \boldsymbol{d} \in \mathbb{R}^s$ and the output variables are the updated control points $\boldsymbol{p} \in \mathbb{R}^{n \cdot m}$. The number of control points $n \cdot m$ is less than the number of FEM nodes s, with

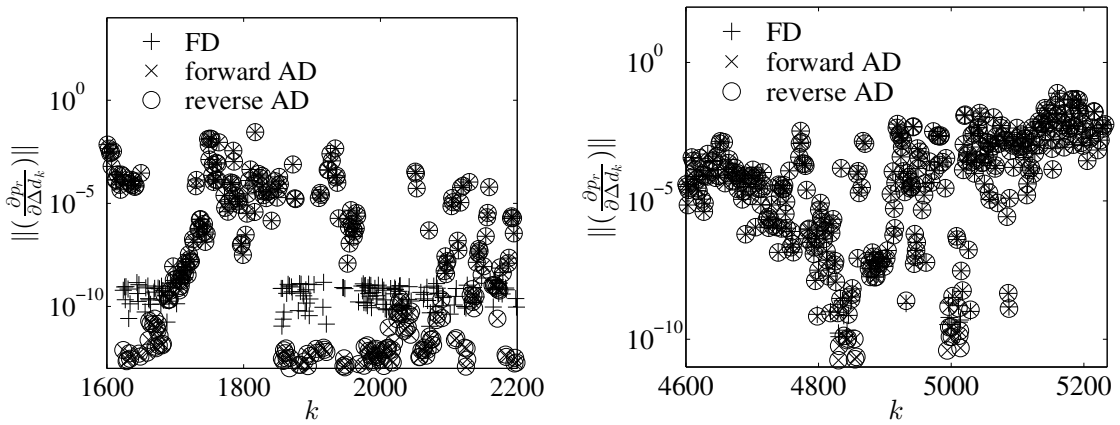
$$n \cdot m = 1053, \quad s = 8149. \tag{18}$$

A finite-difference (FD) approximation of (12) would require $2 \cdot s$ runs of (10) and a forward AD evaluation of the gradient requires s runs of the tangent model (14). On the other hand, computing (12) using reverse AD would require only $n \cdot m$ runs of the adjoint model (11). Thus, a faster run time is expected and also observed for the adjoint AD computation compared to the FD approximations and tangent AD runs (see table 2). The relative run time results are compared to the measured cost of the primal computation of the control point displacements, excluding the FEM node projection, $cost(P) \approx 4.52$ seconds. Note that in the final application of the adjoint cold-to-hot chain, only a single adjoint computation of the cold-to-hot transformation is necessary for each output parameter of the entire chain.

	FD	forward AD	reverse AD
absolute run time [min]	83.57	98.06	11.94
relative run time	$1110.42 \cdot cost(P)$	$1303.04 \cdot cost(P)$	$158.61 \cdot cost(P)$

Table 2: run time comparison of FD vs forward AD vs reverse AD for the 1st order sensitivities (12), with $cost(P) \approx 4.52$ seconds

Figures 5(a) and 5(b) present a comparison between gradients computed using FD, forward AD, and reverse AD for selected rows r and columns k of $\frac{\partial \boldsymbol{p}_r}{\partial \Delta \boldsymbol{d}_k}$. The plots show a good agreement between the FD and AD computed values with discrepancies between FD and AD at values of around $\mathcal{O}(10^{-10})$ in figure 5(a). A maximum discrepancy of $\mathcal{O}(10^{-6})$ and a mean discrepancy of $\mathcal{O}(10^{-13})$ between FD and AD was computed for the entire matrix (12). A visualization of the 1st-order sensitivities of a single control point with respect to the FEM displacements $\Delta \boldsymbol{d}$ is shown in figure 6.



(a) row $r = 100$ (b) row $r = 200$

Figure 5: Comparison of FD vs forward AD vs reverse AD on a semi-logarithmic plot row of $\frac{\partial \boldsymbol{p}_r}{\partial \Delta \boldsymbol{d}}$

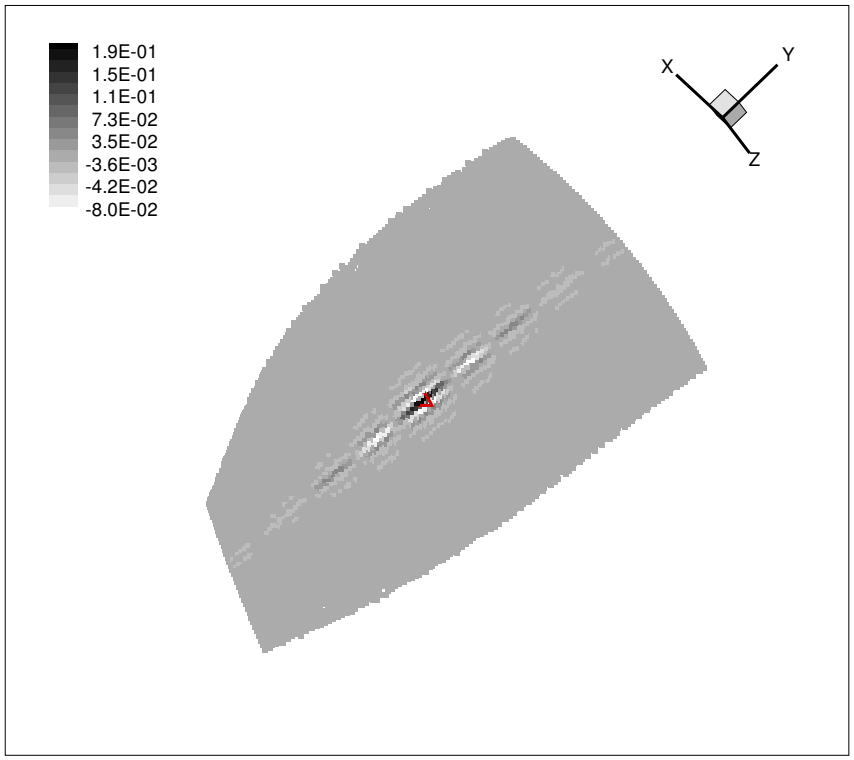


Figure 6: 1st-order sensitivities with respect to $\Delta \boldsymbol{d}$ of center control point, denoted by Δ

6 CONCLUSIONS

With the implementation and adjoint differentiation of the cold-to-hot transformation, an important component of the cold-to-hot chain was addressed. The AD computed sensitivities agree well with the FD computed sensitivities, paving the way towards an adjoint multidisiplinary optimization for turbomachinery components. Further work could involve a local error reduction in the cold-to-hot transformation, as well as the inclusion of pressure forces in the cold-to-hot transformation to couple the fluid and solid computations. An adjoint differentiation of a CSM solver is the next key component of the cold-to-hot chain that is to be implemented.

7 ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support of the IODA (*Industrial Optimal Design using Adjoint CFD*) project, which led to the work presented in this paper. The IODA project is funded by the European Commission within the Horizon 2020 program.

8 NOMENCLATURE

$J \in \mathbb{R}^c$	cost function
$J_{(1)} \in \mathbb{R}^c$	adjoint cost function
$\boldsymbol{x},\ \boldsymbol{\alpha} \in \mathbb{R}^d$	design parameters
$\boldsymbol{x}_{(1)} \in \mathbb{R}^d$	adjoint design parameters
$S,\ S_c,\ \widehat{S} \in \mathbb{R}$	B-spline surface (initial, cold state, deformed)
$N^p,\ N^q \in \mathbb{R}$	B-spline basis functions
$p,\ q \in \mathbb{N}$	B-spline basis function orders
$u,\ v \in \mathbb{R}$	B-spline foot points
$\boldsymbol{p}^0_{i,j},\ \boldsymbol{p}_{i,j} \in \mathbb{R}^3$	control points (initial, displaced)
$\boldsymbol{p}^{(1),i,j} \in \mathbb{R}^3$	tangent control points
$\boldsymbol{p}_{(1),i,j} \in \mathbb{R}^3$	adjoint control points
$\Delta\boldsymbol{p}_{i,j} \in \mathbb{R}^3$	control point displacements
$n \cdot m \in \mathbb{N}$	number of control points
$\boldsymbol{d}^0_k,\ \boldsymbol{d}_k \in \mathbb{R}^3$	FEM nodes (initial, displaced)
$\Delta\boldsymbol{d}_k \in \mathbb{R}^3$	FEM node displacements
$\Delta\boldsymbol{d}^{(1),k} \in \mathbb{R}^3$	tangent FEM node displacements
$\Delta\boldsymbol{d}_{(1),k} \in \mathbb{R}^3$	adjoint FEM node displacements
$s \in \mathbb{N}$	number of FEM nodes
$\mathbf{K}_{edge}$	set of FEM nodes along surface edge
$\mathbf{K}_{inner}$	set of inner FEM nodes
$\sigma_{max} \in \mathbb{R}$	maximum von Mises stres
Ω_s	solid mesh
$F_c \in \mathbb{R}^3$	centrifugal forces
$\vec{\boldsymbol{i}} \in \mathbb{R}^{n \cdot m}$	i-th unit vector
$\vec{\boldsymbol{j}} \in \mathbb{R}^s$	j-th unit vector
$\Delta_k,\ \Delta_{max},\ \Delta_{mean} \in \mathbb{R}$	distance error (k-th, maximum, mean)

REFERENCES

[1] T. Albring, M. Sagebaum, and N.R. Gauger. Development of a consistent discrete adjoint solver in an evolving aerodynamic design framework. *AIAA 2015-3240*, 2015.

[2] Christopher Chahine, Tom Verstraete, and Li He. Multidisciplinary design optimization of an aero-engine fan blade with consideration of bypass and core performance.

[3] Mihai C Duta, Shahrokh Shahpar, and Michael B Giles. Turbomachinery design optimization using automatic differentiated adjoint code. In *ASME Turbo Expo 2007: Power for Land, Sea, and Air*, pages 1435–1444. American Society of Mechanical Engineers, 2007.

[4] Gerald Farin. *Curves and surfaces for computer-aided geometric design: a practical guide*. Elsevier, 2014.

[5] Lilia Gaun, Dieter Bestle, and André Huppertz. Hot-to-cold cad geometry transformation of aero engine parts based on b-spline morphing. In *ASME Turbo Expo 2014: Turbine Technical Conference and Exposition*, pages V02BT45A020–V02BT45A020. American Society of Mechanical Engineers, 2014.

[6] MB Giles. Some thoughts on exploiting cfd for turbomachinery design. *University of Oxford Technical Report*, 1998.

[7] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Siam, 2008.

[8] Antony Jameson. Aerodynamic design via control theory. *Journal of scientific computing*, 3(3):233–260, 1988.

[9] Antony Jameson, Sriram Shankaran, and Luigi Martinelli. Continuous adjoint method for unstructured grids. *AIAA journal*, 46(5):1226–1239, 2008.

[10] Jiaqi Luo, Chao Zhou, and Feng Liu. Multipoint design optimization of a transonic compressor blade by using an adjoint method. *Journal of Turbomachinery*, 136(5):051005, 2014.

[11] Dimitri J Mavriplis. Discrete adjoint-based approach for optimization problems on three-dimensional unstructured meshes. *AIAA journal*, 45(4):741–750, 2007.

[12] Lasse Mueller, Zuheyr Alsalihi, and Tom Verstraete. Multidisciplinary optimization of a turbocharger radial turbine. *Journal of Turbomachinery*, 135(2):021022, 2013.

[13] Uwe Naumann. *The art of differentiating computer programs: an introduction to algorithmic differentiation*, volume 24. Siam, 2012.

[14] Les Piegl and Wayne Tiller. *The NURBS book*. Springer Science & Business Media, 2012.

[15] Olivier Pironneau. On optimum design in fluid mechanics. *Journal of Fluid Mechanics*, 64(01):97–110, 1974.

[16] James Reuther and A Jameson. Control theory based airfoil design using the euler equations. In *5th Symposium on Multidisciplinary Analysis and Optimization, American Institute of Aeronautics and Astronautics, Reston*, pages 206–222, 1994.

[17] Tom Verstraete. Multidisciplinary turbomachinery component optimization considering performance, stress, and internal heat transfer. *University of Ghent, Ph. D. Thesis*, 2008.

[18] Benjamin Walther and Siva Nadarajah. Constrained adjoint-based aerodynamic shape optimization of a single-stage transonic compressor. *Journal of turbomachinery*, 135(2):021017, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOWARDS AN OUTPUT-BASED RE-MESHING FOR TURBOMACHINERY APPLICATIONS

Mateusz Gugala[1], Marcus Meyer[2] and Jens-Dominik Müller[3]

[1] Queen Mary, University of London The School of Engineering and Material Science
Mile End Road, London, E1 4NS, UK,
e-mail: *m.gugala@qmul.ac.uk

[2] Rolls-Royce Deutschland Ltd & Co KG
Eschenweg 11, 15827 Blankenfelde-Mahlow,
e-mail: Marcus.Meyer@rolls-royce.com

[3] Queen Mary, University of London The School of Engineering and Material Science
Mile End Road, London, E1 4NS, UK,
e-mail: j.mueller@qmul.ac.uk

Keywords: Mesh adaptation, Adjoint method, Multi-grid, Output-based sensor, Truncation error, Re-meshing, Turbomachinery.

Abstract. *A methodology for estimating the truncation error between a fine computational grid and topologically inconsistent coarse grid from a multi-grid sequence is presented. This sensor is then weighted with the adjoint solution to obtain a goal-based refinement sensor which provides local scales for a re-meshing procedure. The truncation-based sensor, as well as an output-based indicator, were implemented in the proprietary Rolls-Royce code Hydra. The output-based re-meshing refinement is first applied to the simple cube case with an inviscid flow. The truncation-based sensor and the output-based sensor are then evaluated for a turbine stator case as a first step towards the turbomachinery application of the re-meshing refinement methodology. The results are presented and discussed.*

1 INTRODUCTION

The adjoint method [1] is well-established as the most efficient method for an aerodynamic shape optimisation with CFD. Adjoint sensitivities can be also used to obtain a robust adaptation sensor for an objective function which drives a solution-adaptive mesh refinement process that computes a more accurate objective function at the lower computational cost compared to error-estimators without adjoint weighting or compared to heuristic sensors. The adjoint-weighted truncation error known also as an output-based indicator gained popularity as an effective driver for an adaptation process - see e.g. [2, 3, 4]. Through the adjoint weighting, the mesh adaptation is very effectively targeted to those areas of the computational domain where the objective function is highly sensitive to mesh resolution and the local error estimate is large. That is the key advantage of an output-based indicator as compared to other approaches such as e.g. gradient/Hessian-based sensors or pure truncation-error-based sensors. While the latter at least attempts to estimate the actual errors, both of these methods apply the refinement to all errors, regardless of whether they are relevant to the computation of the objective function or not.

Assuming that the adjoint solver is available, the remaining challenge in obtaining output-based sensor lies in the estimation of the truncation error. A popular approach in the literature has been presented by Venditti and Darmofal [5]. The computational grid is locally refined preserving its original topology and the current solution interpolated to the refined grid is used to estimate the truncation error on the computational grid. Further refinement and application of this approach can be found in [2]. Even though no full solution is computed on the refined mesh, some work as higher-order interpolation or a few iterations have to be done on the refined grid and the computational cost is not negligible.

As an alternative, we could consider using the difference between finest computational grid and a coarsened grid to estimate the local error. In this variant the computational effort is significantly reduced and in case of fully coarsened geometric multi-grid methods can even be at no additional effort since the coarse grid is computed anyways. On the other hand, the error estimate is then only valid for the coarse grid, and care has to be taken to make the extrapolation to the fine grid valid.

Using geometric multi-grid levels for truncation error estimation has been proposed by Fraysse and Ponsin [4]. In this approach, the truncation error is estimated on the finest grid level using a topologically consistent coarser mesh. In our approach, we explore topologically inconsistent fine and coarse grids as arising from unstructured grid coarsening [6].

The proposed output-based sensor estimation methodology described in section 2 was implemented in the Rolls-Royce proprietary CFD code Hydra. A small number of explicit smoothing iterations (see appendix A) is applied to the obtained refinement indicator in order to regularise unwanted high-frequency modes arising mainly from the topological inconsistency between meshes used for truncation error estimation. The obtained sensor fields can either be used for hierarchical refinement, leading to topologically consistent grids, or be used in as local sizing fields in a re-meshing procedure as is the case in this work. Obtained fields are used to drive re-meshing process using BoxerMesh[1] [7]. The application of the output-based re-meshing refinement to the simple cube case with inviscid flow is presented in section 4.1. The preliminary results i.e. truncation sensor and output sensor plots are presented for a turbine stator case and discussed in section 4.2.

[1] http://www.cambridgeflowsolutions.com/en/products/boxer-mesh/

2 Description of Methods

2.1 Basic Quantities and Output Error

The discretisation error δU_h is defined as the difference between the exact solution U of the continuous PDE system $R(U) = 0$ and its discrete approximation U_h, where the latter is resulting from the discrete PDE system solve $R_h(U_h) = 0$. The notation $\cdot|_h$ indicates that the exact quantity is either evaluated at the discrete space with characteristic size h, or interpolated onto this space, not to be confused with the discrete quantity U_h denoted only with subscript h. More formally it can be written using a symbolic interpolation operator $U|_h = \mathcal{I}^h U$. However, to keep the derivations more transparent the following notation is used in this section.

$$\delta U_h = U|_h - U_h \tag{1}$$

The truncation error δR_h can obtained using Taylor series expansion (2) which also shows the relation between the discretisation error and the truncation error.

$$R(U)|_h - R_h(U|_h) = \left(\frac{\partial R}{\partial U}\right)_h \delta U_h + \dots \tag{2}$$

An output error δJ_h due to the inexact solution can be similarly expanded using Taylor series:

$$J_h(U|_h) - J_h(U_h) = \left(\frac{\partial J}{\partial U}\right)_h \delta U_h + \dots \tag{3}$$

Replacing δU_h in (3) with the discretisation error derived from (2) and skipping higher order terms, one can arrive at the formula (4):

$$\delta J_h \approx \left(\frac{\partial J}{\partial U}\right)_h \left(\frac{\partial R}{\partial U}\right)_h^{-1} \delta R_h = v_h^T \, \delta R_h \tag{4}$$

The adjoint variable v for the objective function J, e.g. lift, drag, translates the truncation error δR into the error in cost function δJ. In this manner, the information on how the truncation errors in each control volume contribute to the error in the cost function is obtained. Summation of all contributions together gives a scalar variable i.e. the output correction, which can be used to increase cost function estimation accuracy. The discrete adjoint approach is used in order to obtain adjoint variable ν_h. Tapenade[2] algorithmic differentiation tool is used for flow solver differentiation. The adjoint system is presented in equation (5). Using discrete adjoint allow obtaining exact gradients that correspond to the cost function evaluated on the discrete space J_h.

$$\left(\frac{\partial R}{\partial U}\right)_h v_h^T = \left(\frac{\partial J}{\partial U}\right)_h \tag{5}$$

Assuming that an adjoint solution is available, the remaining task is to get a good estimate of the truncation error.

2.2 Truncation Error Estimation

The truncation error TE_H is the difference between a mathematical model (PDE, denoted $R(U) = 0$) and its discrete approximation $R_H(U|_H)$, or in other words, it is the error due

[2]AD tool developed at Inria http://www-sop.inria.fr/tropics/

to the truncation of the continuous model. First, the exact system $R(U)$ is expanded using Taylor series around a point in space H where the higher order components are denoted TE_H - equation (6). The truncation error can be evaluated using discrete operator R_H and an exact solution U, where the former is evaluated at $|_H$, the discrete space H - equation (8). The operator $R_H(U|_H)$ is a finite-volume residual which is an integral quantity over each control volume. The truncation error is defined for PDE so the mentioned residual has to be divided by the volume. In this work the TE_H^Ω indicates the undivided finite-volume residual scaled with the volume, and $TE_H = TE_H^\Omega/\Omega$ denotes the truncation error in PDE - as per its definition.

$$R(U) = R_H(U|_H) + TE_H^\Omega \tag{6}$$

$$TE_H^\Omega = R(U) - R_H(U|_H), \;\; R(U) = 0 \tag{7}$$

$$TE_H^\Omega = -R_H(U|_H), \; TE_H = TE_H^\Omega/\Omega_H \tag{8}$$

The truncation error calculated using formula (8) will be called exact truncation error and it is applicable only when the exact solution U is known. In practical applications the truncation error has to be estimated. For this purpose the exact solution U is replaced with its approximate i.e. discrete solution U_h resulting from the discrete system solve $R_h(U_h) = 0$.

$$TE_H = -R_H(\mathcal{I}_h^H U_h) \tag{9}$$

In order to get the truncation error estimate on the finest mesh the prolongation operator is used:

$$TE_h = \mathcal{I}_H^h TE_H \tag{10}$$

The list of simplifications that lead to errors in the estimated TE using presented methodology are following:

- The fine grid solution is used as an approximation of the exact one.
- In practice, a finite precision interpolation operators are used.
- It is assumed that the truncation error estimated on space H can be used as an approximation of truncation errors on space h which should be valid as long as both meshes lie within the asymptotic convergence range.

As mentioned in the last point, both meshes, i.e. coarse (H) and fine (h), should lie within the asymptotic convergence range of solution U in order to get a good error estimate. However in practice, it was shown that even for cases which do not meet this criterion the refinement driven by an adaptation sensor based on the method described can be effective, see e.g. Fidkowski and Darmofal [2].

Within the context of geometric multi-grid solvers, inter-grid transfer operators between fine and coarse grids are available in the solver, and the error estimation can be implemented with a very low effort. The complete procedure used for truncation error estimation is presented in algorithm 1.

Algorithm 1 Truncation Error Estimation Using Geometric Multi-grid

1: Solve discrete system for mesh (h)	U_h		
2: Restrict solution from fine (h) to coarse (H) grid	U_H^h	$\leftarrow$	$\mathcal{I}_h^H U_h$
3: Estimate truncation error on the coarse grid	TE_H	$\leftarrow$	$R_H(U_H^h)/\Omega_H$
4: Prolong TE_H to fine grid, add remaining (h) residual	TE_h	$\leftarrow$	$R_h(U_h)/\Omega_h - \mathcal{I}_H^h TE_H$

2.3 Adaptation Sensors

Two adaptation sensors are used for the re-meshing algorithm: a) a truncation-error-based sensor TS and b) an output-error-based indicator OS. The variable N_{eq} denotes the number of equations to solve (e.g. 6 for the RANS with Spalart-Allmaras turbulence model). The truncation error sensor TS is a sum of the absolute values of the truncation error from each equation as shown in (11). The sensor is multiplied with the characteristic size of each cell i.e. $h_i = \Omega^{1/3}$ which plays a role of scaling factor in order to prevent infinite refinement in the regions where the errors decrease at a very low rate or diverge (e.g. at shocks) see e.g. Fraysse [8].

$$TS_{h,i} = h_i \sum_{j=1}^{N_{eq}} |TE_{h,i,j}| \tag{11}$$

The output sensor (OS) is the sum of the undivided local truncation errors weighted by the adjoint solution as shown in (12). It hence is the absolute value of the sum of the contributions of each control volume to the error in the objective function.

$$OS_{h,i} = \left| \sum_{j=1}^{N_{eq}} OE_{h,i,j} \right| = \left| \sum_{j=1}^{N_{eq}} v_{h,i,j}\, TE_{h,i,j}^{\Omega} \right| \tag{12}$$

Five explicit smoothing iterations are applied to obtained sensors in order to regularise unwanted high-frequency modes - section A.

3 Re-meshing Strategy

The re-meshing approach using BoxerMesh[3] [7] and the output-based sensor defined in equation (12) is used to drive the adaptation process. The mesher uses octree cut-cell algorithm to create an initial mesh respecting defined areas of refinement. Created octree mesh is then fitted to the geometry defined by the user and in the final step the boundary layer is extruded. An example cross-section of the stator blade mesh is presented in figure 1.

[3] http://www.cambridgeflowsolutions.com/en/products/boxer-mesh/

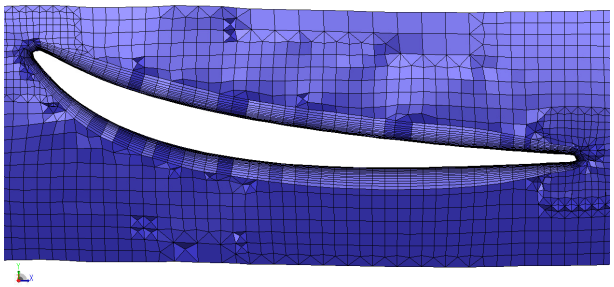


Figure 1: *Cross-section of the stator mesh generated in Boxer.*

The procedure for single re-meshing step is as follows:

1. Obtain flow solution (U_h).
2. Estimate truncation error (TE_h) as presented in algorithm 1.
3. Obtain adjoint solution (ν_h).
4. Evaluate output-based sensor (OS_h) - eqn. (12).
5. Perform 5 explicit smoothing iterations (see appendix A) on obtained sensor (OS_h) to damp unwanted high-frequency modes.
6. Use Paraview to extract mesh region for refinement.
    - Use the 'Threshold' option to mark region for refinement, figure 2.
    - Extract surface and output an STL file.
7. Import surface to Boxer and specify new refinement region for octree mesher.
8. Generate new mesh and re-run the case.

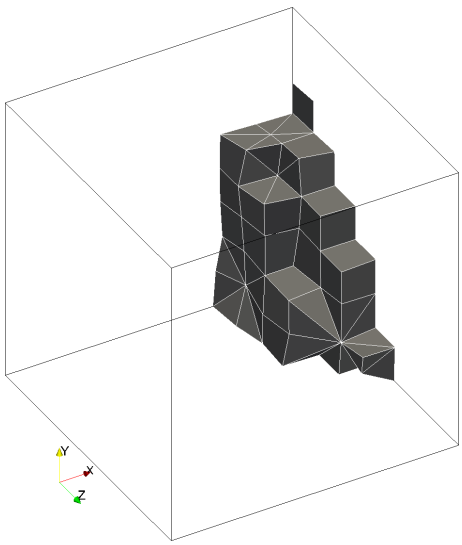


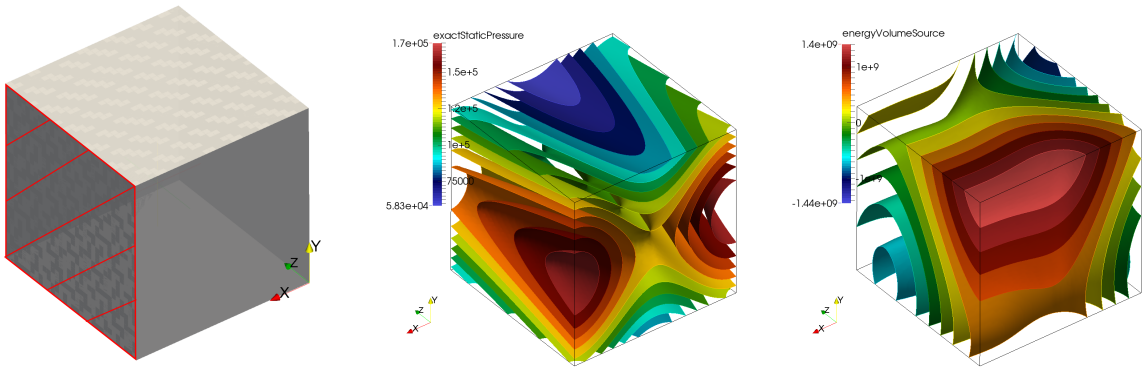
Figure 2: *Cells forming region for refinement obtained using 'Threshold' feature in Paraview.*

Currently steps (5-8) requires manual operations from the user.

4 Application

4.1 Cube with 3D Manufactured Solution

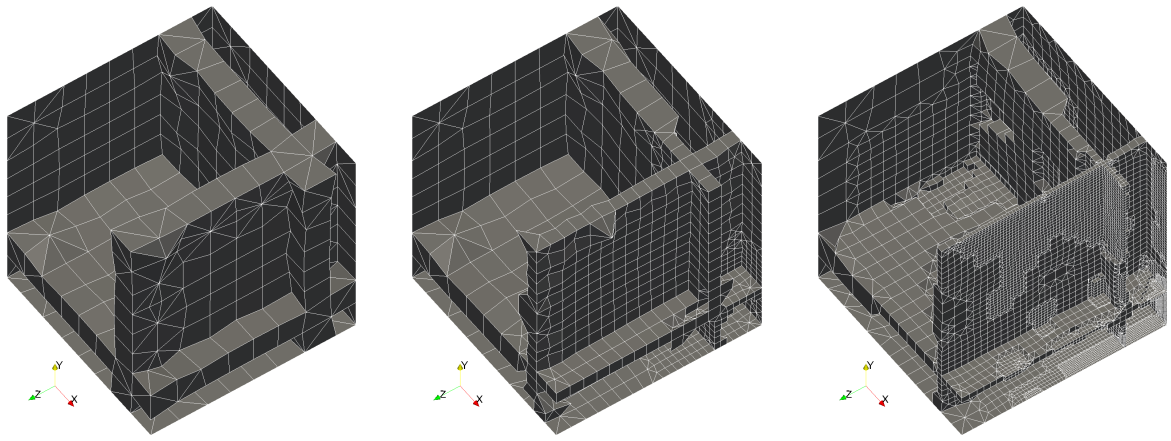
As the first output-based re-meshing example the cube case with 3D manufactured solution by Roy [9] is used. Although the case is physically meaningless it is challenging for the solver as it uses highly nonlinear manufactured solution. It is a compressible, supersonic Euler flow where the example pressure field and corresponding manufactured source term are presented on figure 3. The objective function is a drag force integrated over on one of the cube sides marked in figure 3a.



(a) *Patch for objective evaluation* (b) *pressure (p)* (c) *energy src* (f_e)

Figure 3: The 3D supersonic manufactured solution.

The initial grid was generated using Boxer and is of mixed cell type. The coarse grid for truncation error estimation was generated using internal Rolls-Royce edge-collapsing tool. The re-meshing was performed according to the procedure described in section 3. Total two re-meshing steps were applied and the resulting refined meshes are presented on figure 4. The complex and non-intuitive refinement structures can be easily noticed.



(a) *Initial mesh* (b) *Re-mesh step 1* (c) *Re-mesh step 2*

Figure 4: *Re-meshing process driven by the output-error-based sensor.*

Figure 5 shows the comparison of the achieved error in objective function for the uniformly refined set of grids - 3x3x3 to 129x129x129, regular hex where each refinement stage was achieved by halving the coarser grid edges, and the output-based re-meshed grids. The improved convergence slope for the later approach is obtained - between h^3-h^4. The summary table 1 shows that the same objective function accuracy can be obtained for almost an order of magnitude lower mesh size when the output-based sensor is used, even though a rather crude re-meshing methodology and the topologically inconsistent meshes for truncation error estimation were used. The topological inconsistency between grids is what distinguishes truncation error estimation methodology used in this work as compared to nested grids used by Ponsin and Fraysse [4] as well as Venditti or Fidkowski [2, 3].

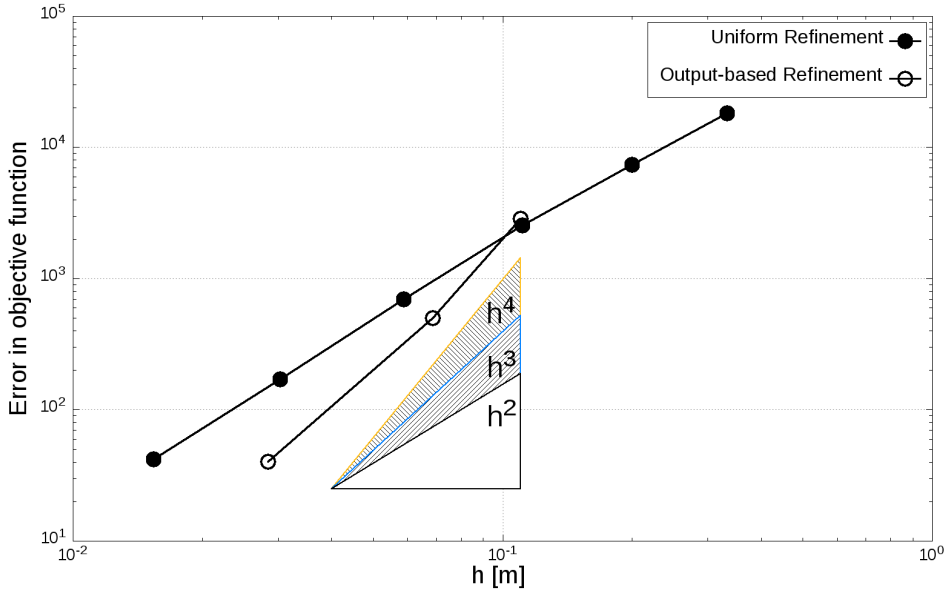


Figure 5: *Re-meshed vs uniformly refined regular mesh - error convergence comparison.*

Re-meshing stage	$\delta L/\widetilde{L}$ [%]	N^{OS}_{DoF}	N^{U}_{DoF}	N^{OS}_{DoF} / N^{U}_{DoF}
0	2.11	754	660	0.87
1	0.37	3082	12100	3.9
2	0.03	43349	335000	7.7

Table 1: Quantitative comparison of achieved objective accuracy between re-meshed grids using output-based sensor and uniformly refined regular hex meshes. DoF - degrees of freedom, N^{OS} - DoF for output-based refinement, N^{U} - DoF for corresponding uniformly refined grid.

4.2 Preliminary Application Results for the TurboLab Stator[4] Case from TU Berlin

The turbine stator from TurboLab at TU Berlin is used The boundary conditions are 42 degrees of swirl angle at the inlet, and outlet static pressure adjusted to keep the mass flow

[4]http://aboutflow.sems.qmul.ac.uk/events/munich2016/benchmark/testcase3/

rate at 9.0 kg/s. The objective function is total pressure loss weighted by mass flow. The flow solution is presented in Figure 6 which shows static pressure on the hub and blade whereas velocity profiles are presented in axial and radial sections.

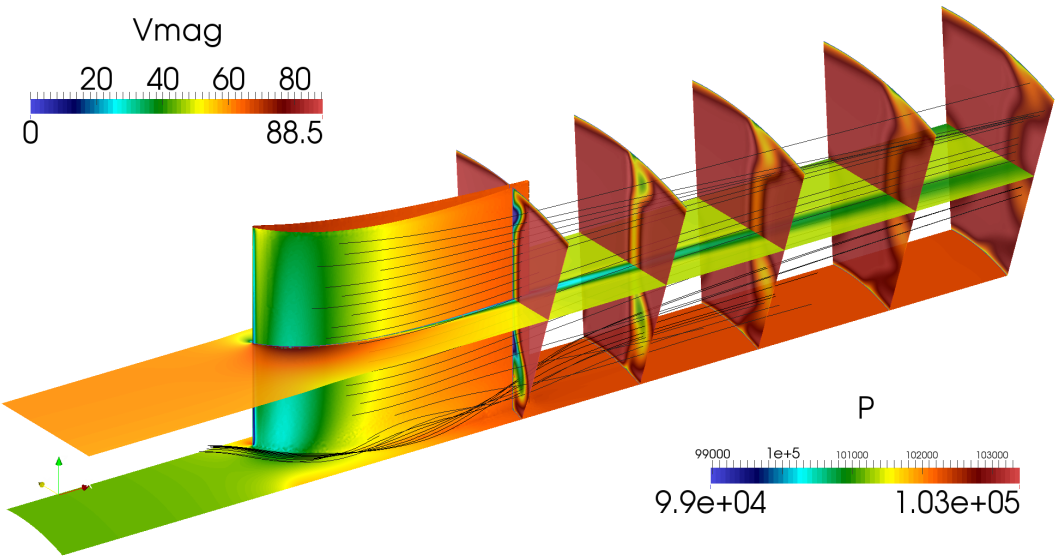


Figure 6: *Flow around the stator blade.*

Figure 7 shows the iso-volume that encloses regions within the domain with the truncation sensor value (11) above a user specified threshold.

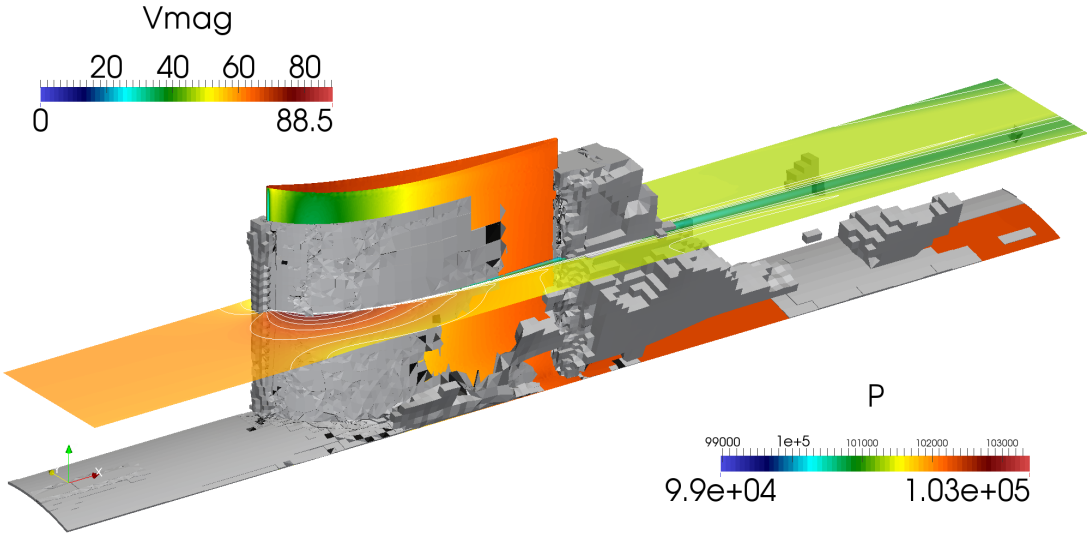


Figure 7: *Iso-volume of truncation sensor.*

Similarly, figure 8 presents iso-volumes for the output sensor (12). The iso-fields were clipped to hide the casing and make the picture more transparent. The fields differ significantly. The truncation sensor targets mainly leading and trailing edges and regions close to the blade. The output-sensor however marks mainly coarse regions of mesh at the proximity of the

inlet and regions in the wake. Both sensors capture the region of a spinning horseshoe vortex between the hub and the blade suction side - visible in figure 6.

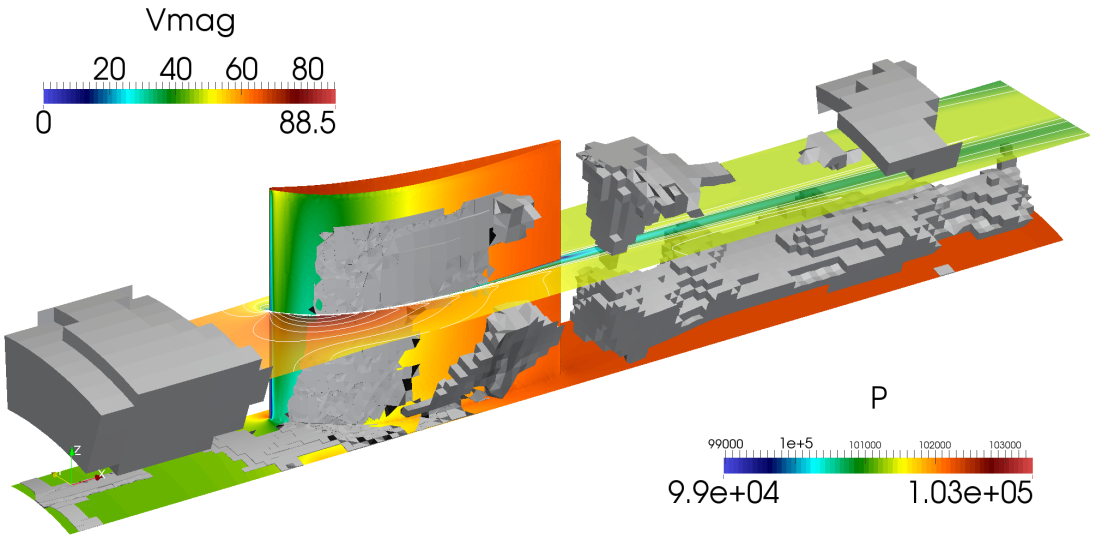


Figure 8: *Iso-volume of output sensor.*

At the moment, only the sensor fields were generated. Application of the re-meshing to the stator case is an ongoing work.

5 Summary

The geometric multi-grid method with topologically inconsistent fine and coarse grids was used for truncation error estimation. The methodology was successfully applied to the simple cube case with a highly nonlinear manufactured solution. Roughly an order of magnitude grid size reduction was obtained for the re-meshed grid as compared to uniformly refined case keeping the same cost function accuracy. The preliminary results for the turbine stator case were presented and the obtained truncation sensor and output sensor patterns compared and discussed. The re-meshing refinement application of the obtained sensor fields to the presented turbine stator case is and ongoing work.

Acknowledgement

This research utilised Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1.

The authors gratefully acknowledge the permission of Rolls-Royce Deutschland to publish this paper. This work has been conducted within the AboutFlow project[5], funded by the European Union FP7 Programme for research, technological development and demonstration under Grant Agreement No. 317006.

The author gratefully acknowledges and thank B. Diskin and F. Fraysse for valuable comments on author's questions (electronic correspondence) regarding their work and publications.

[5]http://aboutflow.sems.qmul.ac.uk

REFERENCES

[1] Pironneau, O. On optimum design in fluid mechanics. **64**, 97–110 (1974).

[2] Fidkowski, K. J. and Darmofal, D. L. Review of output-based error estimation and mesh adaptation in computational fluid dynamics. **49**(4), 673–694 (2011).

[3] Darmofal, D. and Venditti, D. Anisotropic grid adaptation for functional outputs: application to two-dimensional viscous flows. *JCP* **187**, 22–46 (2003).

[4] Ponsin, J., Fraysse, F., Gómez, M., and Cordero-Gracia, M. An adjoint-truncation error based approach for goal-oriented mesh adaptation. *Aerospace Science and Technology* **41**, 229–240 (2015).

[5] Venditti, D. A. and Darmofal, D. L. Grid adaptation for functional outputs: application to two-dimensional inviscid flows. *Journal of Computational Physics* **176**(1), 40–69 (2002).

[6] Moinier, P., Müller, J.-D., and Giles, M. B. Edge-based multigrid schemes and preconditioning for hybrid grids. **40**(10), 1954–60 (2002).

[7] Demargne, A., Evans, R., Tiller, P., and Dawes, W. N. Practical and reliable mesh generation for complex, real-world geometries. In *52nd AIAA Aerospace Sciences Meeting-AIAA Science and Technology Forum and Exposition, SciTech 2014*. American Institute of Aeronautics and Astronautics Inc., (2014).

[8] Fraysse, F., Valero, E., and Ponsin, J. Comparison of mesh adaptation using the adjoint methodology and truncation error estimates. *AIAA journal* **50**(9), 1920–1932 (2012).

[9] Roy, C. J., Smith, T. M., and Ober, C. C. Verification of a compressible cfd code using the method of manufactured solutions. *energy* **2**, s2 (2002).

[10] Desbrun, M., Meyer, M., Schröder, P., and Barr, A. H. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 317–324. ACM Press/Addison-Wesley Publishing Co., (1999).

Appendix

A Weighted Explicit Laplacian Smoothing

The Laplacian equation has the form presented in equations 13. It can be easily discretised using e.g. and explicit time stepping and an edge-based data structure as presented in equation (14). The constant $\beta = 1.0$ was used which is related to the maximum allowed time step, constrained by the stability condition of an explicit scheme. The edge-length-weighting is applied in order to prevent distortions due to the grid size, see e.g. Desbrun [10].

$$\frac{\partial \phi}{\partial t} = \Delta \phi \tag{13}$$

$$\phi_i^{n+1} = \phi_i^n + \beta \left(\sum_{j=1}^{m} \frac{\phi_j^n - \phi_i^n}{l_{ij}} \right) \Bigg/ \sum_{j=1}^{m} l_{ij} \tag{14}$$

Figure 9 shows the eigenvalues of explicit smoothing system matrix for a simple 1D case. Each eigenvalue corresponds to the eigenvector i.e. shape mode. The magnitude of the eigenvalue tells how much given shape mode will be damped when the smoothing is performed. The graph confirms that the explicit smoothing is very effective in filtering the high-frequency modes while having a minor influence on the the low-frequency modes.

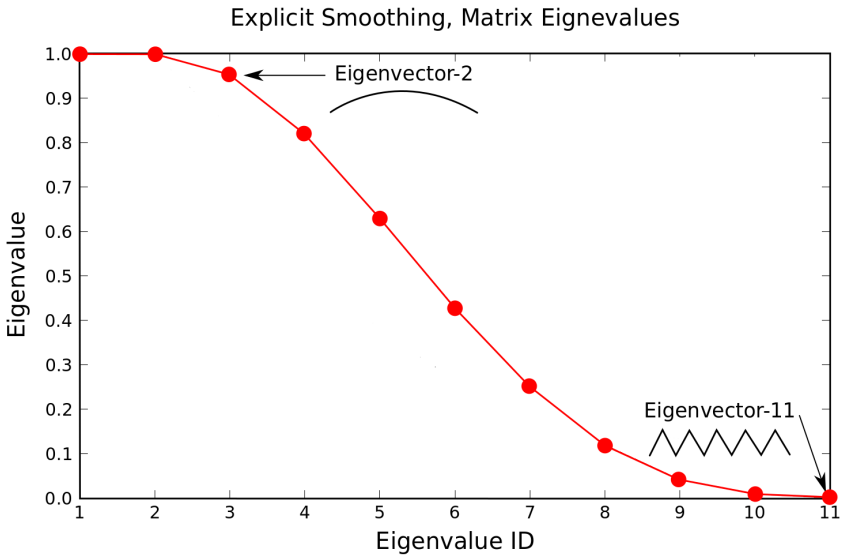


Figure 9: *Eigenvalues of explicit smoothing system matrix for simple 1D case.*

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

OPTIMISATION OF A U-BEND USING A CAD-BASED ADJOINT METHOD WITH DIFFERENTIATED CAD KERNEL

Salvatore Auriemma[1], Mladen Banovic[2], Orest Mykhaskiv[3], Herve Legrand[1], Jens-Dominik Müller[3], Tom Verstraete[3] and Andrea Walther[2]

[1]OpenCascade
1 Place des Frères Montgolfier, 78280 Guyancourt, France
e-mail: salvatore.auriemma@opencascade.com

[2] University of Paderborn
Warburger Str. 100, 33098 Paderborn, Germany
e-mail: mladen.banovic@math.uni-paderborn.de

[3] Queen Mary University of London
Mile End Road, London, E14NS, UK
e-mail: o.mykhaskiv@qmul.ac.uk

Keywords: Adjoint Shape Optimisation, CAD Sensitivities, Automatic Differentiation.

Abstract. *In order to optimise the shape of a three-dimensional CAD-based model using the computationally efficient adjoint methods, the calculation of shape sensitivities, the derivatives of the surface position with respect to the design parameters, is required. This sensitivity is usually not available with CAD systems, but can be obtained by applying the Finite Difference method to CAD-system. Finite-Differences or part-analytic differentiation have been proposed to obtain sensitivities, but have their drawbacks. If source code is available, automatic differentiation can provide accurate derivatives without incurring topology changes or requiring hand-differentiation.*

This paper proposes the differentiation of the open-source CAD kernel - OpenCascade Technology (OCCT) with AD software tool ADOL-C (Automatic Differentiation by Overloading in C++). As a case study we consider the optimisation of pressure loss in a U-bend pipe. The geometry of the U-bend is parametrised in OCCT with a number of cross-sections lofted along a guiding path line. The corresponding geometric derivatives are used in CFD optimisation loops with the resulting shape outperforming the initial design.

1 INTRODUCTION

In industrial shape optimisation the computationally efficient gradient-based methods are used extensively. In particular, the adjoint method [1, 2] is the most effective approach as it allows to compute the sensitivities to an arbitrary number of control variables in a single computation. To complete the chain rule of derivatives, we also need the derivative of the shape parametrisation. A wide range of parametrisations have been developed, here it is useful to distinguish between CAD-free and CAD-based methods. CAD-free methods [3] also often referred as mesh- or lattice-based, optimise the positions of computational grid points using either a globally interpolated distortion field from radial basis functions (RBF), an auxiliary grid defining perturbations such as Free-Form deformation (FFD) or lattices of Hicks-Henne bumps. Mesh-based approaches use the surface mesh of the CFD grid [4] to impose a displacement. This design space is actually too rich for the CFD computation, high-frequency oscillations are not damped adequately, and hence require regularisation. A major drawback of all the CAD-free methods is that the optimised shape exists only as a mesh. Importing this shape back to CAD for further analysis or manufacturing is an unsolved challenge and typically incurs significant approximation. As a consequence the quality of the optimum is impaired.

As an alternative, CAD-based methods work with the CAD in the design loop and use CAD parameters or variables as the design variables. The main advantage is that CAD geometry is taken as an input and CAD geometry is produced as an output, however obtaining derivatives is more challenging. Xu et al. [5, 6] propose the NSPCC approach which uses the NURBS control points in the CAD-native boundary representation (BRep) as degrees of freedom. Additional constraints need to be imposed to retain the desired continuity between NURBS patches or to respect thickness, radius or build-space constraints. These constraints are evaluated numerically and the design space is the kernel of the constraint matrix which is computed using SVD. Selecting the SVD cutoff for non-singular modes provides an effective preconditioner of the design space. The NURBS geometry engine is implemented in source and derivatives are obtained by application of automatic differentiation. The main advantage of the NSPCC approach is that it is vendor neutral and only considers the BRep of the standardised STEP file. As a consequence, it is agnostic to any design parametrisation set up in the CAD system.

Robinson et al. [7] use the design parametrisation and internal variables set up in a closed-source CAD system as design variables and obtain derivatives with finite differences. To avoid issues with patch re-numbering and disappearance due to the finite-size displacements, the geometry is projected onto an STL approximation of the surface and the finite-differences of the displacements of grid nodes are evaluated on this STL, which is a computationally expensive process and further affects accuracy of the gradients. However, the method does allow to define a design space with constraints through the CAD parametrisation.

Dannenhoffer and Haimes [8] use the open-source CAD-kernel OpenCascade Technology (OCCT) as a geometric engine. They apply analytic differentiation to known simple shapes such as circles and cylinders defined by origins, radii and axes. Remaining derivatives are evaluated using finite differences.

In this paper we present the differentiation of the entire geometric kernel of OCCT with Automatic Differentiation (AD) by the ADOL-C AD tool [9]. Compared to other CAD-based approaches this results in a number of significant advantages. Similar to [7, 8] the design space is defined by the parametrisation of the CAD system which allows to build in geometric constraints, while in the NSPCC approach [6] these constraints have to be reimposed. As opposed to the finite-difference approaches [7, 8] the geometric sensitivities are exact and not affected by

truncation error. Most importantly, there is no finite-size displacement of the geometry during the differentiation, hence this approach is not affected by patch re-numbering or disappearance and a projection onto an intermediate surface is not necessary. This should significantly reduce the computational cost of the method. Finally, automatic differentiation can also be performed in reverse mode [10], which has the potential to dramatically reduce the computational cost of derivative computation as used for the flow solver (cf. Sec. 2.1). While we do not use the reverse-mode differentiation in this paper, the successful demonstration of forward-mode differentiation presented here is a step toward this.

The paper is structured as follows. Sec. 2 presents the governing equations for flow and its adjoint, as well as the assembly of the relevant derivatives. Sec. 3 describes the differentiation of the OCCT CAD system. Sec. 4 shows the U-bend testcase and the definition of the parametrisation, followed by computational results in Sec. 5.

2 CAD-DRIVEN SENSITIVITY

2.1 Primal and adjoint flow equations

To optimise a scalar cost function J which describes the aerodynamic performance of the system of interest, the optimisation problem can be stated as

$$\min_{\alpha} \; J(U(X(\alpha)), X(\alpha), \alpha) \,, \tag{1}$$

$$R(U(X(\alpha)), X(\alpha)) = 0 \,. \tag{2}$$

Equation (2) denotes the system of steady-state Reynolds-Averaged Navier-Stokes equations, where the residual R is driven to zero. R is a function of the state variable U and the mesh coordinates X, both depending on the design parameters α. The objective function J could correspond to drag, lift, total pressure losses, etc. Application of the chain rule to the system yields

$$\frac{dJ}{d\alpha} = \Big[\frac{dJ}{dX} + v^T f\Big] \frac{\partial X}{\partial \alpha} \,. \tag{3}$$

Here v represents the solution of adjoint equations:

$$\Big(\frac{\partial R}{\partial U}\Big)^T v = \frac{\partial J}{\partial U} \,, \tag{4}$$

where

$$f = -\frac{\partial R}{\partial X} \,. \tag{5}$$

The algorithm of calculating the surface mesh sensitivities consists of evaluating the volume sensitivities $\frac{dJ}{dX}$ and then projecting them onto design surfaces. This is performed after primal and adjoint CFD runs, followed by projections which use the mesh-perturbation algorithms. For instance, the spring-based or elasticity-based mesh deformation algorithms detailed in [6] could be used. In other words, the mesh perturbation algorithm creates a correspondence between the movements of volume and surface mesh points. Hence, denoting the surface grid points with X_S, the mapping $X_S \rightarrow X(X_S)$ allows to rewrite the sensitivity in equation (3) in terms of displacements of the surface grid points

$$\frac{dJ}{d\alpha} = \frac{dJ}{dX_S} \frac{dX_S}{d\alpha} \,, \tag{6}$$

$$\frac{dJ}{dX_S} = \Big[\frac{dJ}{dX} + v^T f\Big]\frac{\partial X}{\partial X_S}\,. \tag{7}$$

The gradient obtained in (6) is used in an iterative optimisation process to find optimal design parameters α by updating

$$\alpha^{(n+1)} = A(\alpha^{(n)}, \frac{dJ}{d\alpha}(\alpha^{(n)}))\,, \tag{8}$$

were A represents an optimisation algorithm of choice.

2.2 Assembling CAD-based sensitivity

In the previous section we considered a more general nature of the design parameters, from now on we will refer to α as the parameters of a given CAD-model. In this context the two terms in equation (6) correspond to the flow and geometrical (CAD) sensitivities, respectively. These terms, which are composed of derivatives in each of the n surface mesh points, could be calculated independently from each other and afterwards assembled in a global sensitivity by

$$\frac{dJ}{dX_S} = \Big[\frac{dJ}{dX_{S,i}}\Big]_{i=1,\dots,n}. \tag{9}$$

Similarly, the CAD sensitivity w.r.t. one of the m design parameters $\alpha = (\alpha_1, ..., \alpha_m)$ could be written as:

$$\frac{dX_S}{d\alpha_j} = \Big[\frac{dX_{S,i}}{d\alpha_j}\Big]_{j=1,\dots,m}^{i=1,\dots,n}. \tag{10}$$

For a 3D CAD-model the mesh points X_S have three components, hence the two matrices listed above have dimensions $3 \times n$. Therefore, in order to calculate the gradient components, one primal and adjoint run of a CFD solver is required, and m computations of the matrix in equation (10). Finally, the m components of the gradient are obtained from the scalar multiplication of two previously written matrices as in equation (6):

$$\frac{dJ}{d\alpha_j} = \sum_{i=1}^{n} \frac{dJ}{dX_{Si}}\frac{dX_{Si}}{d\alpha_j}\,. \tag{11}$$

The process of finding the CAD derivatives (second term) in the last equation will be described in detail in the following sections.

The optimisation algorithm consists of the following steps: for the initial design parameters, the CAD-model and the corresponding mesh are built. Afterwards, at each iteration one solves primal and adjoint flow equations and assembles the gradient as described above. Then, the CAD-model is updated with new parameters and the mesh deformation step is performed for the computational grid to match the updated CAD geometry.

3 AUTOMATIC DIFFERENTIATION OF OPENCASCADE TECHNOLOGY

3.1 Introduction to ADOL-C

ADOL-C is a software tool that facilitates the computation of first and higher derivatives of vector functions that are defined by computer programs written in C/C++. It uses the *operator overloading* concept, which means that it is not generating intermediate source code as it is the case for the source transformation approaches [9]. ADOL-C supports the following modes of differentiation:

- Forward (options: *trace-based* or *traceless*),
- Reverse (*trace-based*),

where *trace* is an internal representation of the function to be differentiated, produced by operator overloading. The difference between *trace-based* and *traceless* mode is that in *traceless* mode the derivative computation is done at the same time as the function evaluation. For computing the derivatives using *trace-based* variants, one has to call driver functions provided by ADOL-C after the trace generation. For the purpose of this article, the *traceless forward differentiation*, which computes first order derivatives in scalar mode and in vector mode, is considered.

3.2 Introduction to OCCT differentiation

A key ingredient for automatic differentiation by overloading is the concept of an *active* variable - which is named `adouble` in ADOL-C. All variables that may be considered as differentiable quantities at some time during the program execution must be of an `active` type. Therefore, the integration of the ADOL-C library into a certain code is done by injection of its specific `adouble` type instead of the native `real` type. This kind of integration is not simple when facing complicated *object-oriented* code like OCCT.

Several possible ways of source code modification were considered for ADOL-C integration, but one was taken as a way to proceed with the full sources - the `typedef` approach. The `typedef` approach is the most intrusive way of integrating ADOL-C into OCCT because of replacing all `double` by `adouble`, by using an existing `typedef` which is named `Standard_Real` in OCCT. It is the fastest way of integration because code modification should be as minimal as possible, while the drawback is about sacrificing memory and efficiency to some extend since all `double` variables, even ones not needed for differentiation, will be `adouble` objects.

Although the idea looks simple, it is not as straightforward as one would expect. The differentiation involved a significant amount of code modification and even after successful compilation, a large number of run-time errors had to be resolved during the testing phase. These issues will not be explained here, but will be set out in detail in a future paper that will be mainly focused on the automatic differentiation of OCCT.

After fixing the parts of the code that were giving the errors, a major part of the full differentiated kernel is working. There are still some run-time errors that have to be resolved, but they are not related to the parts of OCCT executed in the application considered here. Before using the differentiated kernel, AD has been verified against Finite Differences (by a *central difference scheme*) in the required OCCT methods for the U-bend construction.

4 PARAMETRISATION OF THE U-BEND

4.1 Geometry

The U-bend under investigation is a typical internal cooling channel for the turbine blades in turbomachinery applications [11]. The geometry is shown in Figure 1. It consists of a circular U-bend with a hydraulic diameter $D = 0.075m$.

4.2 Parametrisation

The parametrisation is based on a cross-sectional design approach which takes N-curves/N-slices as inputs in order to construct a final surface. Each slice consists of a closed wire com-

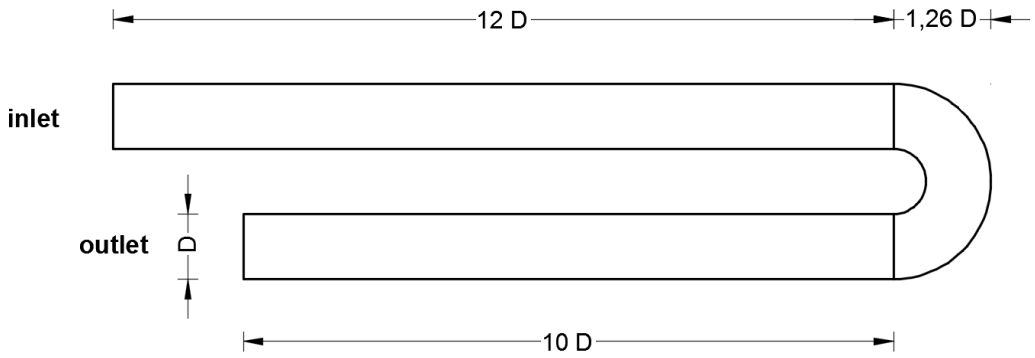


Figure 1: U-bend Dimensions

posed by 4 Bezier curves; every Bezier curve is characterized by 4 control points and shares the first and the last control point with the previous and the following Bezier curve, therefore having in total 12 control points, as shown in Figure 2. This choice allows the section to assume a wide variety of shapes that the optimiser could impose during the optimisation loop.

The plane where the slice lies is constructed in the following way:

- define a B-spline planar curve called *pathline* that drives the U-part of the U-bend in the 3D space.
- Take a point P on the pathline and the vector V tangent to the pathline in P.
- Construct the plane that holds the 12 control points of the section passing through the point P defined as the origin of the plane's (x, y, z) axis, orthogonal to V (whose direction is defined as the z axis) and with y axis orthogonal to the plane where the pathline lies.

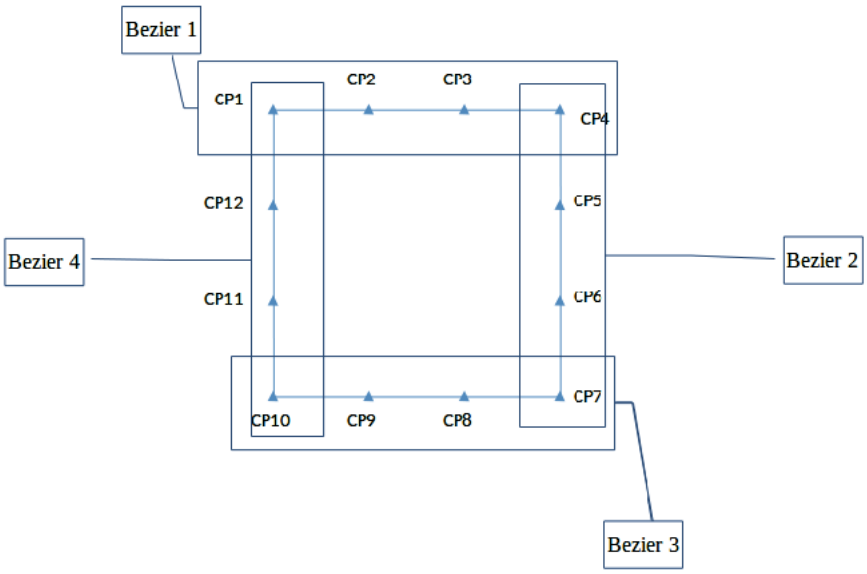


Figure 2: U-bend slice

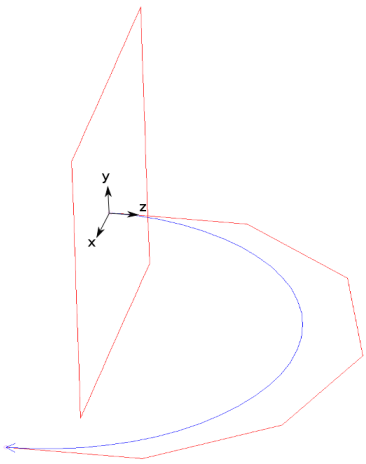
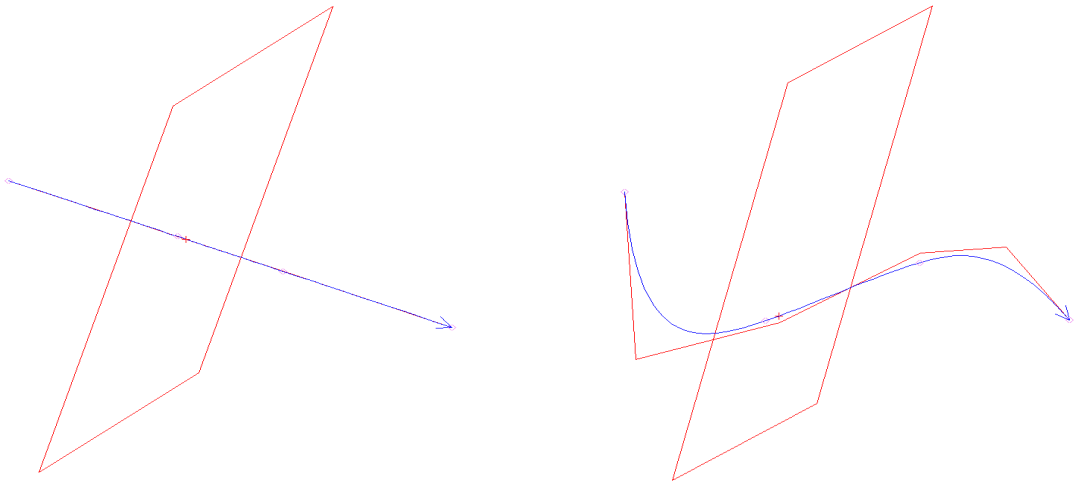


Figure 3: Construction of the first plane

An example of the constructed plane in shown in Figure 3. The slice of the U-part is free to move onto the plane constructed as explained above. Each control point of the section is characterised by a law of evolution along the pathline; in particular, the control point laws are described by B-spline curves in a (ϵ,η,τ) modelling space. These B-spline laws consist of 8 control points whose (ϵ,η) coordinates are the final design parameters of the simulation; the τ coordinates are not considered as design parameters - their values constrain the first and last control point to correspond to the first and the last slice, respectively. The laws are intersected N-times by a plane that is always parallel to plane (ϵ,η) and that passes through a point of coordinates (0,0,P_{length}), where P_{length} is the length of the curve pathline at the point of intersection between the slice and the pathline; this approach in particular is called "Clipping" technique. For each intersection point between the plane and the law, the coordinates of the point of intersection (ϵ_n,η_n) are assigned as coordinates (x_n,y_n) of the control point of the section to which the law is referred. Two examples of the *clipping* technique are shown in Figure 4.
Furthermore to assure the tangency constraint between the U-part and the vertical pipes of the U-bend during the optimisation, for every law the coordinates of the first two and the last two control points are not considered as design parameters of the simulation. In general it would be possible to allow the second and the penultimate law control points to move in τ direction (which means to assign the τ coordinates of these control points as design parameters) but at this stage we preferred to simplify the simulation making the assumption that the optimisation does not consider τ coordinates as design parameters. In a future article an optimisation will be presented that considers all the coordinates of the law control points.

Once that all the control points of a slice assume their position on the plane as described above, it is possible to get the N-slice of the U-bend. The N-slices are then used to obtain the final U-bend B-spline surface. To achieve this result, an Open Cascade cross-sectional design technique [12] that approximates the 4 *NURBS* surfaces (one for every Bezier curve of the section) has been used. Finally the 4 surfaces are merged using an Open Cascade Topology tool.

To complete the U-bend for the test case, the first and the last wire of the U-part are extruded with an Open Cascade tool to get the vertical pipes that corresponds to the inlet and the outlet pipes, as shown in Figure 5.



(a) Clipping before optimisation (b) Clipping after optimisation

Figure 4: Clipping before and after optimisation

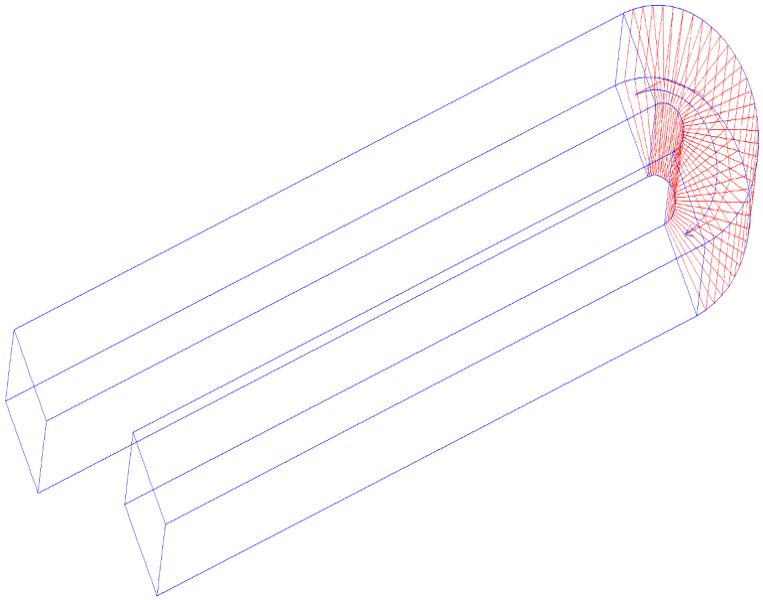


Figure 5: Final U-bend

5 OPTIMISATION RESULTS FOR THE U-BEND

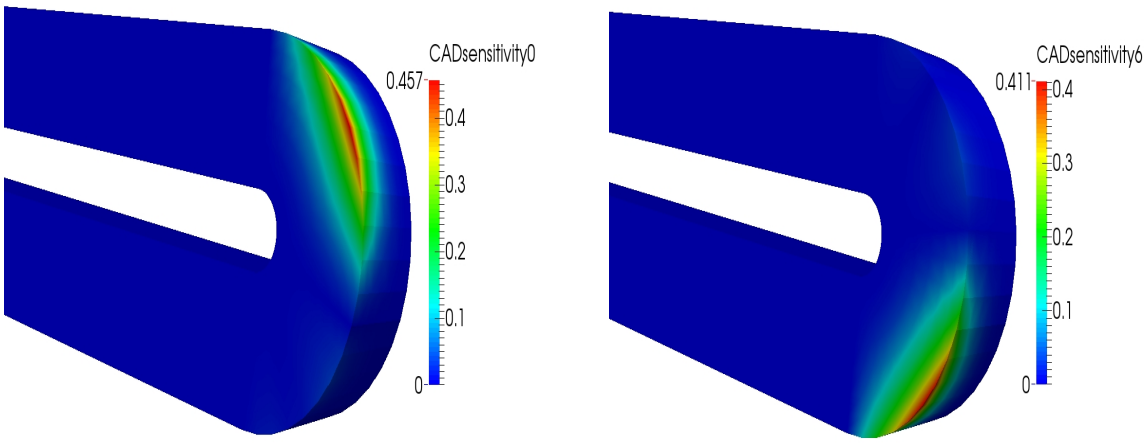
The U-bend model described in the previous section will be optimised to reduce total pressure losses between the inlet and outlet:

$$J = \frac{\int_{inlet} P(\mathbf{u}\cdot\mathbf{n})dS + \int_{outlet} P(\mathbf{u}\cdot\mathbf{n})dS}{\int_{inlet}(\mathbf{u}\cdot\mathbf{n})dS}\,. \tag{12}$$

Only the U-part is subject to design changes. This testcase was initially proposed in [11]. Since we wanted to point out in particular the strength of CAD-based optimisation approach using the differentiated OCCT, a simplified version of the U-bend was chosen to reduce dramatically the computational cost of the CFD. In comparison with [11] much shorter lengths of outlet and

inlet legs were considered. Furthermore, the horizontal outlet and inlet part were not subject to design changes. While longer legs assured fully developed flow at the design part (U-part), shorter legs reduced the size of computational mesh. To further accelerate CFD computations, turbulent flow was replaced with a laminar one. We want to emphasize that the latter does not affect in any way the proposed methodology and more correct physical flow conditions could be applied without any further restrictions.

The CAD-model parametrised in differentiated OCCT depends on 96 design parameters (12 laws of evolution $\times$ 4 control points $\times$ 2 directions), as described in section 4.2. In Figure



(a) Sensitivity Parameter 0 (b) Sensitivity Parameter 6

Figure 6: CAD sensitivities

6 the magnitudes of the CAD sensitivities of two parameters are shown. Since the design parameters correspond to control points positions of the law's B-spline curves, these sensitivities also inherited their local influence feature.

The structured mesh was created to match the given geometry. This could be achieved either by creating a mesh from corresponding CAD-file (STEP, IGES, etc.) or by creating a mesh directly matching the prescribed sizes (feasible in case of non-complex geometries).

For the flow simulation the in-house CFD solver *mgopt* with incompressible setting was used. The solver is based on the vertex-centred Finite Volumes scheme and facilitates a geometric multi-grid method. It includes the differentiated code provided by AD tool Tapenade [13] to implement the discrete adjoint method. To increase stability and robustness of CFD runs the solver makes use of novel implicit time stepping JT-KIRK scheme [14].

The flow is considered to be incompressible with low free stream Mach number $Ma = 0.04566$. Following boundary conditions were imposed: subsonic inlet at the end of the longer leg, and subsonic outlet at the shorter one. On all other boundaries no-slip wall conditions were specified. The air properties at ambient conditions are: density $\rho = 1.204 kg/m^3$, viscosity $\mu = 1.813 \times 10^{-5} kg/(sm)$, pressure $P = 101300 Pa$ and temperature $T = 293.15 K$.

As the optimisation algorithm the Conjugate Gradient method with Armijo-type line search was used. For this particular problem this method has the additional advantage of a direct control over design parameters update, which influences robustness of the mesh perturbation step. The gradient (11) which drives the optimisation was tested with the Finite Differences for few design parameters, showing mutual agreement.

In Figure 7 the decrease of cost function is shown during the course of optimisation. This demonstrates that optimiser components work as expected while the optimised design outperforms the initial one by approximately 4%. The corresponding geometry update is shown in Figure 8.

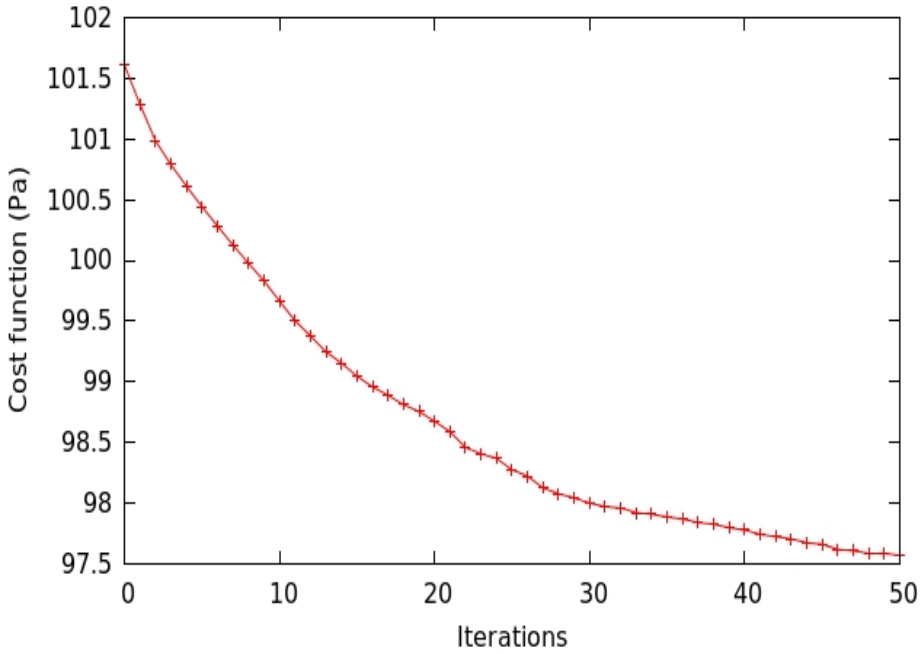


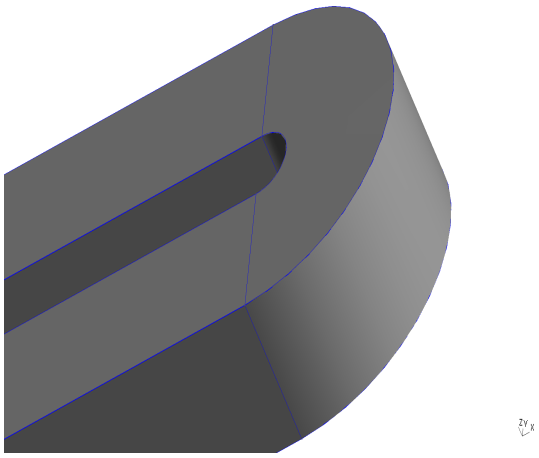
Figure 7: Optimisation loop history

The relatively small improvement can be explained with the case settings. Laminar conditions simplified the flow behaviour in comparison with the turbulent. Fixed inlet and outlet legs also reduced the possible design change. In this setting pressure losses are related mostly to frictional forces and the flow separation occurring after the U-part. In Figure 9 the change in velocities before and after the optimisation are shown. The optimiser managed to suppress the separation bubble and reduce the cost function mainly by increasing the outer U-part hydraulic diameter and a small decrease of the inner circular surface diameter.

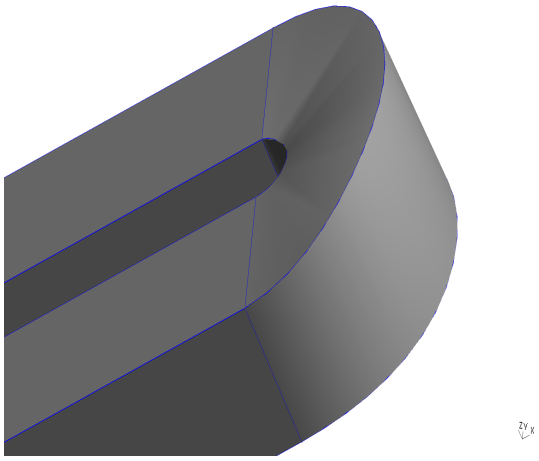
6 CONCLUSIONS AND FURTHER WORK

This paper shows that the automatic differentiation of a fully developed CAD system is feasible, as well as its integration into the design loop. The CAD-model was built in OCCT and its derivative information was used in the global gradient assembly driving the optimisation process. Despite relatively small improvements, the advantages of CAD-based optimisation were exploited. First, the output of the optimisation process was the updated CAD-model. Second, the direct use of the parametrisation ensured a seamless process of parameters update without appearance of cross-patches discontinuities. This is an advantage in comparison with other CAD-based approach, when the control points of the surfaces are taken directly as design parameters. Nevertheless, the obtained CAD-software sensitivities were used in the simplification of the testcase mentioned above resulted in relatively small design changes and cost function decrease. The further investigation on this testcase is planned to solve CFD on a finer mesh and for correct flow physics including additional design domains. Together with more advanced

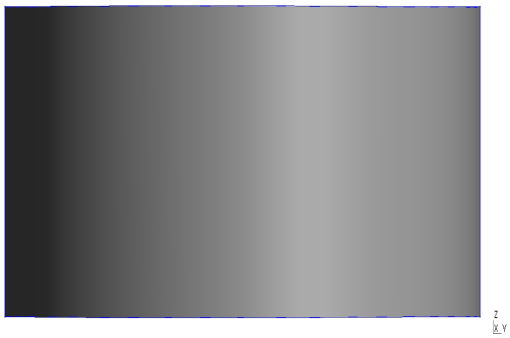
optimisation methods the problem could result in more apparent design update.



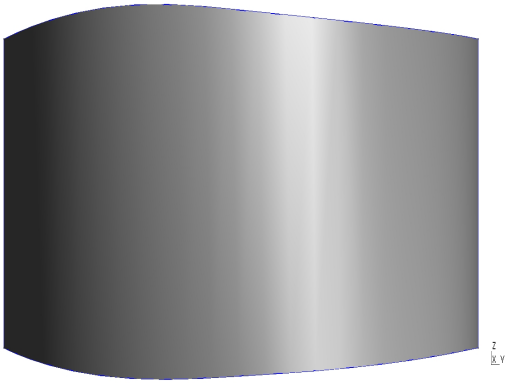
(a) Initial design



(c) Optimised design

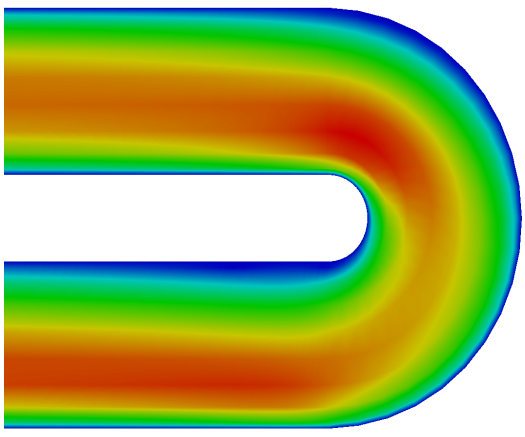


(b) Initial outer circular U-part surface

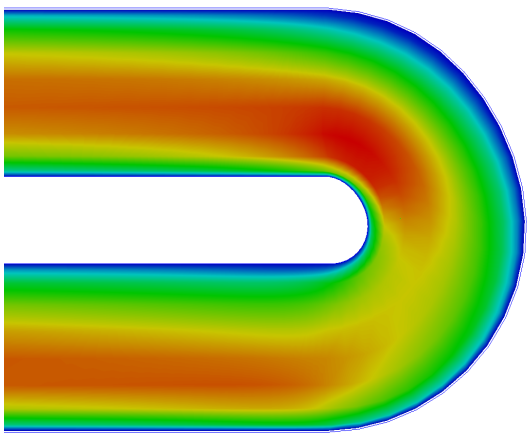


(d) Optimised outer circular U-part surface

Figure 8: Modified Geometry



(a) Initial design



(b) Optimised design

Figure 9: Velocity Magnitude

ACKNOWLEDGMENTS

The authors are very thankful to *Mr. Sergey Slyadnev (OpenCASCADE)* for his support related to the OpenCASCADE Technology, its differentiation and parametrisation.
This research is a part of the IODA project - Industrial Optimal Design using Adjoint CFD. IODA is *Marie Sklodowska-Curie Innovative Training Network* funded by European Commission under Grant Agreement No. 642959.

REFERENCES

[1] A. Jameson, Aerodynamic Design via Control Theory. *JSC*, 3: 233–260, 1988.

[2] M. B. Giles and M. C. Duta and J.-D. Müller and N. A. Pierce, Algorithm Developments for Discrete Adjoint Methods. *AIAA Journal*, 41(2) : 198 –205, 2003.

[3] J.A. Samareh, Aerodynamic shape optimization based on free-form deformation. *AIAA*, 4630: 1–13, 2004.

[4] A. Jameson and J. Vassberg, Studies of alternative numerical optimization methods applied to the brachistochrone problem. *Computational Fluid Dynamics Journal*, 9(3), 281 – 296, 2000.

[5] S. Xu, J. Wolfram, J.-D. Müller, CAD-based shape optimisation with CFD using a discrete adjoint. *Int. J. Numer. Meth. Fluids*, 74: 153–168, 2013.

[6] S. Xu, D. Radford, M. Meyer, and J.-D. Müller, CAD-Based Adjoint Shape Optimisation of a One-Stage Turbine with Geometric Constraints. ASME Turbo Expo 2015: Turbine Technical Conference and Exposition, 2C: Turbomachinery.

[7] T. Robinson, C. Armstrong, H. Chua, C. Othmer and T. Grahs, Optimizing Parameterized CAD Geometries Using Sensitivities Based on Adjoint Functions. *Computer-Aided Design and Applications,* 9(3), 253-268, 2012.

[8] J. Dannenhoffer and R. Haimes, Design Sensitivity Calculations Directly on CAD-based Geometry. *53rd AIAA Aerospace Sciences Meeting, AIAA SciTech*, AIAA 2015–1370.

[9] A. Griewank and A. Walther. *Getting Started with ADOL-C*, 181 – 202. *Combinatorial scientific computing*, 181 – 202, 2009.

[10] A. Griewank and A. Walther, Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. *SIAM*, 2008.

[11] T. Verstraete, F. Coletti, J. Bulle, T. Vanderwielen, and T. Arts, Optimization of a U-Bend for Minimal Pressure Loss in Internal Cooling Channels - part 1: Numerical Method. *ASME Turbo Expo 2011: Power for Land, Sea and Air*, GT2011-46541.

[12] L. Piegl, *On NURBS: a Survey*. Int. IEEE Computer Graphic and Application, 55 - 71, 1991.

[13] L. Hascoët, V. Pascual, Tapenade 2.1 users guide. *Technical Report 0300*, INRIA, 2004.

[14] S. Xu, D. Radford, M. Meyer, J.-D. Müller, Stabilisation of discrete steady adjoint solvers. *Journal of Computational Physics*, 99: 175 – 195, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A CONTINUOUS ADJOINT APPROACH FOR VEHICLE INTERIOR NOISE REDUCTION

Christos Kapellos[1] **and Michael Hartmann**[1]

[1]Volkswagen AG, Group Research, CAE Methods
Letter Box 1777, D-38436, Wolfsburg, Germany
e-mail: {christos.kapellos,michael.hartmann2}@volkswagen.de

Keywords: continuous adjoint method, vibroacoustics, automotive noise reduction

Abstract. *In this paper the continuous adjoint method is developed for a vibroacoustic model that predicts the interior noise of a vehicle induced by the airflow. The model simulates the front side window vibration, excited by the acoustic and hydrodynamic pressure load, and the resulting sound wave propagation into the cabin. Targeting interior noise reduction, the continuous adjoint formulation is used to derive the adjoint to the state equations, namely the bending wave and the wave equations, whilst taking into consideration their coupling. The developed method is applied to a generic vehicle test body for the minimisation of the interior noise.*

1 INTRODUCTION

When a vehicle travels at high speeds wind noise dominates the noise perceived in the passenger cabin. The turbulent flow field around the A-pillar and side mirror induce flunctuating acoustic and hydrodynamic pressure load on the side window. The vibrational response in form of bending waves of the window generates in turn sound waves which propagate to the interior and contribute to the overall sound pressure level.

In [1], the external flow generated sound and its propagation into the interior of a model passenger vehicle was numerically and experimentally investigated. The vehicle model, which is also used in this paper, was a SAE Type 4 car body and was built in such a way that only sound transmission through the front side and only window is allowed. In [2] a vibroacoustic model was developed to simulate the window vibration induced by pressure load and afterwards the sound propagation in the interior of the same vehicle.

In this paper the continuous adjoint method is developed for the aforementioned vibroacoustic model, in order to compute gradients of functionals related to the interior noise of a vehicle. Continuous and discrete adjoint methods [3, 4] have been widely used to solve optimisation problems in various research fields, due to their advantage of computing cost function gradients at a computational cost nearly independent of the number of the design variables. In the continuous formulation, the adjoint equations and sensitivity derivatives are derived by differentiating the objective function, augmented by the field and time integral of the product of the state equations and the adjoint variables. The adjoint equations are discretised and numerically solved to compute the adjoint fields and, through them, the sensitivity derivaties. In contrast, the discrete adjoint equations are derived directly from the discretised state equations. This paper is dealing exclusively with the continuous adjoint method.

2 VIBROACOUSTIC MODEL FOR THE INTERIOR NOISE OF A VEHICLE

The governing equation of the window deflection driven by a given pressure load is the two-dimensional bending wave equation,

$$R^w = \frac{\partial^2 w}{\partial t^2} + \frac{D}{m'}\nabla^4 w + \eta_1 \frac{D}{m'}\frac{\partial}{\partial t}\nabla^4 w + \eta_2 \frac{\partial w}{\partial t} + \eta_3 \sqrt{\frac{D}{m'}}\frac{\partial}{\partial t}\nabla^2 w - \frac{p'}{m'} = 0 \tag{1}$$

where w is the window deflection, p' the pressure load, D the window bending stiffness, m' the normalised window mass and η_i the damping coefficients. The initial and boundary conditions for a pinned at its edges window are

$$w = 0 \quad \text{and} \quad \frac{\partial w}{\partial t} = 0 \ \text{ for } t = 0s \text{ and } \mathbf{x} \in S_{wi} \tag{2}$$

$$w = 0 \quad \text{and} \quad \nabla^2 w = 0 \ \text{ for } \mathbf{x} \in \partial S_{wi} \text{ and } t > 0s \tag{3}$$

where S_{wi} is the window surface.

The window vibration generates sound waves which propagate to the vehicle interior. The sound radiation is governed by the wave equation,

$$R^p = \frac{\partial^2 p}{\partial t^2} - c^2 \nabla^2 p = 0 \tag{4}$$

The initial and boundary conditions are

$$p = 0 \text{ and } \frac{\partial p}{\partial t} = 0 \quad \text{for} \quad t = 0s \text{ and } \mathbf{x} \in \Omega \tag{5}$$

$$\nabla p \cdot \mathbf{n} = \rho_0 \frac{\partial^2 w}{\partial t^2} \quad \text{for} \quad \mathbf{x} \in S_{wi} \subset \partial\Omega \text{ and } t > 0s \tag{6}$$

$$\nabla p \cdot \mathbf{n} = -\frac{1}{c}\frac{\partial p}{\partial t} \quad \text{for} \quad \mathbf{x} \in S_{wa} \subset \partial\Omega \text{ and } t > 0s \tag{7}$$

where Ω is the interior domain and $\partial\Omega = S_{wi} \cup S_{wa}$. On the window S_{wi}, the acoustic pressure is coupled with the window accelaration. At any other interior wall the propagated sound waves are reflected and a convective boundary condition is imposed.

The equations are programmed and solved in OpenFOAM, version 1.6-ext. For the solution of the bending wave equation a fourth order scheme for gradients, a first order scheme for second order time derivatives and an explicit time marching scheme were used. Moreover, in order to take into account the difference of propagation speeds of waves in the window and in the air, the bending wave equation is solved with a smaller timestep.

3 THE CONTINUOUS ADJOINT FORMULATION FOR THE VIBROACOUSTIC MODEL

The objective function to be minimised is defined as the integral over a time interval T of the squared pressure at a specific point P

$$J = \frac{1}{T}\int_T p_P^2 \, dt \tag{8}$$

The implentation of more sophisticated objective functions, such as functionals in frequency domain, is possible. However, without loss of generality, a simplified objective function was chosen here so as to clearly focus on the results of the developed method.

The augmented function L is defined as the sum of J and the time-space integrals of the product of the state equations and the corresponding adjoint variables. So,

$$L = J + \int_T \int_{S_{wi}} z R^w dS dt + \int_T \int_\Omega q R^p dV dt \tag{9}$$

where z and q are the adjoint deflection and pressure respectively. Next step is to compute the variation of L

$$\delta L = \delta J + \int_T \int_{S_{wi}} z \delta R^w dS dt + \int_T \int_\Omega q \delta R^p dV dt \tag{10}$$

The development using the Gauss divergence theorem of steady and unsteady terms similar to the ones of eq. 10 has been thoroughly covered in literature [5], and is therefore omitted here. Due to the self-adjoint nature of the state equations, the adjoint equations will have the same differential operators, now applied on the adjoint variables, apart from the first order derivatives, which will appear with the opposite sign in the adjoint equations.

More focus is laid upon the the derivation of the boundary conditions and sensitivity derivatives on the window and how the coupling between the two adjoint equations is achieved. The surface integrals expressed on the window derived by the expansion of $\int_T \int_\Omega q \delta R^p dV dt$ read

$$c^2 \int_T \int_{S_{wi}} \left[\nabla q \cdot \mathbf{n} \, \delta p - q \, \delta \left(\nabla p \cdot \mathbf{n} \right) \right] dS dt \tag{11}$$

The first term is set to zero, by imposing a zero Neumann boundary condition for adjoint pressure q on the window. Using eq. 6 , the second term expands further to

$$\rho_0\, c^2 \int_T \int_{S_{wi}} -q \left(\frac{\partial^2 \delta w}{\partial t^2} \right) dS dt = \rho_0\, c^2 \int_{S_{wi}} \left[-q \frac{\partial \delta w}{\partial t} + \frac{\partial q}{\partial t} \delta w \right]_{T_{start}}^{T_{end}} dS + \rho_0\, c^2 \int_T \int_{S_{wi}} -\frac{\partial^2 q}{\partial t^2} \delta w \, dS dt \tag{12}$$

All terms of eq. 12 include now the total variation of the deflection w. The first integral gives the initial conditions for the adjoint wave equation, while the second one will be the source term for the adjoint bending wave equation.

The **adjoint equations** and their respective **boundary conditions** eventually read

$$R^q = \frac{\partial^2 q}{\partial t^2} - c^2 \nabla^2 q + 2 p_P = 0 \tag{13}$$

with

$$q = 0 \text{ and } \frac{\partial q}{\partial t} = 0 \quad \text{for} \quad t = T_{end} \text{ and } \mathbf{x} \in \Omega \tag{14}$$

$$\nabla q \cdot \mathbf{n} = \frac{1}{c} \frac{\partial q}{\partial t} \quad \text{for} \quad \mathbf{x} \in S_{wa} \subset \partial\Omega \text{ and } t < T_{end} \tag{15}$$

$$\nabla q \cdot \mathbf{n} = 0 \quad \text{for} \quad \mathbf{x} \in S_{wi} \subset \partial\Omega \text{ and } t < T_{end} \tag{16}$$

and

$$R^z = \frac{\partial^2 z}{\partial t^2} + \frac{D}{m'} \nabla^4 z - \eta_1 \frac{D}{m'} \frac{\partial}{\partial t} \nabla^4 z - \eta_2 \frac{\partial z}{\partial t} - \eta_3 \sqrt{\frac{D}{m'}} \frac{\partial}{\partial t} \nabla^2 z - c^2 \rho \frac{\partial^2 q}{\partial t^2} = 0 \tag{17}$$

with

$$z = 0 \quad \text{and} \quad \frac{\partial z}{\partial t} = 0 \text{ for } t = T_{end} \text{ and } \mathbf{x} \in S_{wi} \tag{18}$$

$$z = 0 \quad \text{and} \quad \nabla^2 z = 0 \text{ for } \mathbf{x} \in \partial S_{wi} \text{ and } t < T_{end} \tag{19}$$

The **sensitivity derivatives** expression of the objective function with respect to any design variables b_m which affect the input pressure load is given by

$$\frac{\delta J}{\delta b_m} = - \int_T \int_{S_{wi}} \frac{z}{m'} \frac{\delta p'}{\delta b_m} dS dt \tag{20}$$

In the adjoint chain, the procedure is performed backwards in time and in reverse order. The variation of the objective function acts as a monopole source for the adjoint wave equation and the information propagates to the interior. Then, the second derivative of the adjoint pressure on the window is used as a source term for the adjoint bending wave equation. In the end, the sensitivity derivatives computed at the window express the change in the objective function induced by a change in pressure load p'.

In case there is a closed-form expression for p', as in the test case shown in section 5, the gradient of the pressure load can be analytically calculated. Otherwise, when p' is derived from the solution of another set of equations, the right hand side term in eq. 20 serves as a boundary condition to its corresponding adjoint. For example, in case p' is the hydrodynamic pressure computed by solving the incompressible Navier-Stokes equations, the integrad of 20 is used for imposing the boundary condition for the adjoint Navier-Stokes on the window.

4 VALIDATION OF THE METHOD

The accuracy of the adjoint solution was validated first for each equation separately and then for the coupled problem, thus for three different cases. The vehicle model used is the SAE Body, seen in fig. 1.

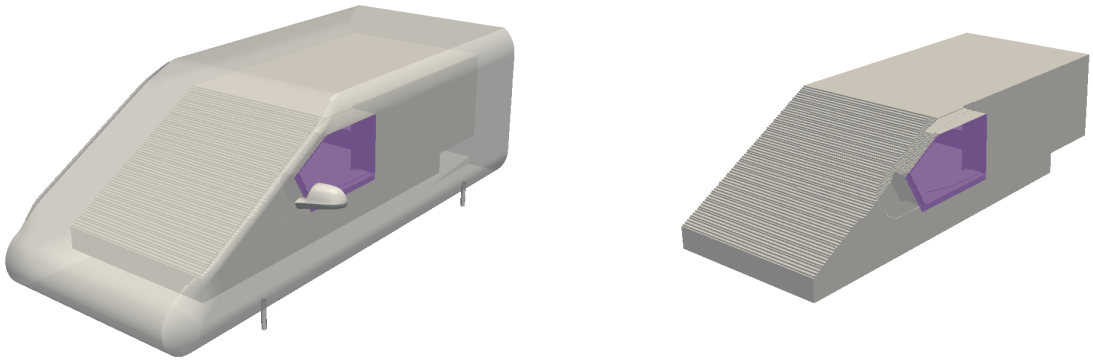


Figure 1: SAE Type 4 geometry: Exterior (left) and interior (right) vehicle walls. The only window of the vehicle, through which sound is transmitted to the interior, is coloured in magenta.

For the window vibration a plane wave excitation is investigated, targeting at minimising the time averaged square of the deflection,

$$p'_{vib} = A\sin(k_x x - \omega t) \text{ and } J_{vib} = \frac{1}{T}\int_T \int_{S_{wi}} w^2 dSdt \tag{21}$$

For sound radiation in the interior a plane wave profile is given for w in eq. 6 , targeting at minimising the objective function in 8.

$$w_{prop} = A\sin(k_x x - \omega t) \text{ and } J_{prop} = \frac{1}{T}\int_T \int_{S_{wi}} p_P^2 dSdt \tag{22}$$

For the coupled problem the plane wave excitation of eq. 21 was used.

$$p'_{coupl} = A\sin(k_x x - \omega t) \text{ and } J_{coupl} = \frac{1}{T}\int_T \int_{S_{wi}} p_P^2 dSdt \tag{23}$$

The gradient of the respective objective function for each of the three aforementioned testcases with respect to the input amplitude were computed with the continuous adjoint method and compared with finite differences.

As seen in table 1, the continuous adjoint method computes the gradient of the objective function with sufficient accuracy, since at most occasions the relative error is lower than 1%. In the sound radiation case, the higher error can be explained by taking into consideration the boundary condition on the window. Since the given deflection w_{prop} does not fulfil a zero Dirichlet condition on the window edges, a discontinuity arises there, which leads to higher numerical errors. However, the high accuracy computation in the coupled problem means that the gradient can be confidently used in a gradient-based optimisation.

	Window Vibration	Sound propagation	Coupled Problem
Continuous Adjoint	2.4391E-15	6.236	3.115E-10
Finite Differences	2.4394E-15	6.486	3.085E-10
Relative Error (%)	0.013	3.87	-0.964

Table 1: Comparison between objective function gradient values computed with the continuous adjoint method and finite differences.

5 MINIMISING THE INTERIOR NOISE OF THE SAE BODY

The developed method was applied for the minimisation of the perceived noise near the driver's ear, in the SAE body. To keep a focus only on the vibroacoustic phenomena and their adjoint, the pressure load on the window was induced by the sound radiation from a dipole source in a moving medium, rather than by a fully developed turbulent flow. The induced pressure is given analytically by

$$p'(\mathbf{x},t) = -\rho_0 \left[\frac{\partial}{\partial t} + U_0 \nabla\right] \phi \tag{24}$$

where ρ_0 is the air density, U_0 the vehicle velocity and ϕ the acoustic potential

$$\phi(\mathbf{x},t) = \nabla \left\{ \frac{A}{4\pi R^*} e^{i\omega\left(t-\frac{R^+}{c_0}\right)} \right\} \mathbf{d} \tag{25}$$

with amplitude A and directivity vector $\mathbf{d}$. Variables R^+ and R^* are related to the distance between the dipole and each receiver point where the induced pressure is calculated [6], namely each computational face of the window, S_{wi}.

Optimising with respect to its position, the dipole source is only allowed to move on the surface of an ellipsoid resembling a vehicle mirror, as seen in fig. 3. Moreover, its directivity is defined by the normal to the surface vector. The design variables of the optimisation problem are variables u and v which parameterise the ellipsoid surface by

$$\begin{aligned} X_{dip} &= X_c + a\cos u \cos v \\ Y_{dip} &= Y_c + b\cos u \sin v \\ Z_{dip} &= Z_c + c\sin u \end{aligned} \tag{26}$$

with $-\frac{\pi}{2} \leq u \leq \frac{\pi}{2}$ and $-\pi \leq v \leq \pi$

By differentiating eqs. 24 ,25 and 26, a closed form expression is derived for the derivative of the pressure load with respect to the design variables in expression 20.

A linesearch strategy was used for updating the design variables, starting from the initial value $\mathbf{b}^{init} = (0\ 0)$. The quasi-Newton BFGS method was used for computing the search direction and an interpolation based algorithm for choosing an appropriate step length [7]. In total 4 optimisation cycles were performed and the objective function was reduced by 34.3%. As seen in fig. 4, the dipole source is moved by the optimisation algorithm in a way that the induced pressure amplitude is reduced and that less area is affected by the dipole source. The first is achieved by increasing the distance between the source and the window and the second by pointing the directivity parallel to the z-axis.

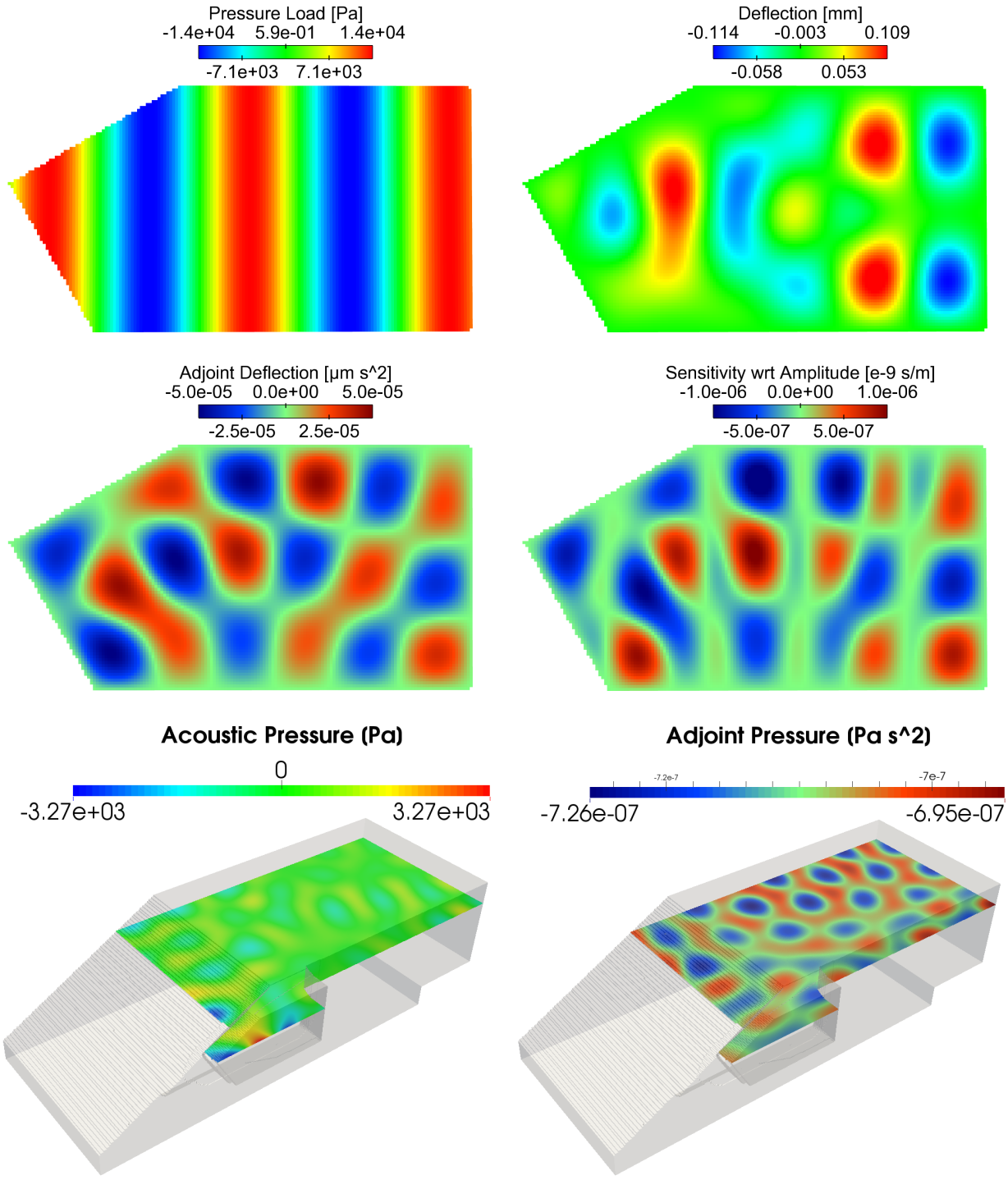


Figure 2: Validation of the adjoint solution. Bending wave equation: Instantaneous fields of the input pressure load p'_{vib} (top left), window deflection w (top right), adjoint deflection z (middle left) and sensitivity derivative w.r.t. input amplitude (middle right). Wave equation: Instantaneous fields of pressure (bottom left) and adjoint pressure (bottom right).

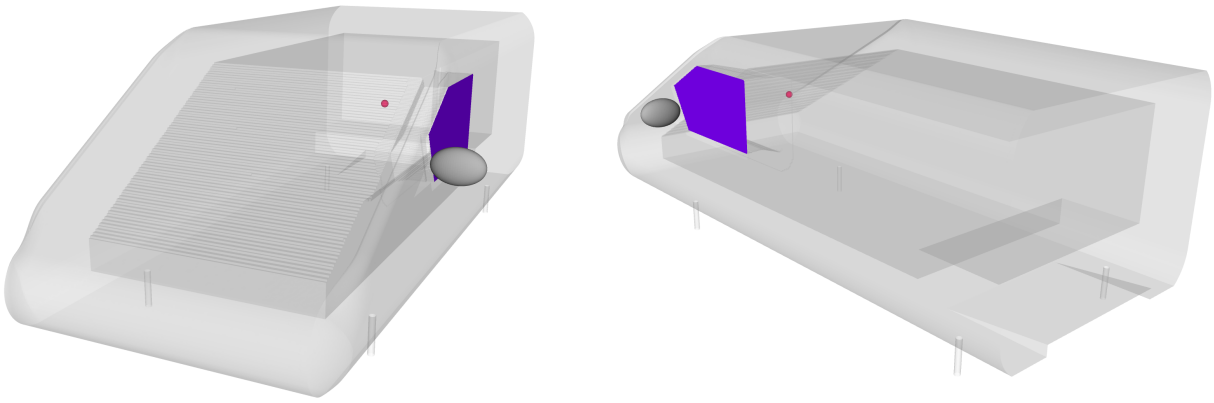


Figure 3: Minimising the interior noise of the SAE body: Pressure load on the window is induced by a dipole source which moves on the surface of the ellipsoid resembling a mirror, with directivity normal to the surface. The optimisation loop will find the optimal position of the dipole source, so that the perceived noise at the point near the driver's ear (red sphere) is reduced.

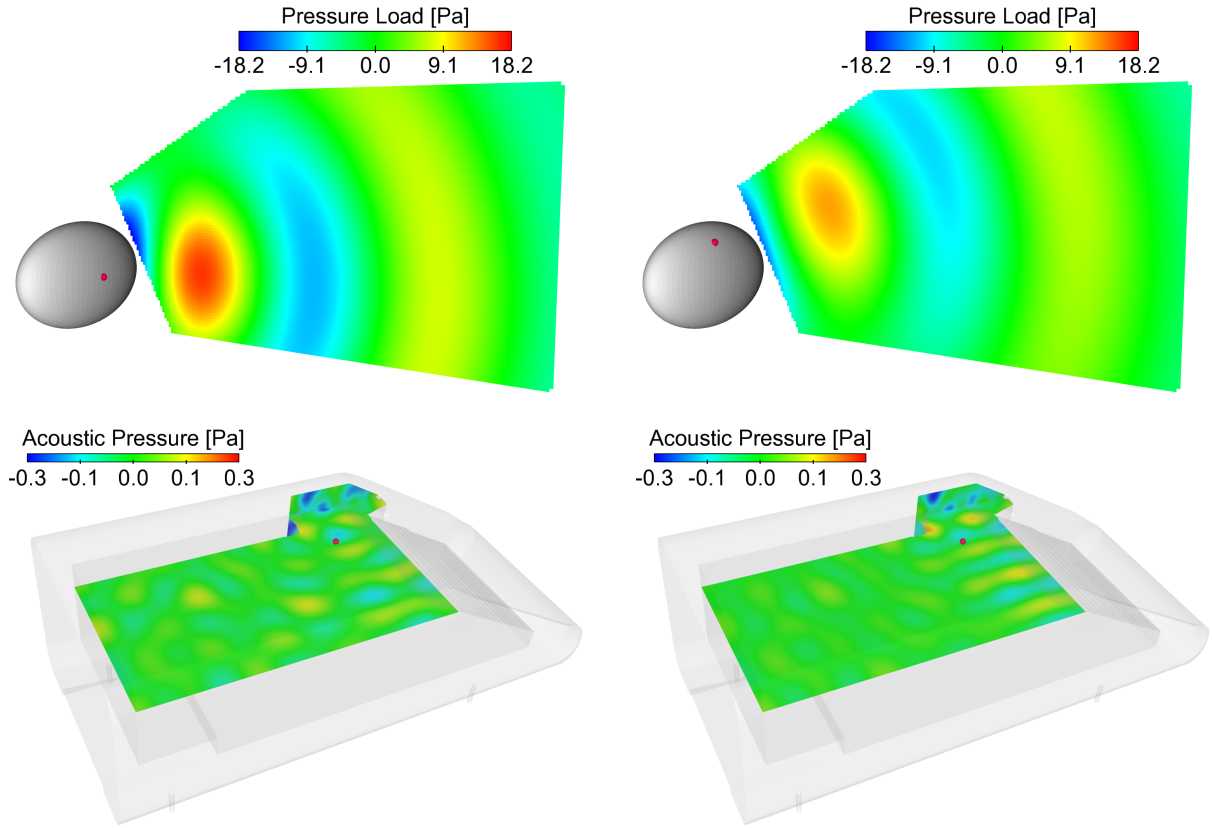


Figure 4: Minimising the interior noise of the SAE body: Comparison between the starting (left) and the optimised (right) setup. The instantaneous pressure load p' (top row) induced by the dipole source (red sphere) has reduced amplitude and acts on a smaller area of the window in the optimised case. The amplitude of the acoustic pressure in the interior computed at a time instance (bottom row) is also reduced for the optimal position of the source.

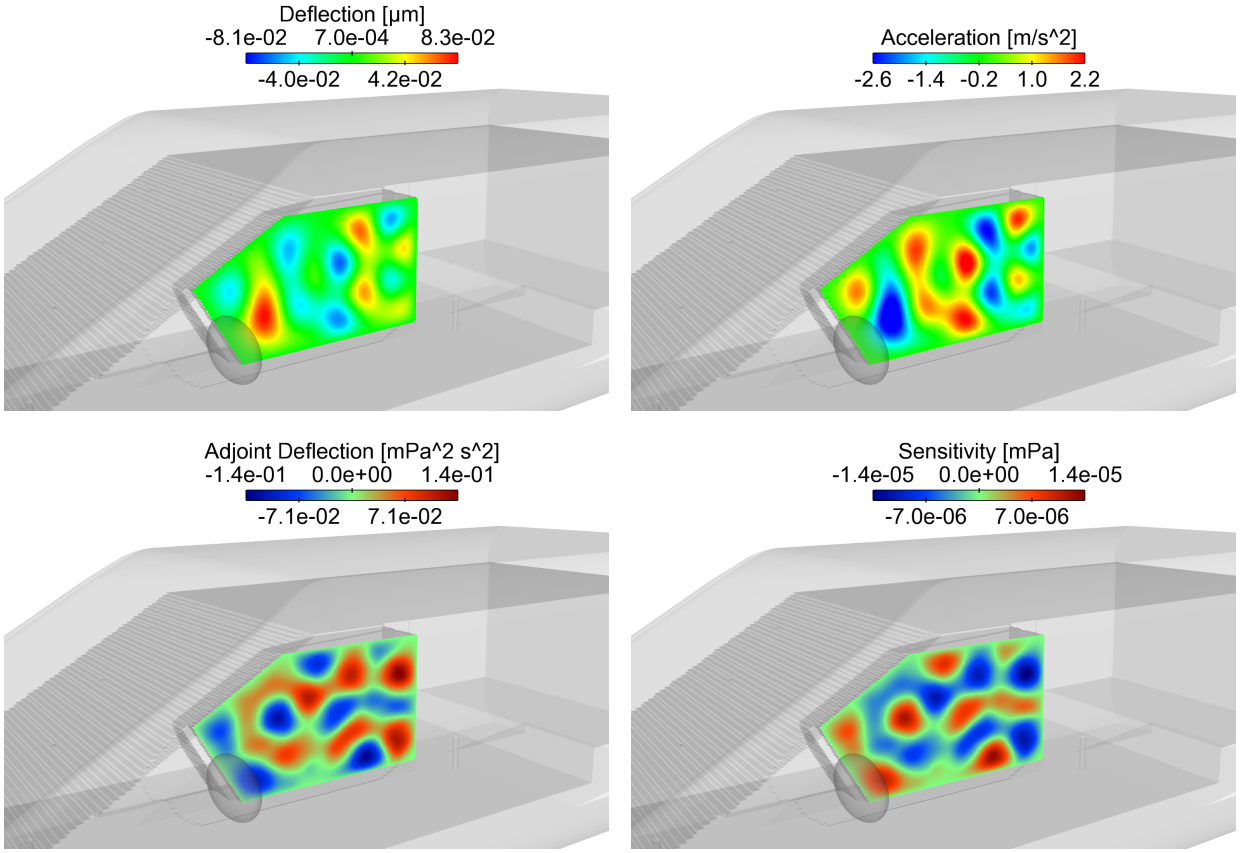


Figure 5: Minimising the interior noise of the SAE body: Instantaneous fields of deflection (top left) and acceleration (top right) of the window, as well as adjoint deflection (bottom left) and sensitivity (bottom right) computed at the first optimisation cycle.

6 CONCLUSIONS

The continuous adjoint method was formulated for a vibroacoustic model, which predicts the vehicle interior noise induced by a pressure load on the side window. The primal equations include the bending wave equation, which simulates the vibrational response of the window to the pressure load, and the wave equation, which describes the sound radiation to the interior. The adjoint to the state equations and the corresponding boundary conditions which couple the two equations were derived.

The method was validated with finite differences and applied to interior noise reduction in a generic vehicle model, the SAE body. In this case, the pressure load was induced by the sound radiation from a dipole source located at the area of the side mirror and the optimisation algorithm searched for the optimal position of the source. It was shown that the developed method computes gradients with sufficient accuracy and therefore can be integrated to larger aeroacoustic chains, in which the pressure load on the window is computed by solving different sets of equations, such as the Navier-Stokes equations.

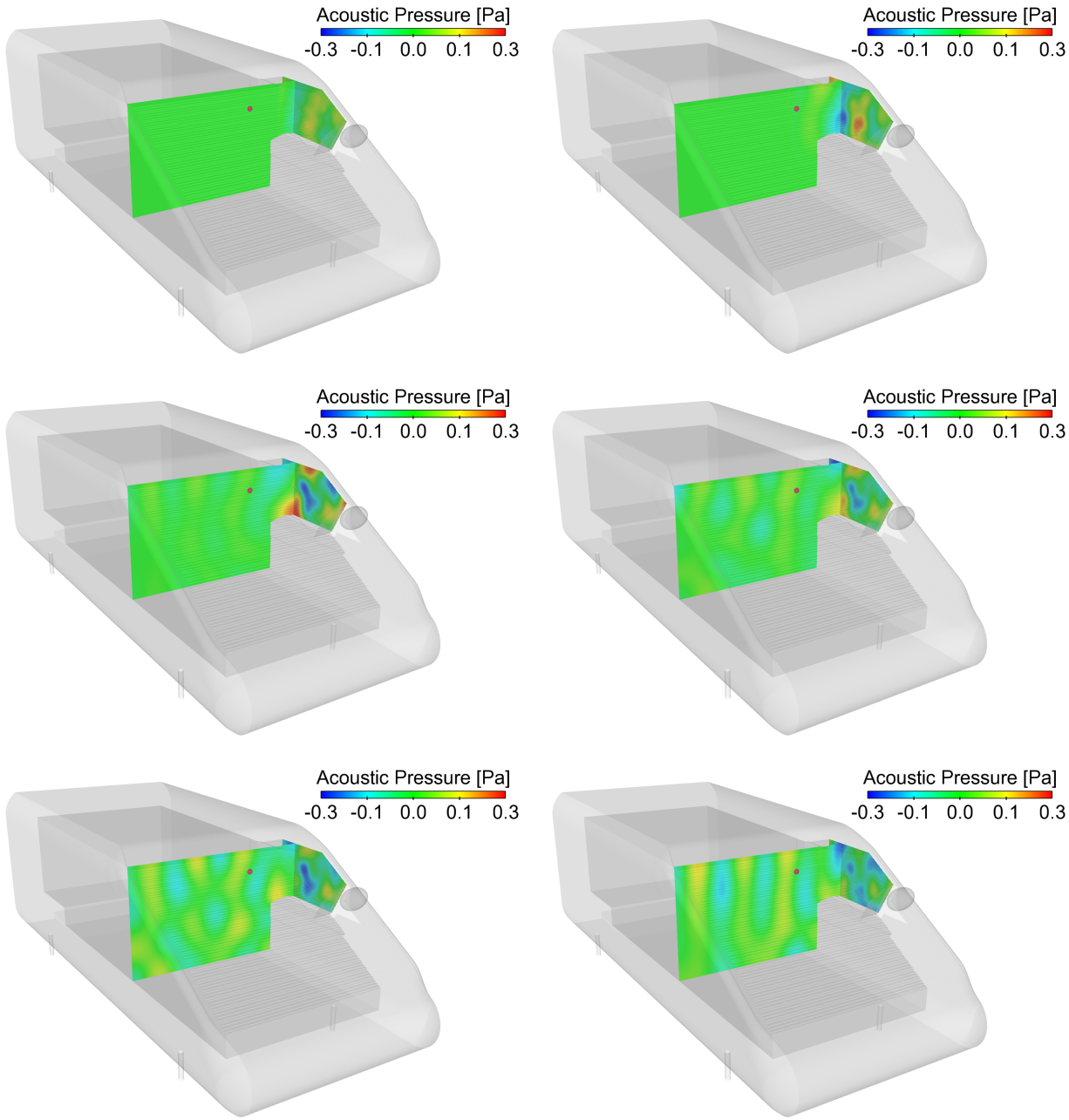


Figure 6: Minimising the interior noise of the SAE body: Six time snapshots during the sound radiation in the interior, computed at the first optimisation cycle. Sound waves generated by the window vibration propagate to the interior and are partly reflected by the vehicle interior walls.

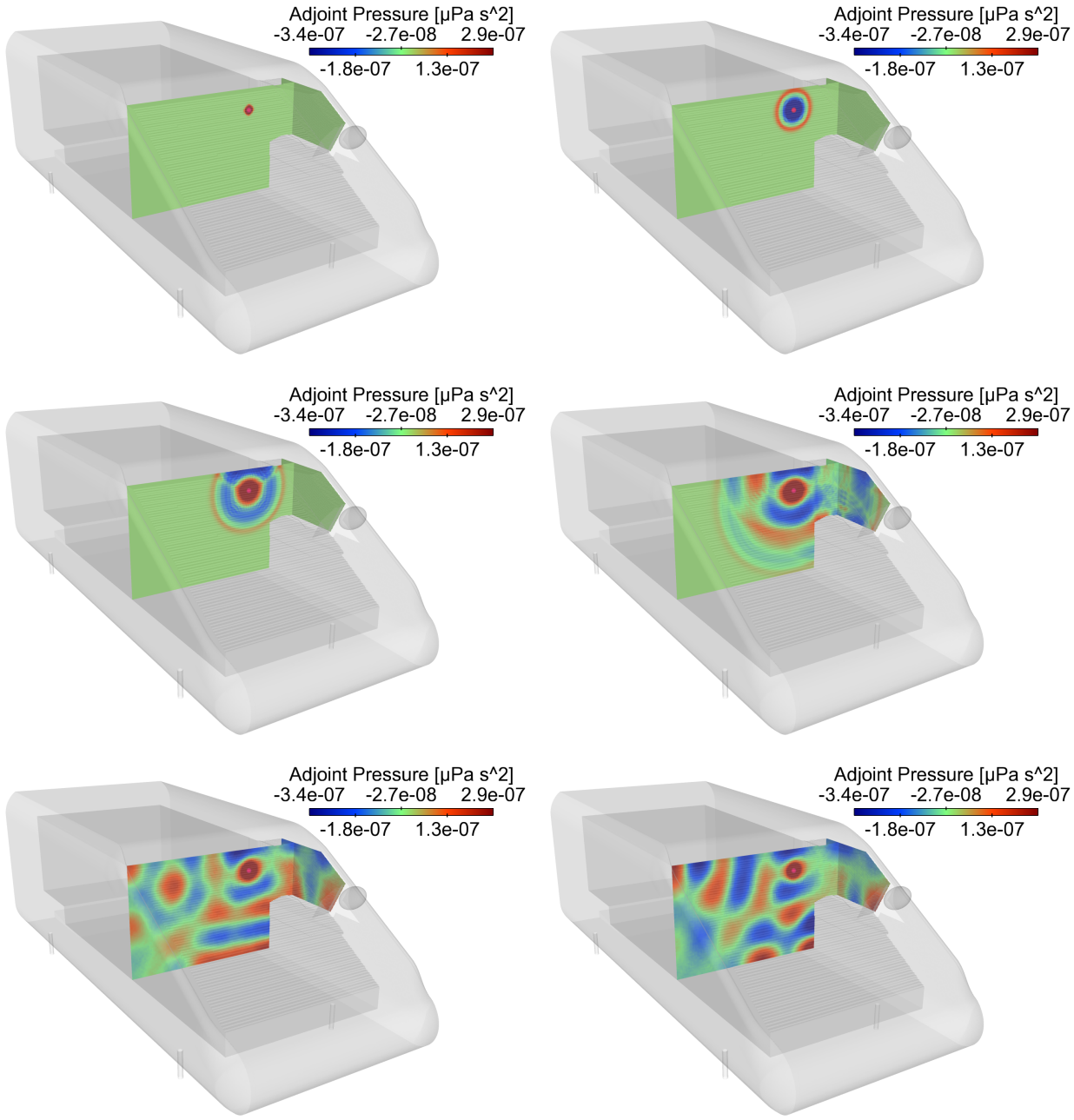


Figure 7: Minimising the interior noise of the SAE body: Six time snapshots of the computed adjoint pressure. The point where the objective function is defined (red sphere) acts as a monopole source for the adjoint wave equation. The adjoint pressure flunctuations propagate to the interior and reach the window. The second time derivative of the presented field at the window is the source for the adjoint bending wave equation.

REFERENCES

[1] M. Hartmann, J. Ocker, T. Lemke, A. Mutzke, C. Schwarz, H. Tokuno, R. Toppinga, P. Unterlechner, G. Wickern, Wind Noise caused by the A-pillar and the Side Mirror flow of a Generic Vehicle Model. *18th AIAA/CEAS Aeroacoustics Conference (33rd AIAA Aeroacoustics Conference)*, 2012, Colorado Springs, USA

[2] Alexander Kabat vel Job, Analysis of the vibrational response of a non-laminated glass plate for predicting the aeroacoustic performance in time domain. *The 2nd OpenFOAM User Conference 2014*, Berlin Germany

[3] S. Nadarajah, A. Jameson, Studies of the continuous and discrete adjoint approaches to viscous automatic aerodynamic shape optimisation. *AIAA Paper*, 25(30), 2001

[4] M. Giles, N. Pierce, An introduction to the adjoint approach to design. *Flow, Turbulence and Combustion*, 65,393-415, 2000

[5] E. Papoutsis-Kiachagias, K. Giannakoglou, Continuous adjoint methods for turbulent flows, applied to shape and topology optimization: Industrial applications. *Archives of Computational Methods in Engineering*, 2014, 1-45

[6] A. Najafi-Yazdi, F. Bres, L. Mongeau, An acoustic analogy formulation for moving sources in uniformly moving media. *Proceedings of the Royal Society A*, 2010

[7] J. Nocedal, S. Wright, Numerical Optimization. *Springer*

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A MIXED OPERATOR OVERLOADING AND SOURCE TRANSFORMATION APPROACH FOR ADJOINT CFD COMPUTATION

Zahrasadat Dastouri[1], **Sinan M. Gezgin**[1], **Uwe Naumann**[1]

[1]LuFG Informatik 12
Software and Tools for Computational Engineering (STCE)
RWTH Aachen University, DE 52056 Aachen
(dastouri,gezgin)@stce.rwth-aachen.de

Keywords: Algorithmic Differentiation, Computational Fluid Dynamics, Operator Overloading, Source Transformation

Abstract. *Adjoint based calculation of sensitivities pertaining to a Computational Fluid Dynamics (CFD) Solver has proven to be a vital tool in tackling large scale problems often found in industrial applications. There are two basic approaches for applying Algorithmic Differentiation (AD) to a CFD solver, namely, Operator Overloading and Source Transformation. Unfortunately, in both cases, it is still necessary to invest a significant amount of manual coding in order to get to an application that performs acceptably in terms of memory consumption and runtime. In this paper reverse mode of AD has been applied to an unstructured pressure-based steady Navier-Stokes solver. We explore the feasibility of combining both kinds of AD approaches to show where and how the advantages of each method can be exploited in order to reach a suitable compromise between performance, simplicity and efficiency. Additionally, we propose a methodology to automate the implementation of an adjoint software in order to minimize work the developer must carry out to produce the desired derivative. We investigate the effectiveness of this approach for relevant flow test cases.*

1 INTRODUCTION

Algorithmic Differentiation (AD) [1, 2] employs the rules of differential calculus in an algorithmic manner to determine accurate derivatives of a function defined by computer programs. For a given implementation of the flow model $F : \mathbb{R}^n \to \mathbb{R}^m$ over a computational grid, the computer program is developed to simulate the functional dependence of one or more objectives $\mathbf{y} \in \mathbb{R}^m$ on a potentially large number of input variables $\mathbf{x} \in \mathbb{R}^n$ by simulation of:

$$F : \mathbb{R}^n \to \mathbb{R}^m, \quad \mathbf{y} = F(\mathbf{x}) \tag{1}$$

AD enables us to compute the corresponding derivatives $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ in forward (forward mode) or backward (reverse mode). For a given implementation of the primal function in Equation 1, the function $F_{(1)} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$, defined as

$$(\mathbf{y}, \mathbf{x}_{(1)}) = F_{(1)}(\mathbf{x}, \mathbf{y}_{(1)}) \equiv \left(F(\mathbf{x}), \left(\frac{d\mathbf{y}}{d\mathbf{x}} \right)^T \cdot \mathbf{y}_{(1)} \right) \tag{2}$$

is referred to as the *adjoint* model of F. The adjoint model implementation yields the objective $\mathbf{y}$ and the product $\mathbf{x}_{(1)} \equiv \nabla F(\mathbf{x})^T \cdot \mathbf{y}_{(1)}$ of its transposed jacobian at the current point $\mathbf{x} \in \mathbb{R}^n$.

There are two main methods for implementing AD: by source code transformation (S-T) or by use of derived data types and operator overloading (O-O). In O-O AD the code segments and arguments of the primal code are stored inside a memory structure called tape during the forward run of the primal. In reverse mode the stored values on the tape are interpreted to get the resulting adjoints , while the S-T approach parses the code at compile time and generates the actual derivative code.

Prior to this paper [3], [4], we have discussed the implementation of the AD tool `dco/fortran`[1] *(Derivative Code by Overloading in Fortran)* to an unstructured pressure-based steady Navier-Stokes solver. The solver is an incompressible flow solver with cell-centered storage and face-based residual assembly. It works on unstructured meshes with collocated variables and uses the SIMPLE [5] pressure correction algorithm in a pseudo time stepping scheme towards a steady solution. It is written in Fortran 90-95 (7,000 lines) as a test-bed for developing adjoint Navier-Stokes fields [6]. The case study that is used for flow model simulation and sensitivity studies is the S-bend channel flow case which is a simplified vehicle climatisation duct that is presented in [7]. We have addressed proper solution algorithms adapted to the code for the improvement of efficiency of the adjoint code by optimizing the checkpointing scheme for the iterative solver and development of symbolically differentiated of linear solver. In addition, we investigated the benefit of the reverse accumulation technique [8] for the fixed point iterative construction in the primal code.

In this paper we combine the flexibility and robustness of operator overloading with the efficiency of source transformation by coupling `dco/fortran` and `TAPENADE` [9]. The latter is used for the derivation of computationally expensive kernels. Our emphasis is to automate the implementation of an adjoint software to decrease the development time of the differentiated code. The overall design is presented in Sections 3 and 4 illustrating the transformation procedure. Numerical results demonstrate the computational efficiency in terms of memory

[1] developed at the institute a*Software and Tools for Computational Engineering* at RWTH Aachen University implementing AD by overloading in Fortran [1]

consumption and speed while preserving the flexibility of an O-O approach applied to a CFD code.

2 REVIEW OF AD TOOLS

Algorithmic Differentiation is generally implemented using one of two strategies: source code transformation or operator overloading. Here we present a brief review of both tools.

2.1 Operator overloading

Using AD by operator overloading to evaluate directional derivatives of an existing program is comparatively easy. The main idea is to replace all relevant floating point data types within a given program with corresponding enhanced types of the particular operator overloading tool. The enhanced data types then keep track of any *tangent*- or *adjoint*-information of the specific variables declared as floating point data types throughout the forward execution of the program. This is accomplished by defining special operators for all intrinsic functions like +, -, sin(), exp(), etc. , hence the name *operator overloading*. In case of a *tangent-linear* model derivative information is acquired alongside the forward execution of the program. Using the *adjoint* model this is not the case. When an intrinsic function is called during the forward execution of the program, the required information to compute the adjoint information of all involved active(i.e. tracked) variables does not yet exists. To show that, consider the assignment $y = x_1 \cdot x_2$. The adjoints for both inputs are $x_{(1),1} += y_{(1)} \cdot x_2$ and $x_{(1),2} += y_{(1)} \cdot x_1$ respectively. Unfortunately, the value for $y_{(1)}$ is not yet available when the assignment is executed. Therefore, the derivative information cannot be computed at that point in time. A two step approach to solve this problem first stores all required information for every operation during the forward execution of the program. The second step is running through the operations in reverse order and calculating adjoint values of its inputs (x_1, x_2) from adjoint values of its outputs $y_{(1)}$ which are now available. Storing of such information is usually realized by creating log entries on a stack also referred to as tape. In case of `dco/fortran`, an entry in such a tape stores a code for the type of operation, virtual addresses of operands x_1 and x_2 as well as floating point values for its value y and adjoint $y_{(1)}$. This method automatically fulfils requirements of *data-flow* and *control-flow-reversal* for an adjoint model without any extra effort by the tool or the developer. Understandably, such an approach may require a very large amount of available memory in order to store the entire tape for programs of even small to moderate complexity. Its big advantage is its ease of implementation. There are different O-O AD tools for Fortran programs such as ADF [10], ADOL-F [11], AUTO_DERIV [12], etc.

2.2 Source transformation

AD by source transformation analyses the source code of a particular function and produces a new source code which, when executed, calculates the original function value and its directional derivative. The source code for a function is replaced by an automatically generated source code that includes statements for calculating the derivatives interleaved with the original instructions. The advantage is that the resulting code can be compiled into an efficient program due to compile time optimizations having a greater effect. This is a valid approach for all programming languages. The drawback is that this is an enormous transformation that usually cannot be done by hand. Tools are needed to perform this transformation correctly and rapidly. For a *tangent-linear* model, the transformation is straight forward because the derivative statements are in the same order as the original statements. The adjoint model requires control flow reversal which

turns it into a challenging problem for S-T tool developers. Tapenade [9] engine is just one such Automatic Differentiation tool that uses source transformation.Other examples of S-T tools that provide automatic differentiation for Fortran programs are ADG [13], ADFOR [14], TAF [15], OPENAD [16].
The relative advantages and dis- advantages of using the two different tools are summarized in Table 1. The + sign indicates that the corresponding AD tool possesses the desired characteristic.

AD tool	O−O	S−T
Ease of implementation	+	−
Changing the original source code	+	−
Memory consumption	−	+
Compile time optimizations	−	+
Compatibility of numerical gradients with the discrete PDE	+	+
Flexibility to handle arbitrary functions	+	−
Robustness	+	−

Table 1: Comparison operator overloading and source transformation AD tools

3 MIXED APPROACH: SOURCE TRANSFORMATION VIA OPERATOR OVERLOADING TOOL

We apply S-T via O-O tool using a feature of `dco/fortran` called External Function. This interface provide adjoints of specific functions of a program to `dco/fortran` or `TAPENADE`. It is typically used to provide analytically derived adjoints for computationally intense kernels. Instead of recording the operations of such a kernel during the forward run a function pointer is registered on tape. It points to a function capable of providing the adjoint information of the kernel. Basically, if the adjoint model of a function f is known, this information can be used during the interpretation in order to remove the necessity to record f during the forward execution of the program. Depending on the complexity of f, this may have a drastic impact on memory consumption. The following small example shows how this feature is used in order to mix different approaches to adjoint computation.
Consider a function $\mathbf{z} = g(\mathbf{x}, p)$ with $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m$ which at any point during its execution calls $\mathbf{y} = f(\mathbf{x})$ (1). Let $f_{(1)}$, described in (2), be its known first order adjoint model and let P be a program that computes the adjoint $\mathbf{z}_{(1)}$ of g using `dco/fortran`. Including $f_{(1)}$ into P takes two function implementations. First, a function $f\_wrap(\mathbf{x})$ that hides the process described in figure 1 denoted by *X-Forward* from the rest of P. Second, a function f_ext which includes the steps *X-Reverse* and calls the desired implementation of $f_{(1)}$ in between steps 3 and 4.

1-Forward The state of every non-constant variable which is needed to evaluate f has to be saved. In addition, independent inputs for f are registered using the checkpointing features of `dco/fortran` in order to have the correct variable addresses available later in step 4.

2-Forward The output of f is saved and the implemented function $f_{(1)}$ is registered as the external function that is later called by the interpretation process to evaluate adjoints $\mathbf{x}_{(1)}$.

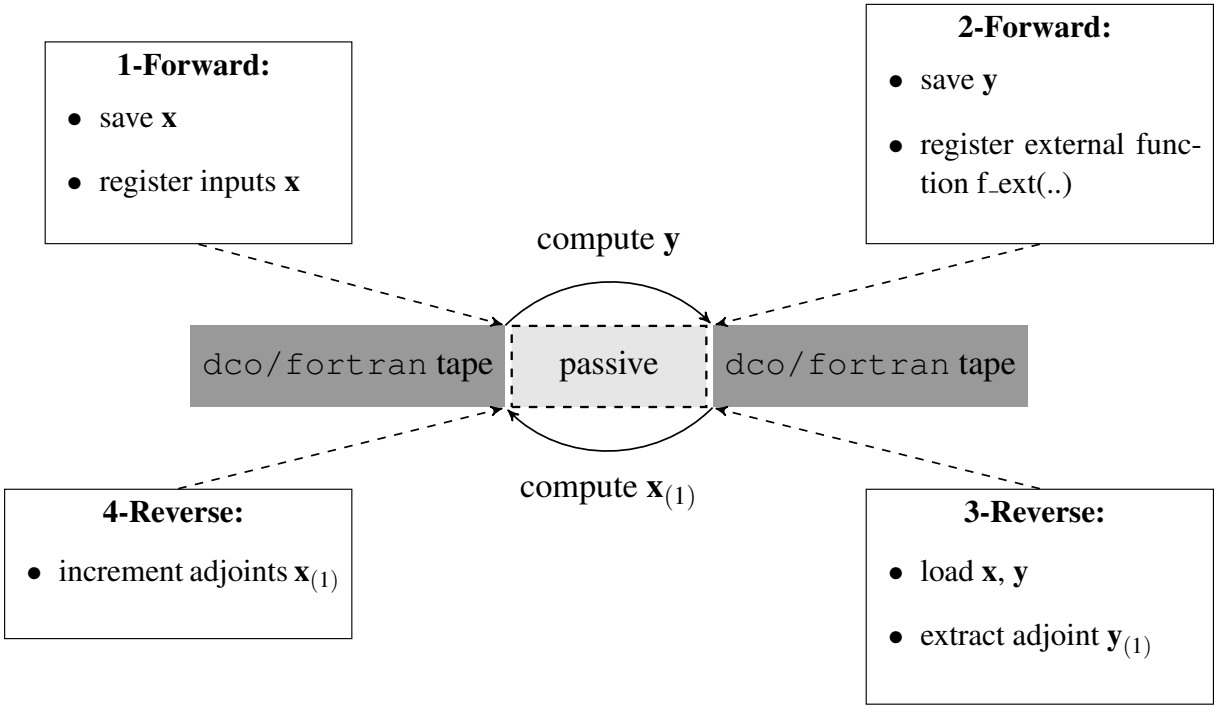


Figure 1: External Function Concept

3-Reverse The stored data is loaded and the adjoint $y_{(1)}$ is read from tape.

4-Reverse The computed adjoint information has to be added back into the tape at their corresponding positions. This is done by incrementing adjoint values of all registered variables in reverse order.

4 AUTOMATION METHODOLOGY

The idea of the automation process described here assumes that there already exists an overloaded program derived by using `dco/fortran`. This section lists steps that need to be taken in order to produce a combined (`dco/fortran` and `TAPENADE`) solution. The overview of the method is presented in figure 2. In addition, a basic approach for automatic source code generation of wrapper- and external functions is presented in 4.2.1 and 4.2.2 respectively. This section uses some abbreviations for the different versions of a function:

f	:	The original *overloaded* function of the black-box `dco/fortran` version
$f\_passive$	:	The original function of the original program
f_b	:	The adjoint function derived by `TAPENADE`
f_wrap	:	The wrapper function corresponding to f
f_ext	:	The function called by the interpreter to compute $f_{(1)}$

Table 2: Function abbreviations

4.1 Kernel identification

The first and vital step is to choose the correct functions for extraction. Using a profiling tool like gprof[17] or intel's vtune[18] is more than sufficient to provide a list of computationally

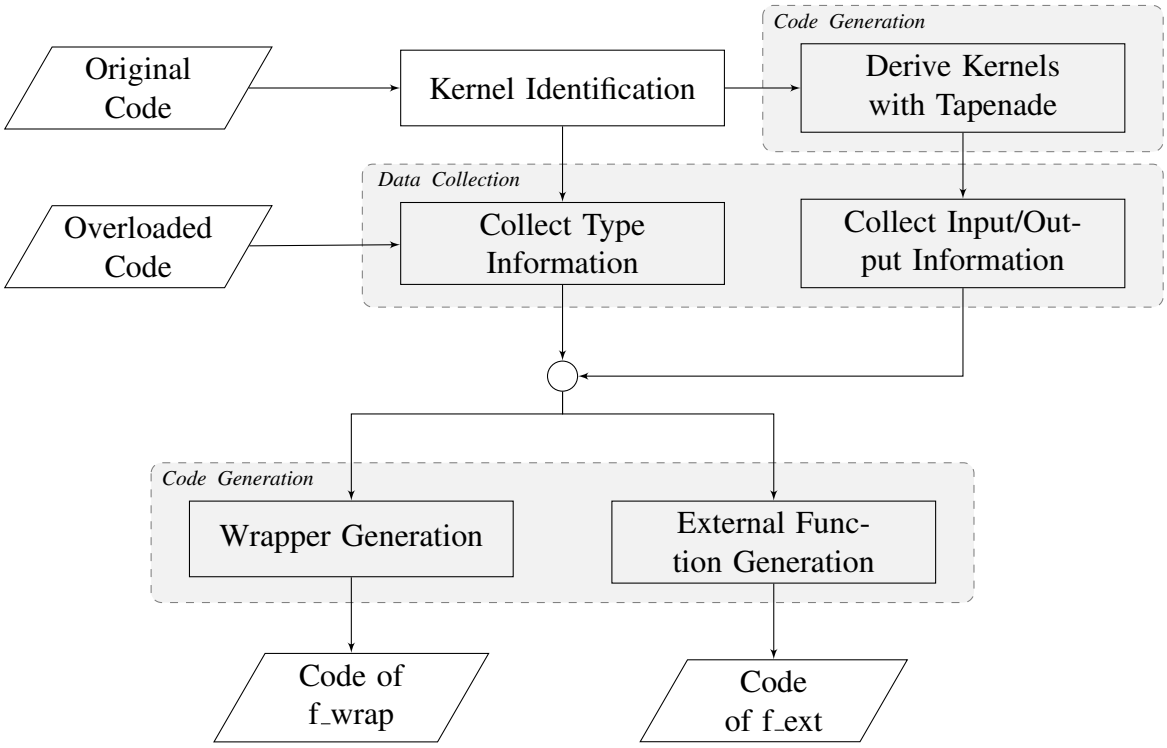


Figure 2: Automation of Source Code Generation

expensive functions. Another important information is how often those have been called during one execution of the original program. The reason for collecting that particular information is that using the external function feature involves storing/copying of data. This overhead may negate any benefit when the computational effort per function call is too small. Thus, the optimal candidates for extraction are functions with a significant self time and minimal number of calls. In addition, it is beneficial if such a function is low in the hierarchical calling order so as to avoid passing the entire program on to `TAPENADE`. Once such functions are identified they can be passed on to `TAPENADE` in order to generate the corresponding overloaded versions.

4.2 Code generation

This sections describes how the development of wrapper- and external functions can be automated. The focus is on source code generation. Not considered is the automated integration of newly created functions into a build system due to the variety in which such a system can be realized.

4.2.1 Wrapper generation

A Wrapper function *f_wrap* of an arbitrary function f has a fixed structure outlined in Algorithm 1. Due to this fixed structure, it is possible to define the complete source code of *f_wrap* using only a few additional informations about the arguments of f. Once all required information is collected a Python preprocessor could generate wrappers, register routines and replace the code. The source code is determined knowing the following information about each variable appearing in the argument list of the function f:

1. Input/Output/Parameter

2. Array Dimension(s)

3. Active/Basic/Derived Type

An active type is a specific derived type, namely `dco_a1s_type`. Knowing the input/output status of an active variable defines whether it needs to be registered as such. It also determines when its value has to be saved. Inputs and parameter values are saved before calling f whereas outputs naturally have to be saved after their evaluation. Array dimensions determine the loop structures encasing calls of functions like *register_input*. All basic types are classified as parameters by default and only their current values need to be stored. A special case occurs when processing derived types. This is due to the fact that a derived type may hold all kinds (input/output/parameter) of variables inside. Consequently, this information needs to be collected for all members of a derived type independently. In addition, a derived type may also have a member which is a derived type itself. A program/script capable of processing arbitrary types therefore has to have a part handling nested derived types. One way to realize such a program is to split it in two parts. The first part collects all required information and fills a data structure similar to Table 4 for each argument and each basic/active member variable of any derived types appearing in the argument list of f. Then, for each step of Algorithm 1 it is checked for each variable if code has to be added to the current section corresponding to the current variable. This can be done by adding references of all *var* structures into three lists, *inputs*, *outputs* and *parameters*, depending on their `var.intent` value.

Algorithm 1: wrapper function generation

Input :
- module dependency information of the overloaded adjoint function $f_{(1)}$
- three lists of var structs: *inputs*, *outputs*, *parameters*

Output:
- source code of function f_wrap

Algorithm:
generate function head
foreach *var* ***in*** *inputs* **do**
| generate block *register input*
end
foreach *var* ***in*** *inputs & parameters* **do**
| generate block *checkpoint*
end
generate call of original function $\mathbf{y} = f\_passive(\mathbf{x}, \mathbf{p})$
foreach *var* ***in*** *outputs* **do**
| generate block *checkpoint*
end
foreach *var* ***in*** *outputs* **do**
| generate block *register output*
end
generate closure statements

4.2.2 External function generation

Like the concept of the wrapper function, the external function also follows a specific sequence of tasks that need to be executed in a fixed order. The sequence is presented in algorithm 2. The information required to generate external function code is equal to that of a wrapper function when making one simplification. Every array is assumed to be *allocatable*, meaning that all data structures are allocated when loading the checkpointed data in *f_ext*. Not having this assumption just means adding a allocatable flag to the *Var* structure and distinguishing the two cases during code generation.

Algorithm 2: external function generation

Input :
- module dependency information of *f_passive*
- three lists of var structs: *inputs*, *outputs*, *parameters*

Output:
- source code of function f_ext

Algorithm:
generate function head
foreach *var* ***in*** *inputs & parameters* **do**
| generate block *read checkpoint*
end
foreach *var* **do**
| generate block *allocate adjoint*
end
foreach *var* ***in*** *outputs* **do**
| generate block *get_input_adjoint*
end
generate call to f_b
foreach *var* ***in*** *inputs* **do**
| generate block *increment_adjoint*
end
generate closure statements

4.3 Data collection

The collection of necessary information can be done in many ways. The method proposed here consists of parsing the output of `gfortran -fdump-parse-tree` and the result of the data flow analysis of `TAPENADE`. The latter is necessary for the identification of in- and outputs of the function where as the former is used to fill data type and shape information. The `TAPENADE` data flow analysis can be found on top of the source code of the derived function. For example, let a function `f_passive(x,y)` implement $y = \sum_{i=1}^{n} x_i^2$. The corresponding analysis of `TAPENADE` produces the following comments on top of the derived function `f_b(x,xb,y,yb)`:

```
!  Differentiation of f_passive in reverse (adjoint) mode (with
   options noISIZE):
```

State	meaning of $var_{(1)}$ for f_b	meaning of *var* for f	code blocks for f_wrap
:zero	output, $var_{(1)} := 0$	parameter	save_in
:incr	input, $var_{(1)}$ incremented	input	rgstr_in, save_in
:in-zero	input, $var_{(1)} := 0$ at end	output	rgstr_out, save_out
:in-out	input & output	input & output	rgstr_in, rgstr_out, save_in
:out	output, $var_{(1)} := 0$ at start	input & output	rgstr_in, save_in

Table 3: Intent states of active variables for f_b

```
!   gradient     of useful results: x y
!   with respect to varying inputs: x y
!   RW status of diff variables: x:incr y:in-zero
```

The input/output status of an argument can be extracted from the *RW status of diff variables* line. Arguments appearing here can have the following states listed in Table 3. Arguments of *f* not appearing in this line are parameters and only need to be stored to ensure correct execution of `f_b`. If any variable appearing in that line is a member of a derived type the corresponding part denoting the *RW status* looks like

```
*(<var>.<member>):<state>.
```

A convenient solution to parse nested types is to construct each `var.name` like the string inside the round brackets. This way, it is easy to produce the Fortran source code accessing said member by just exchanging every "." with a "%".

5 NUMERICAL RESULTS

Numerical results are described for the S-bend test case in three dimensions for different problem sizes ranging from 47k up to 500k cells. Sensitivity analysis results of the adjoint code using operator overloading have been presented in earlier work [4] to calculate the sensitivity of the pressure loss objective function with respect to surface boundary nodes of the S-bend. These results have been obtained for the whole iteration process for the steady solver up to the convergence point of the flow solver. Applying only operator overloading zo the flow solver leads to significant memory usage presented in Figure 3 (-◇- on tape). The performance has been improved by symbolically differentiating the linear solver. However, this is only applicable to problems where the function to be differentiated is simple enough. Nevertheless, it is a very good baseline for performance measurements.

Here we apply the alternative approach presented in this paper by applying source transformation for repetitive parts of the solver where the operator overloading interconnect the upper level. Numerical results are verified and the cost is compared for most expensive repetitive part of the CFD solver.

Var	
name	string
type	string
rank	integer
shape	$[dim_1, dim_2, \ldots, dim_{rank}]$
intent	{'in','out', 'in-out', 'parameter'}

Table 4: Members of Var struct

Table 5: Verification of results with finite difference

version	error
`dco/fortran`	$4.622\mathbf{63870374081}E-008$
`dco/fortran` +`TAPENADE`	$4.622\mathbf{73583679148}E-008$

5.1 Verification

We verify the sensitivity results for adjoint solver with finite difference for all cells as presented in Equation 3. The value of step size increment is determined by smallest relative error between the derivatives of adjoint code and finite difference, see [4]. The error comparison in Table 5 show the accuracy of applying `dco/fortran` and `TAPENADE` tools for generating derivatives.

$$err = \sum_{i=1}^{n} | sen_{fd}(i) - sen(i)_{version} | \tag{3}$$

5.2 Test environment

The test case is a S-bend ventilation duct for incompressible steady-state flow as presented in [4] . The objective function computes the loss of pressure at the outlet of the channel and the derived program adds computation of sensitivities of surface nodes with respect to the objective function (pressure loss at outlet). All programs are compiled and run on a machine with hardware specifications shown in Table 6.

CPU	Intel(R) Xeon(R) CPU E5-2630
Memory	DDR3 @ 1600 Mhz
OS	Debian GNU/Linux 8 (jessie)
Fortan Compiler	GNU Fortran (Debian 4.9.2-10) 4.9.2
C++ Compiler	gcc (Debian 4.9.2-10) 4.9.2
S-T tool	`TAPENADE` 3.9

Table 6: Machine Specifications

5.3 Overloaded program variants

The process described in section 4 aims to combine the overloaded `dco/fortran` program with kernels derived by `TAPENADE`. The kernel identification exposed the linear solver as, once again, being the function doing most of the computation. Thus, it is chosen to be replaced using the process presented in sections 4.2.1 and 4.2.2. Two program variants are compiled which differ in one option for `TAPENADE`. The option is *-nocheckpoint* which splits the recording of the variable and control flow stacks from the computation of the adjoint values into two separate functions `f_fwd` and `f_bwd` [9]. This allows a more detailed performance comparison. As a baseline for performance the symbolic differentiation of a linear solver is used and incorporated using the external function feature producing a third program variant.

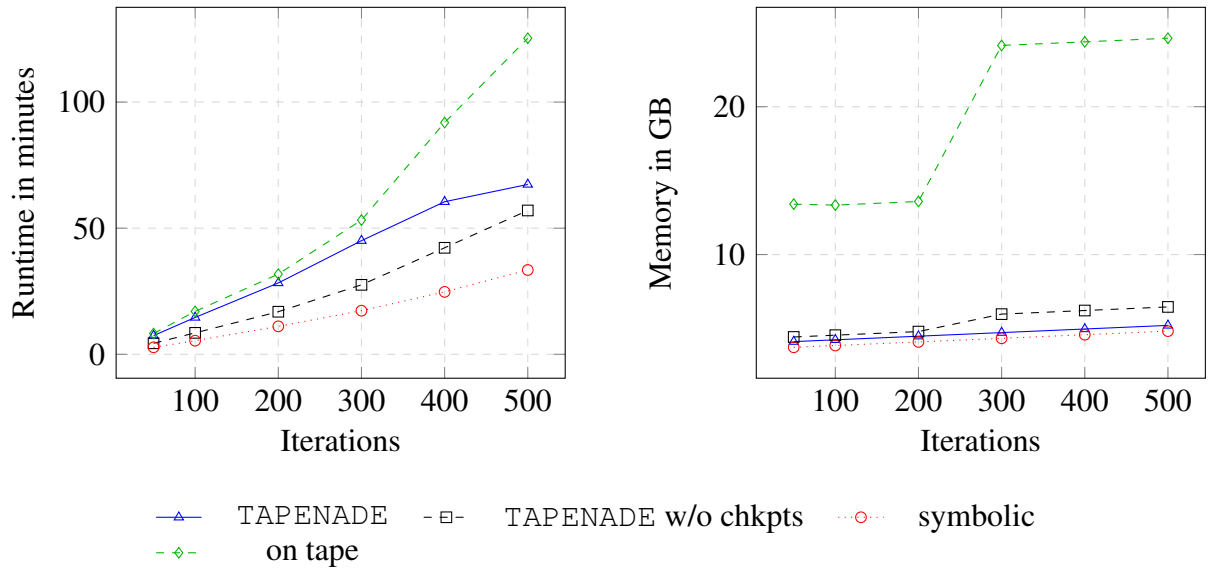


Figure 3: Runtime and Memory for ventilation duct with 47k elements

5.4 Performance

Figure 3,4 and 5 show performance comparison results for the above mentioned test case with 47k, 130k and 500k elements respectively. Runtime and peak memory requirement is measured for computation of sensitivities with a fixed iteration count for the forward computation. The blue solid line is the default variant combining `dco/fortran` and `TAPENADE`. Here, the function derived by `TAPENADE` reruns the forward evaluation of the linear solver during the interpretation of the `dco/fortran` tape. The black, dashed line represents the second combination where temporary data used by `TAPENADE` is acquired during the passive execution of the linear solver, i.e. in between steps 1-Forward and 2-Forward. This saves one linear equation solve during the interpretation of each outer iteration but adds to the peak memory requirement. The theoretical maximum memory requirement for this variant is

$$\text{peak mem} \approx tape + \sum_{i=1}^{n} stack_i \tag{4}$$

where *tape* represents the space needed to store the entire `dco/fortran` tape and $stack_i$ refers to the size of the temporary data of `TAPENADE` generated for calls of the linear solver during the ith outer iteration. The default variant is between two and three times slower than the symbolic variant for all three mesh configurations. As expected, the *-nocheckpoint* version(black, dashed line) is faster than the default variant but requires more space. In particular, for smaller test cases(47k) the peak memory depends more on the iteration count of the linear solver as shown in figure 3. Here, the number of iterations needed to meet a certain tolerance of the linear solve about triples after roughly 250 outer iterations. That is the reason for the sudden jump of the peak memory for the variant combining `dco/fortran` and `TAPENADE` without checkpoints. A similiar but much more pronounced, jump can be observed for the O-O only case. The configuration with 130k elements has a constant high number of iterations of the linear solver and thus does not show a similar jump.

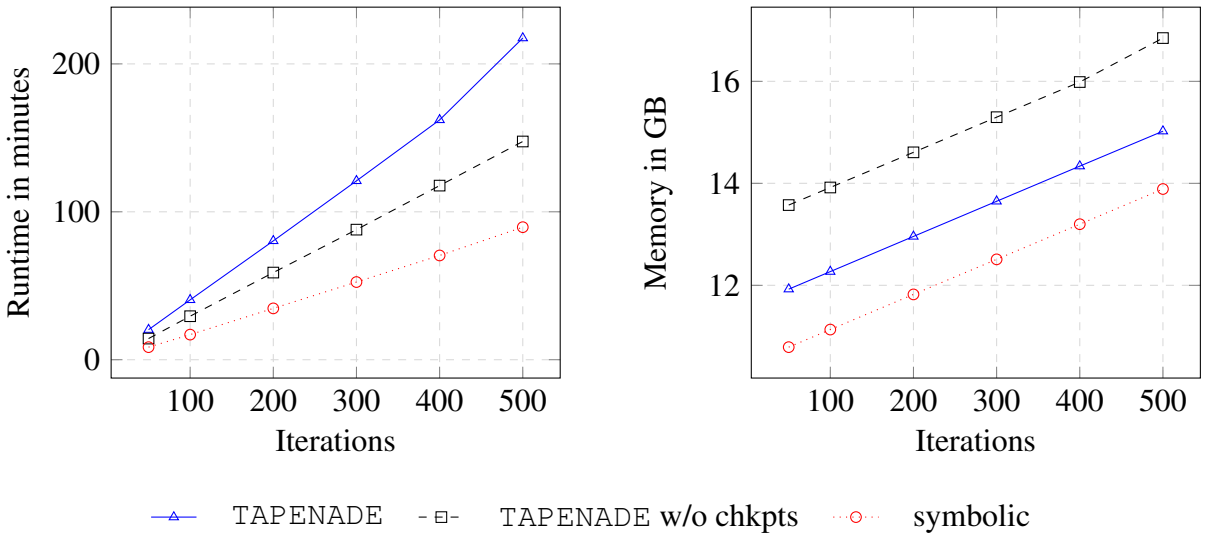


Figure 4: Runtime and Memory for ventilation duct with 130k elements

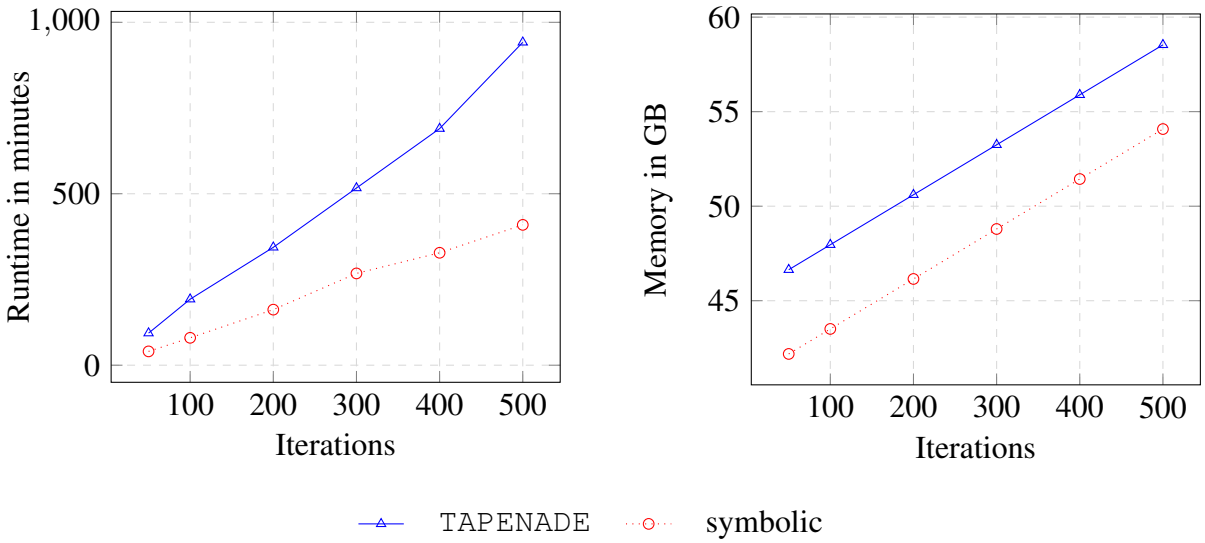


Figure 5: Runtime and Memory for ventilation duct with 500k elements

6 CONCLUSIONS

The method described in this paper employs a source transformation via operator overloading approach such that the resulting derivative code computes the sensitivity results of the given problem in CFD solver. The key aspect of this method is a combination of source-to-source transformation and operator overloading approaches. The numerical experiments are tested for expensive repetitive part of the solver demonstrating significant memory reduction of the resulting combined code compared to the overloaded adjoint version. In terms of memory usage, these results are competitive with symbolic differentiation of a given function. It could be used as a lower bound to decide whether developing a symbolic adjoint is viable. Moreover, the time needed to computer the derivatives speed up by applying the combined approach which is clearly superior for higher number of iteration. The overall design of a generic transformation methodology is presented that is applicable for complicated computer programs.

ACKNOWLEDGEMENT

The presented research is supported by the project **aboutFlow**, funded by the European Commission under grant no.FP7-PEOPLE-2012-ITN-317006.

References

[1] U. Naumann. *The Art of Differentiating Computer Programs. An Introduction to Algorithmic Differentiation*. SIAM, Philadelphia, 2012.

[2] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, Jan 2000.

[3] Z. Dastouri, J. Lotz, and U. Naumann. Development of a discrete adjoint CFD code using Algorithmic Differentiation by Operator Overloading. In *OPT-i: An International Conference on Engineering and Applied Sciences Optimization*, Athens, 2014. National Technical University of Athens.

[4] Z. Dastouri and U. Naumann. Improving efficiency of a discrete adjoint CFD code for design optimization problems. In *EUROGEN: Proceeding of International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems*, Glasgow, UK, 2015.

[5] S.V. Patankar and D.B. Spalding. A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. *International Journal for Heat Mass Transfer*, 15(10):1787–1806, 1972.

[6] D. Jones, F. Christakopoulos, and J-D. Mller. Preparation and assembly of adjoint CFD codes. *Computers and Fluids*, 46(1):282286, July 2011.

[7] C. Othmer and T. Grahs. Approaches to fluid dynamic optimization in the car development process. In R. Schilling et. al., editor, *Evolutionary Methods for Design, Optimisation and Control with Applications to Industrial Problems*, Munich, 2005. FLM.

[8] B. Christianson. Reverse Accumulation and attractive fixed points. *Optimization Methods and Software*, 3:311–326, 1994.

[9] L. Hascoët and V. Pascual. The Tapenade Automatic Differentiation tool: Principles, Model, and Specification. *ACM Transactions On Mathematical Software*, 39(3), 2013.

[10] C. W. Straka. Adf95: Tool for automatic differentiation of a fortran code designed for large numbers of independent variables. *Computer Physics Communications*, 168(2):123–139, 2005.

[11] D. Shiriaev, A. Griewank, and J. Utke. A user guide to ADOL–F: Automatic Differentiation of Fortran codes. *Tech. Report*, (IOKOMO–04–1995), 1996.

[12] S. Stamatiadis, R. Prosmiti, and S. C. Farantos. AUTO_DERIV: Tool for Automatic Differentiation of a FORTRAN code. *Computer Physics Communications*, 127(2&3):343–355, 2000.

[13] C. Qiang, Z. Linbo, and W. Bin. Model adjointization and its cost. *Science in China, Series F–Information Sciences*, 47(5):587–611, 2004.

[14] C. Bischof, A. Carle, G. Corliss, A Griewank, and P. Hovland. ADIFOR: Generating derivative codes from Fortran programs. *Scientific Programming*, 1(1):11–29, 1992.

[15] R. Giering and T Kaminski. Recipes for adjoint code construction. *ACM Transactions on Mathematical Software*, 24(4):437–474, 1998.

[16] U. Naumann, U. Utke, C. Wunsch, C. Hill, P. Heimbach, M. Fagan, N. Tallent, and M. Strout. Adjoint code by source transformation with OpenAD/F. 2006. Available online at `http://proceedings.fyper.com/eccomascfd2006/documents/35.pdf`.

[17] S. Graham, P. Kessler, and M. Mckusick. Gprof: A call graph execution profile. In *SIGPLAN '82 Proceedings of the 1982 SIGPLAN symposium on Compiler construction,*, number ISBN:0-89791-074-5, pages 120–126, ACM New York, NY, USA, 1982.

[18] J. Reinders. Vtune performance analyzer essentials measurement and tuning techniques for software developers. In *Intel Press,*, Hillsboro, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EFFICIENT CALIBRATION OF DISCRETE ELEMENT MATERIAL MODEL PARAMETERS USING LATIN HYPERCUBE SAMPLING AND KRIGING

M. Rackl[1], C.D. Görnig[1], K.J. Hanley[2] and W.A. Günthner[1]

[1]Institute for Materials Handling, Material Flow, Logistics
Technical University of Munich
Boltzmannstraße 15, 85748 Garching, Germany
e-mail: rackl@fml.mw.tum.de, guenthner@fml.mw.tum.de

[2] Institute for Infrastructure and Environment, School of Engineering
The University of Edinburgh
Edinburgh EH9 3JL, United Kingdom
e-mail: K.Hanley@ed.ac.uk

Keywords: discrete element method, calibration, meta model, Kriging

Abstract. *Material model parameter identification for discrete element models (DEM) is typically done using a trial-and-error approach and its outcome depends largely on the experience of the DEM user. This paper describes a work flow which facilitates the efficient and systematic calibration of discrete element material models against experimental data. The described workflow comprises three steps. In the first step, an approach based on the design and analysis of computer experiments (DACE) is adopted in which data is generated for the parametrisation of Kriging models based on Latin hypercube sampling. In the second step, multi-objective optimisation is applied to the Kriging models. This study introduces an additional cost criterion, which includes the Rayleigh time step, in order to reduce the solution set size to one element. In the third step, the optimisation procedure is repeated with the actual DEM models, using the optimal parameter set from the Kriging models as the start value. This final step with the full DEM models refines the parameter set against experimental data. Since DEM material model calibration is time-consuming, the workflow is implemented into an automated process chain. In this paper, the methodology is described in detail and results are shown which illustrate the usefulness and effectiveness of this approach. Initial verification simulations run using the calibrated parameters give good agreement with experimental results.*

1 Introduction

Many advanced modelling methods and tools are available, and the continual increases in affordable computational power allow more ambitious and realistic simulations to be run year-on-year. One remaining problem affecting certain simulation methods, which particularly inhibits their adoption in industry, is the difficulty in choosing suitable input parameters. This is especially true when the simulation input parameter does not have a physical analogue.

This paper focuses on one simulation tool, the discrete element method (DEM). A work flow is described and demonstrated for calibration of key simulation input parameters against experimental data. The objective is to allow DEM users to identify their required input parameters in an efficient, user-friendly and systematic manner.

1.1 Overview of discrete element method

Since its formulation in the 1970s [1], the discrete element method (DEM) has become increasingly popular for simulating complex systems of particulates at the particle scale by specifying a relatively small number of microstructural parameters. The growing interest in DEM among the scientific community is demonstrated by the publication of a number of special issues of journals devoted to DEM within the past five years (e.g., Engineering Computations 26(6); Granular Matter 11(5); Powder Technology 193(3); Powder Technology 248). Literature surveys have also shown a rapid increase in the popularity of DEM since its introduction [2, 3].

In 'soft-sphere' DEM simulations, particles are modelled as bodies of finite size, inertia and stiffness. The geometries of the simulated particles are usually idealised to reduce the computational requirements. The particles are rigid but are allowed to overlap; this interparticle overlap is analogous to the deformation that occurs at real particle contacts. The most common implementation of DEM is based on an explicit, conditionally-stable time-stepping algorithm. In each time step [3], interparticle forces are evaluated at contact points and resultant forces are calculated for each particle. Newton's Second Law is then applied to determine the translational and angular particle accelerations. These acceleration terms are integrated numerically to find particle velocities and displacements. Finally, the displacements are used to update the particle positions during each time step.

1.2 DEM material parameter identification

Even though a typical DEM simulation requires relatively few input parameters, finding these can be challenging. Consider the simplest case in which each idealised, simulated particle represents one physical particle. The rheological parameters for input to the contact models (i.e., force–displacement laws) can be difficult to determine accurately by experiment [4]. The interparticle friction coefficient strongly influences the system response in simulations using spheres [5], but its value is often increased beyond physical measurements of friction in an effort to model the particle irregularity in the real system. Numerical damping [6, 7] is difficult to relate to physical measurements. These problems are exacerbated when irregular particles are simulated: even simple rolling resistance models require at least one non-physical parameter [8]. Other complications of the basic DEM, e.g., crushable fundamental particles, also require parameters which lack a physical basis [9]. 'Coarse graining' may be used to reduce the computational requirements of a simulation. The real particles are replaced by fewer, larger coarse particles, each of which represents many real particles [10, 11, 12]. Thus, none of the input parameters for a coarse-grained simulation may necessarily be obtainable by laboratory testing.

In general, it is not possible to infer a complete set of appropriate parameters for a DEM

simulation directly from measured properties of the physical material. Therefore, some (or all) of the parameters may need to be determined by calibration, i.e., by varying the unknown DEM parameters until a good match is obtained between physical measurements of the response(s) of interest and the simulation results. Although DEM calibration is often used, e.g., [13, 14, 15], the main limitation of existing, commonly-used calibration methods is inefficiency. The parameters are often varied individually by trial-and-error while the user-defined response of interest is monitored. This approach has many disadvantages [4]: the effectiveness of the approach is dependant on the experience of the user, the final parameters obtained may not be optimal, the number of DEM simulations required for calibration is not known in advance, and limited mechanistic insight is gained. As DEM simulations can be very computationally expensive, it is desirable to maximise the efficiency of the calibration process. Several researchers have proposed more efficient DEM calibration approaches using design of experiments (DoE) methods [4, 16, 17, 18] but these have found limited application to date. In subsection 2.4 of this paper, an alternative work flow for DEM calibration is described which is both efficient and user-friendly.

2 MATERIALS AND METHODS

The work flow proposed in this paper is based on Kriging meta modelling, multi-variate optimisation and general concepts of Design and Analysis of Computer Experiments. A detailed description of the process implemented using GNU Octave (Octave, [19]) is given in subsection 2.4.

2.1 Design and Analysis of Computer Experiments

Design and Analysis of Computer Experiments (DACE) is closely related to the Design of Experiments (DoE) for physical experiments. It includes methods and work flows for the characterisation of computationally inexpensive and expensive black box models. In the case of DEM material parameter calibration, the black box models are the material test setups, which were modelled using DEM. The reader is referred to widely-available literature on DoE and DACE for nomenclature, general ideas as well as background of the methods (e. g. [20, 21]).

As with DoE, any DACE workflow can be divided into four steps:

1. Define the factors, factor levels (or factor intervals) and the response to be recorded.
2. Create an experimental plan based on the data of step 1 and the requirements for step 4.
3. Conduct the experiments by running simulations according to the experimental plan and gather the response data.
4. Transform (where required) the data, use them to parametrise meta models and harness the latter for e. g. optimisation of the response.

2.2 Kriging and Latin Hypercube sampling

Kriging, which is also known as Gaussian process regression, is a statistical regression method for meta modelling which is based on covariances. It requires response data from several samples taken from within a previously-defined N-dimensional factor space and estimates the (meta) model parameters via a maximum likelihood approach [22]. Basic ideas for Kriging were developed and applied in [23] and later generalised [24]. In this study, the Kriging

implementation of the *Small Toolbox for Kriging* (STK, [25]) was used under Octave. The baseline options for the Kriging models were set to using an anisotropic Matérn covariance matrix together with a linear trend.

The experimental plan, which prescribes the factor value combinations at which the response data are to be computed, is generated via Latin hypercube sampling (LHS, [26]). LHS generates factor value combinations by randomly, but evenly, distributing points within the specified factor interval limits. To further reduce the variance of the Kriging models, this study uses LHS in conjunction with a MAXIMIN criterion [21, p. 138 ff.].

2.3 Issues addressed by the proposed work flow

Since the calibration of DEM material parameters is greatly dependent on the DEM user, the aim of the proposed work flow was to reduce the user interaction to a minimum. Three major issues with DEM material calibration were selected to be addressed by the work flow.

First of all, the calibration process should become automatic, which means that computations required to calibrate the material parameters are to be set up and run automatically. This also includes the evaluation of DEM results. While the manual evaluation of e. g. the bulk density is fairly easy, the measurement of, for example, the angle of repose is somewhat subjective and its result may vary with the measurement method used. It is therefore important to ensure comparability of different values of the same response, which is solved using an automatised results evaluation.

Secondly, keeping track of how changes of the material parameters affect the DEM results is a challenging task, even for an experienced user. Due to the sheer number of up to 14 parameters[1] and possible interactions, it becomes very difficult to relate parameters to effects when altering several parameters at once. This problem could be resolved by only changing one factor at a time; however this approach does not account for parameter interactions and is highly inefficient [20]. The solution suggested in this study is to combine LHS and Kriging methods to create meta models that account for interactions and corresponding effects.

Thirdly, the computational efficiency of the calibrated material model may not always be the focus of the calibration process. This is especially true since the Rayleigh time step in DEM depends on material parameter values (i. e. altering the Young's modulus affects the maximum time step size). Hence, a further target of the described work flow is to take the Rayleigh time step into account and not only seek precise prediction of measurement results but also larger minimum time step sizes. Within subsection 2.4, a cost function that accounts for the Rayleigh time step is proposed and implemented.

2.4 Work flow of the proposed calibration process

The described workflow was developed and implemented using the open-source DEM software LIGGGHTS [29]. Since DEM material model calibration is time-consuming, the workflow is implemented into an automated process chain using GNU Octave. The user has to integrate custom DEM models of material tests into the folder structure of the Octave scripts. A custom DEM model usually consists of a LIGGGHTS model of a bulk material test along with auxiliary functions which allow for automatic evaluation of the result files. For example, this could be the model of a simple bulk density experiment, including Octave functions to compute the bulk density from the total particle mass and the known volume of the container[2].

[1] DE model with one particulate medium and wall interaction.

[2] Setting up custom models is a non-recurring task.

The only actual user input are the measurement results from physical experiments (e. g. bulk density) and intervals for the material and contact parameters of materials in Octave. These intervals will have to include the final solution set. Of course, the final solution set is unknown at this point, so the intervals can be of any finite width. Subsequently, the algorithm can be started and it will finish autonomously.

The main workflow comprises three steps. In the first step, an approach based on the design and analysis of computer experiments is adopted in which data is generated for the parametrisation of Kriging models based on Latin hypercube sampling. The material and contact parameter values, as defined by the user, are used as free variables for the calibration. The programme will carry out a user-specified number of DEM simulations, which provide the input for the Kriging meta model parametrisation (cf. Figure 1). Depending on the performance of the computer, several simulations may be run in parallel to save time. All obtained results are handed over and saved in Octave.

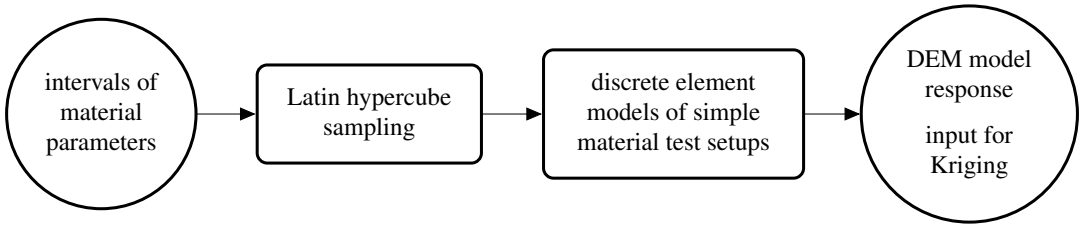


Figure 1: Generation of response data as input for the Kriging models.

In the second step, multi-objective optimisation is applied to the Kriging models (cf. Figure 2). For the optimisation the frontend 'nonlin-residmin' Octave function for non-linear residual minimization is used which is based on the Levenberg-Marquardt algorithm [27, 28]. All results of the previous DEM simulations and the Latin hypercube sampling plan are used as responses to parametrise Kriging models. The Kriging models are then used to optimise the set of variable material and contact parameters such that the user-specified measurement values are attained.

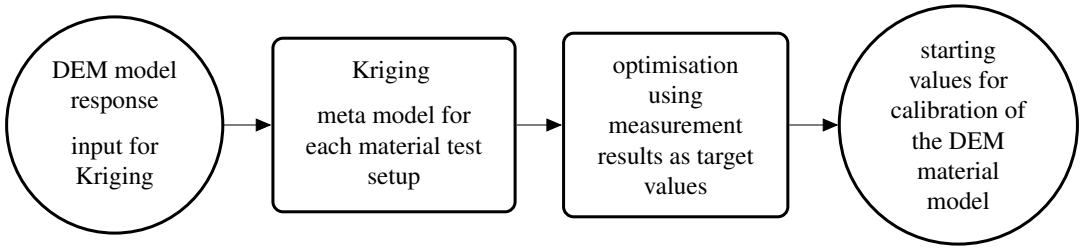


Figure 2: Response data from the LHS are used as input to the Kriging models.

For this purpose, the multi-objective cost function

$$c_i = \frac{s_i - m_i}{m_i} \tag{1}$$

has been implemented where s_i is the result of the simulation and m_i is the corresponding physical measurement. The smaller the difference between the two, the smaller c_i is. The function calculates i proportions of i measurement results and all have the same weighting. To avoid generating computationally-expensive sets of DEM parameters, the Rayleigh time step

will also be optimised. The cost function of the Rayleigh time step is

$$t_i = \frac{r_{max} - dt_r}{r_{max} - r_{min}}, \tag{2}$$

where dt_r is the Rayleigh time step, r_{max} is the best-case of the Rayleigh time step and r_{min} is the worst-case. Best and worst-case refer the minimum and maximum achievable Rayleigh time with respect to the intervals specified. The Rayleigh time step is calculated according to

$$dt_r = \pi \cdot r \cdot \frac{\sqrt{\frac{\rho}{G}}}{0.1631 \cdot \nu + 0.8766}, \tag{3}$$

where ρ is particle density, G is the shear modulus, ν is Poisson's ratio and r is the radius of the smallest particle. The Rayleigh time step depends on the parameter values that have to be optimised. The cost function of the Rayleigh time step is linear and it can take values between 0 and 1.

In the third step, the optimisation procedure is repeated with the actual DEM models, using the optimal parameter set from the Kriging models as start values (cf. Figure 3). This final step only requires a few more runs with the full DEM models and simultaneously refines the parameter set against the experimental measurement data.

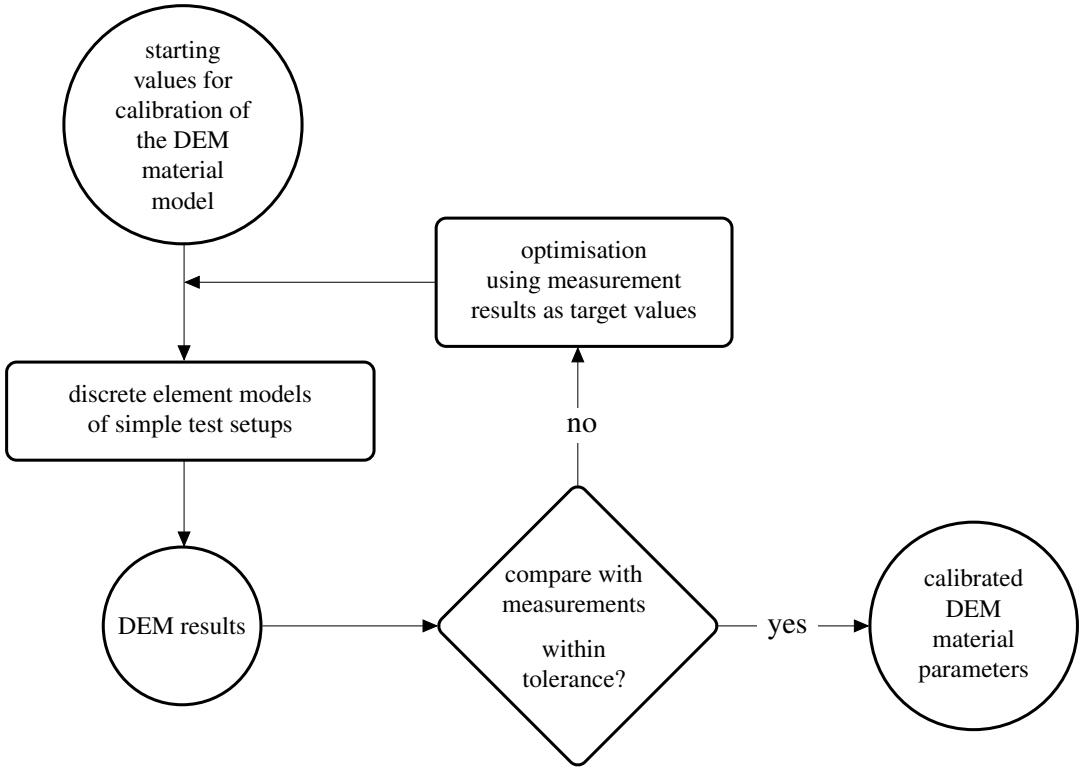


Figure 3: Optimisation procedure for DEM models

2.5 A small study using glass beads

A small initial study was set up in order to test the proposed work flow within a realistic framework and glass beads were chosen as the bulk solid of interest. The overall aim was to

calibrate the density and rolling friction (particle-particle) in such way that the angle of repose and bulk density results match measured values. A bulk density of 1500 kg m^{-3} and an angle of repose of 22° were found in literature for glass beads ([30, p. 16], [31]). The particle diameter was set to 10 mm.

A combined discrete element model was used for simulation of the bulk density and angle of repose. Since measuring the angle of repose can be difficult, an automatic evaluation algorithm, based on image recognition, was developed for this study. The DEM software LIGGGHTS was applied with a Hertz-Mindlin contact model and the elastic-plastic spring-dashpot (EPSD) rolling friction model. Table 1 shows the DEM input parameters used.

material parameter	**particle**	**wall**	**p-p**	**p-w**	**constant?**
Young's modulus (N m^{-2})	3.6×10^5	2.11×10^{11}			yes
Poisson's ratio (-)	0.22	0.30			yes
density (kg m^{-3})	2100	7700			used for calibration
contact parameter					
friction coefficient (-)			0.18	0.27	yes
rolling friction (-)			0.67	0.005	used for calibration
coefficient of restitution (-)			0.79	0.82	yes
viscous RF damping (-)			0.25	0.25	yes

Table 1: Constant DEM input parameters and those used for calibration of the bulk density and angle of repose (p-p: particle-particle; p-w: particle-wall; RF: rolling friction).

In addition to running the described test case, the influence of the factor interval width (FIW) and the random factory seed (RFS) of the DEM models was investigated for the purpose of testing the robustness of the work flow. The FIW prescribes how wide the factor interval of the calibration parameters is, whereas the RFS controls the random generation of particles when starting a DEM simulation in LIGGGHTS. The influence of RFS on the results should only be of a stochastic nature, in theory. Table 2 shows details about the RFS and FIW levels and the experimental plan for this study. For the two non-constant parameters–density and rolling friction– maximum and minimum values (intervals) were specified before each calibration run (±FIW). During each calibration run 26 probes were computed to parametrise the Kriging models.

	FIW (%)	**RFS** (-)
levels	10, 33, 66	3000, 4000
run no.		
1	10	3000
2	10	4000
3	33	3000
4	33	4000
5	66	3000
6	66	4000

Table 2: Factor levels and experimental plan for the study to investigate factor interval width (FIW) and random factory seed (RFS).

3 RESULTS AND DISCUSSION

Two factors, the FIW and the RFS, were altered to initially assess the reliability of the work flow for DEM material parameter calibration. Each of the runs from the experimental plan

converged to a meaningful solution set. It took the multi-variate optimisation algorithm a mean of 10 optimisation runs (min. 9, max. 15) with the actual DEM models.

The desired values for bulk density and angle of repose were closely met, as deviations were within a few percent. Mean differences were 1.3 % (min. 0.45 %, max. 2.7 %) for the angle of repose and 0.92 % (min. 0.13 %, max. 3.1 %) for bulk density. One should also note that the variance measured in physical experiments, especially for the angle of repose, is typically significantly higher than the deviations recorded in this study.

Due to the small number of simulation experiments, it is not fully clear how much the RFS value influences the results. Ideally, it should only induce stochastic noise. According to the results depicted in Figure 4 and Figure 5, it does not generate a systematic trend or offset.

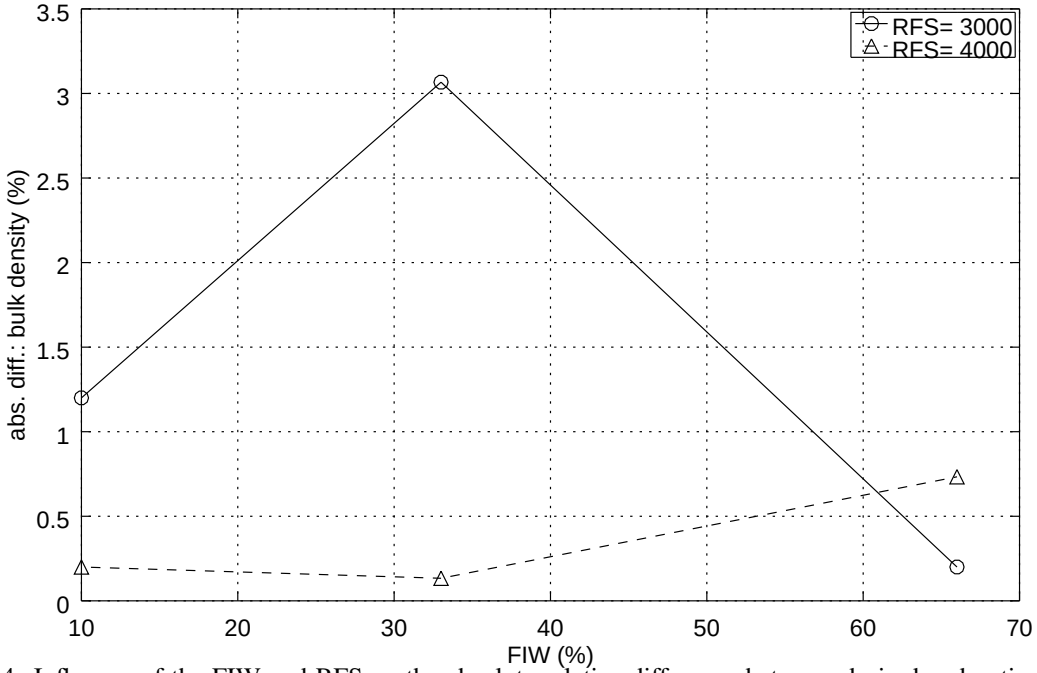


Figure 4: Influence of the FIW and RFS on the absolute relative difference between desired and optimised bulk density.

There is a clear correlation between the number of optimisation runs and the FIW. It can be seen from Figure 6 that a wider FIW results in more optimisation runs being required, independently of the RFS. This behaviour was expected, since the sampling density reduces with an increased FIW and the Kriging models have to interpolate in areas with a greater distance between interpolation points. The Kriging model predictions therefore show a greater variance on those areas and the precision of the estimates for starting values for the DEM optimisation reduces.

4 CONCLUSION

A framework for efficient calibration of discrete element material model parameters was implemented in this study. A total of approximately 38 DEM runs were required to calibrate particle density and rolling friction against measurements of the angle of repose and bulk density. The results from a small study with glass beads show that the general approach works and desired results were well within practical tolerances of the desired values. Nonetheless, more studies will be necessary to further validate the usefulness of the described work flow as well as its robustness, e. g. when unfavourable parameter intervals are set by the user.

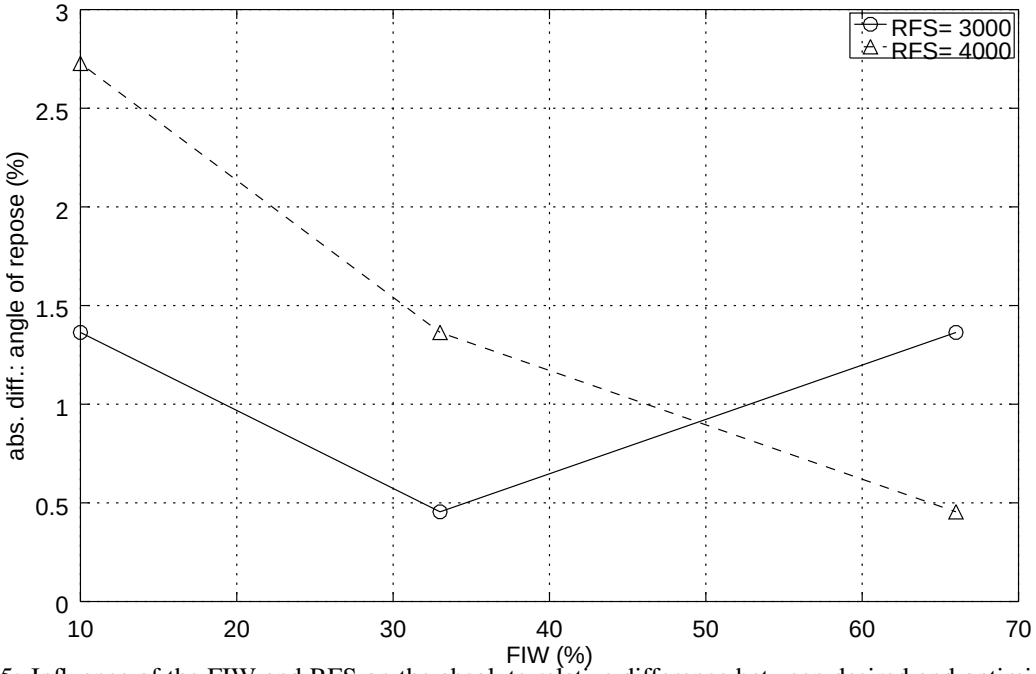


Figure 5: Influence of the FIW and RFS on the absolute relative difference between desired and optimised angle of repose.

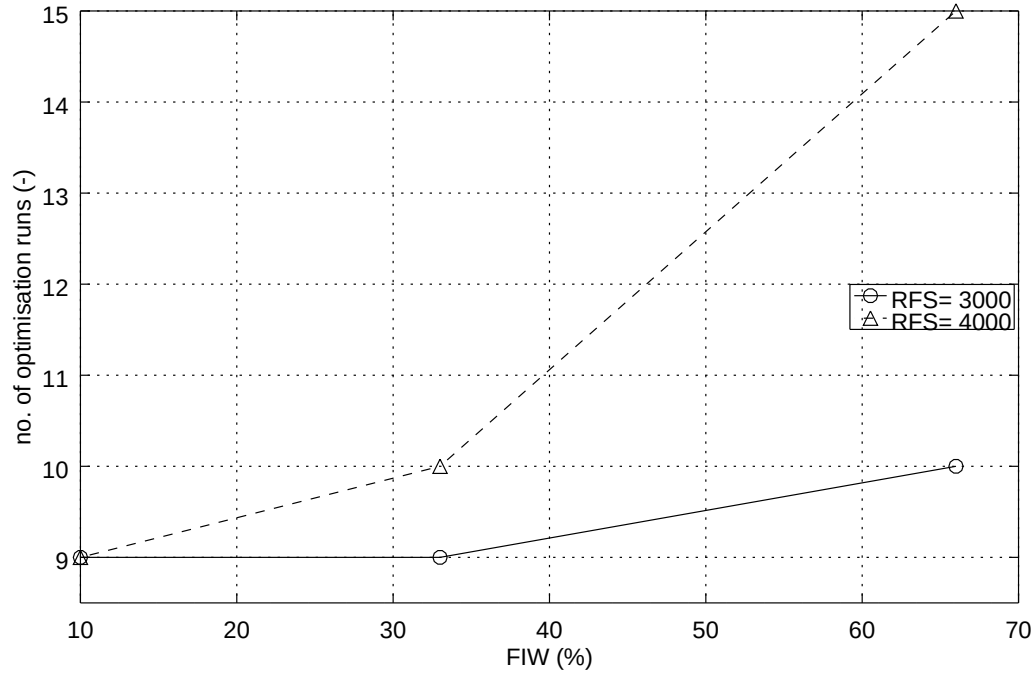


Figure 6: Influence of the FIW and RFS on the number of optimisation runs required.

5 ACKNOWLEDGEMENT

This study received funding by the AiF (no. 18371 N/1), within the programme for sponsorship by Industrial Joint Research (IGF) of the German Federal Ministry of Economic Affairs and Energy based on an enactment of the German Parliament.

REFERENCES

[1] Cundall, P.A. & Strack, O.D.L. (1979). A discrete numerical model for granular assemblies. *Géotechnique*, **29**(1): 47–65.

[2] Zhu, H.P., Zhou, Z.Y., Yang, R.Y. & Yu, A.B. (2007). Discrete particle simulation of particulate systems: Theoretical developments. *Chemical Engineering Science*, **62**(13): 3378–3396.

[3] O'Sullivan, C. (2011). *Particulate discrete element modelling: A geomechanics perspective* (1st ed.), Taylor & Francis, Oxford, UK.

[4] Hanley, K.J., O'Sullivan, C., Oliveira, J.C., Cronin, K. & Byrne, E.P. (2011). Application of Taguchi methods to DEM calibration of bonded agglomerates. *Powder Technology*, **210**(3): 230–240.

[5] Huang, X., Hanley, K.J., O'Sullivan, C. & Kwok, C.-Y. (2014). Exploring the influence of interparticle friction on critical state behaviour using DEM. *International Journal for Numerical and Analytical Methods in Geomechanics*, **38**(12): 1276–1297.

[6] Potyondy, D.O. & Cundall, P.A. (2004). A bonded-particle model for rock. *International Journal of Rock Mechanics and Mining Sciences*, **41**(8): 1329–1364.

[7] Itasca Consulting Group (2008). *PFC3D: Particle flow code in three dimensions User's Guide* (4th ed.), Minneapolis, Minnesota.

[8] Ai, J., Chen, J.F., Rotter, J.M. & Ooi, J.Y. (2011). Assessment of rolling resistance models in discrete element simulations. *Powder Technology*, **206**, 269–282.

[9] Hanley, K.J., O'Sullivan, C. & Huang, X. (2015). Particle-scale mechanics of sand crushing in compression and shearing using DEM. *Soils and Foundations*, **55**(5): 1100–1112.

[10] Sakai, M. & Koshizuka, S. (2009). Large-scale discrete element modeling in pneumatic conveying. *Chemical Engineering Science*, **64**(3): 533-539.

[11] Sakai, M., Abe, M., Shigeto, Y., Mizutani, S., Takahashi, H., Viré, A., Percival, J.R., Xiang, J. & Pain, C.C. (2014). Verification and validation of a coarse grain model of the DEM in a bubbling fluidized bed. *Chemical Engineering Journal*, **244**: 33–43.

[12] Nasato, D.S., Goniva, C., Pirker, S. & Kloss, C. (2015). Coarse Graining for Large-scale DEM Simulations of Particle Flow – An Investigation on Contact and Cohesion Models. *Procedia Engineering*, **102**: 1484–1490.

[13] Doležalová, M., Czene, P. & Havel, F. (2002). Micromechanical modeling of stress path effects using PFC2D code. In: H. Konietzky (Ed.), *Numerical Modeling in Micromechanics via Particle Methods*, Balkema, Lisse, pp. 173–182.

[14] Asaf, Z., Rubinstein, D. & Shmulevich, I. (2006). Evaluation of link-track performances using DEM. *Journal of Terramechanics*, **43**(2): 141–161.

[15] Coetzee, C.J. & Els, D.N.J. (2009). Calibration of discrete element parameters and the modelling of silo discharge and bucket filling. *Computers and Electronics in Agriculture*, **65**(2): 198–212.

[16] Yoon, J. (2007). Application of experimental design and optimization to PFC model calibration in uniaxial compression simulation. *International Journal of Rock Mechanics and Mining Sciences*, **44**(6): 871–889.

[17] Favier, J., Curry, D. & LaRoche, R. (2010). Calibration of DEM material models to approximate bulk particle characteristics. *Proc. 6th World Congress on Particle Technology*, Nuremberg, Germany.

[18] Johnstone, M. (2010). *Calibration of DEM models for granular materials using bulk physical tests*. PhD thesis, University of Edinburgh.

[19] Eaton, J.W., Bateman, D., Hauberg, S., Wehbring, R. (2015). *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*.

[20] Dean, A., Morris, M., Stufken, J., and Bingham, D. (2015). *Handbook of Design and Analysis of Experiments*. CRC Press, Hoboken.

[21] Santner, T.J., Williams, B.J., Notz, W.I (2003). *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer, New York, NY.

[22] Jones, D.R., Schonlau, M., Welch, W.J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

[23] Krige, D.G. (1951). *A statistical approach to some mine valuations and allied problems at the Witwatersrand*. Master's thesis, University of Witwatersrand.

[24] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8):1246–1266.

[25] Bect, J., Vazquez, E. et al (2014). *STK: a Small (Matlab/Octave) Toolbox for Kriging. Release 2.3*.

[26] McKay, M.D., Beckman, R.J., Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239.

[27] D.W. Marquardt. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431-441.

[28] K. Levenberg. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *The Quarterly of Applied Mathematics*, 2: 164-168.

[29] Kloss, C., Goniva, C., Hager, A., Amberger, S., Pirker, S. (2012). Models, algorithms and validation for opensource DEM and CFD-DEM. *Progress in Computational Fluid Dynamics*, 12(2/3):140–152.

[30] Brown, R.L., Richards, J. C., Danckwerts, P. V. (1970). Principles of Powder Mechanics: Essays on the Packing and Flow of Powders and Bulk Solids, volume volume 10 of *International series of monographs in chemical engineering*. Elsevier Science, Burlington.

[31] Wong, A.C.-Y. (2000). Characterisation of the flowability of glass beads by bulk densities ratio. *Chemical Engineering Science*, 55(18):3855–3859.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

UPDATING PRIOR PARAMETERS BASED ON LIKELIHOOD FUNCTION-BAYESIAN METHOD FOR PARAMETER ESTIMATION AT HIGH MEASUREMENT UNCERTAINTY

S. Somasundharam[1], **and K. S. Reddy**[2]

[1]Research Scholar
Heat transfer and Thermal power Laboratory, IIT Madras, Chennai-600 036
e-mail: somukpk@gmail.com

[2] Professor
Heat transfer and Thermal power Laboratory, IIT Madras, Chennai-600 036
e-mail: ksreddy@iitm.ac.in

Keywords: Bayesian Inference, Parameter Estimation, Update Prior Parameters, Markov Chain Monte-Carlo (MCMC), Inverse Heat Conduction.

Abstract. *One approach in statistical analysis that distinguishes between frequentist and Bayesian is the inclusion of available prior information about the data even before measuring/ surveying the data. Many researchers argued that the inclusion of prior information resulted in a better model prediction/parameter estimation. Bayesian inference is repeatedly used in inverse problems to retrieve parameters due to the development of high efficient sampling algorithms such as Markov Chain Monte-Carlo (MCMC). Inverse problems are generally ill-posed in nature. Nevertheless, the inclusion of prior information reduces the ill-posedness of the problem to an extent. Any inverse problem relies on measured data by physical sensors, therefore induced random errors greatly affects the quality of estimation. When the uncertainty of the measured data is high, the inferences made from the resulting sampling distributions are nearly the same as the supplied prior information. The reason is that, probability of samples nearer to prior information is more at higher uncertainties of measurement and has more chance to get repeatedly accepted than the samples that are close to actual value. Therefore, in this work an effort is made to update the prior hyper parameters in each iterations of the MCMC algorithm based on the history of the likelihood function. The applicability of method is demonstrated by retrieving parameters from 1-D fin experiment. Three thermal properties such as thermal conductivity, heat transfer coefficient and emissivity are retrieved simultaneously. The estimation is carried out for both error and errorless temperature measurements and the results show that, the estimated parameters with the proposed method are in excellent agreement with the true parameter value and a maximum of 7% deviation occurs in estimating heat transfer coefficient at a measurement error of* $\pm 0.3K$

1 INTRODUCTION

In recent years, Bayesian inference is predominantly used in many engineering, science, social and behavioral sciences due to the advancement in high computing resources and efficient sampling algorithms. The result of Bayesian is a probabilistic statement regarding the unknown quantities unlike a single point estimate in other methods e.g., least-square parameter estimation. Bayesian treats the measured/surveyed data as a random variable and hence the inferences made by Bayesian are random variables. The uniqueness about Bayesian inference lies in that, a prior knowledge about the unknown parameters could be included in the statistical analysis of the data. The inclusion of prior information in inverse problems reduces the ill-posed nature of the problem and improves the quality of estimation. In general, prior information in the form of Gaussian distribution is commonly employed in inverse heat transfer literatures. Parathasarathy and Balaji [1] studied the different forms of prior model and its effects in the estimation of single and multiple parameters. In their work, they demonstrated that in multi-parameter estimation, Bayesian inference tends to point to alternate feasible solutions when highly correlated parameters are retrieved using non-informative prior models.

Monte-Carlo sampling techniques are widely practiced in order to find the expectations of the probability distribution functions. Among several Monte-Carlo techniques, Markov Chain Monte-Carlo (MCMC) is more frequently used in inverse heat transfer problems [2]-[4]. The hyper parameters (mean and standard deviation) of the prior model are usually taken from previously published results/previous experiments. More frequently these prior parameters are known vaguely. Under such case, when correlated parameters are estimated or when the measurement uncertainty is high (more likely to occur due to unavoidable random errors) the estimated parameters using sampling algorithms are close to the subjected prior parameters.

Single parameter estimation (inverse estimation of thermal conductivity) is carried out at different levels of measurement uncertainty. Refer section 2 and 3 for problem specification and inverse formulation. MCMC sampler is used to sample through the posterior. The estimation results reveal that, as the measurement uncertainty increases, the estimated parameter (mean of the posterior distribution) becomes closer and closer to the subjected prior mean. The reason for this could be explained intuitively as follows: consider the case (a) wherein the generated sample is close to actual parameter. In this case, at low uncertainty the likelihood probability is relatively larger than prior probability, so the combined probability (posterior probability) is more likely greater than the acceptance ratio (refer MCMC algorithm) and hence the sample is more frequently accepted. However, at high uncertainties the likelihood probability is slightly lesser or equal to prior probability, so the combined probability is likely lesser than the acceptance ratio and hence the sample is frequently rejected even though it is close to actual parameter. Now, consider the case (b) wherein the generated sample is close to prior parameter (mean). In this case, at low uncertainty the likelihood probability is relatively less compared to prior probability, so the combined probability is more likely lesser than acceptance ratio and hence the sample is more frequently rejected. However, at higher uncertainties the likelihood probability is slightly lesser or equal to prior probability, so the combined probability is likely greater than the acceptance ratio and hence the sample is accepted frequently even though it is away from actual parameter (nearer to prior mean). Therefore, in this work an attempt is made to modify the prior parameters in each iteration of the MCMC algorithm based on the history of likelihood function to overcome the aforesaid difficulty and study the applicability of modified algorithm to multi-parameter estimation related to inverse heat transfer applications.

2 DEFINITION OF FORWARD PROBLEM

A simple 1-D fin losing heat to the surrounding both by convection and radiation is considered as the forward problem. Mathematically, the forward problem is represented as,

$$k\frac{\partial^2 T}{\partial x^2} - hA_s(T - T_\infty) - \epsilon\sigma A_s(T^4 - T_\infty^4) = C_p\frac{\partial T}{\partial t} \qquad in\ 0 \leq x \leq L \tag{1}$$

subjected to the following initial and boundary conditions:

$$T(x, 0) = T_i = T_\infty \qquad 0 \leq x \leq L$$

at $x = 0$,

$$q = \begin{cases} q_0 & 0 \leq t \leq t_0 \\ 0 & t\ > t_0 \end{cases}$$

In equation 1 k, ϵ and C_p are fin material properties. i.e., thermal conductivity (W/mK), surface emissivity and heat capacity (J/K) respectively. A_s (m^2) is lateral surface of the fin and L (m) is the length of the fin. A step heat input of magnitude q_0 is given at one end ($x = 0$) for time t_0. The experiment is carried out in an ambient of temperature T_∞(K) and heat transfer coefficient, h (W/m^2K)

The solution to equation 1 is the temperature distribution along the length of the fin w.r.t. time. In this work, Finite Volume Method (FVM) is employed to solve the forward problem [5] in which energy balance is applied to each discretized control volume in order to convert the above partial differential equation into set of algebraic equations. Because of the radiation heat loss term, equation 1 is non-linear and hence the discretized equations are also non-linear. Therefore, they are solved simultaneously using Gauss-Seidel iterative technique.

3 BAYESIAN INVERSE FORMULATION FOR PARAMETER ESTIMATION

The objective of the Bayesian inference in inverse problems is to formulate the *posterior probability distribution (ppdf)*, which in Bayesian context of linear/non-linear parameter estimation is *proportional* to product of *likelihood function and prior probability distribution*.

$$p(X|Y) \propto p(Y|X)p(X) \tag{2}$$

In equation 2 X is the unknown parameter vector to be estimated and Y is the measured/ surveyed data. The LHS of equation 2 is the required *ppdf*, whereas the first and second term of RHS is the *likelihood function* and *prior distribution* respectively. Without loss generosity, it is assumed that the errors between the measured and simulated variable is *additive, uncorrelated, normally distributed with zero mean* and *constant standard deviation* and hence the *likelihood function* is given by [6, 7].

$$p(Y|X) = \frac{1}{\sqrt{2\pi}}|V|^{-1/2} exp\left\{\frac{-1}{2}[Y - Y(X)]^T V^{-1}[Y - Y(X)]\right\} \tag{3}$$

In equation 3 Y is the measured quantity, $Y(X)$ is the simulated quantity for the assumed unknown parameter X and V is the covariance matrix of the measurement. As previously mentioned, Bayesian exploits all the information that are available even before the measurements

are taken, this information is usually included in the analysis in the form of *prior distribution p(X)* as given in equation 2. In this work, a normal prior model is assumed and hence it is given by,

$$p(X) = \frac{1}{\sqrt{2\pi}}|V_p|^{-1/2} exp\left\{\frac{-1}{2}[X-\mu_p]^T V_p^{-1}[X-\mu_p]\right\} \tag{4}$$

In equation 4 μ_p and V_p are hyper parameters of the prior distribution (mean and covariance) respectively. Substituting 3 and 4 in equation 2 we get the desired *ppdf*. For the present parameter estimation problem the unknown parameters are thermal conductivity, heat transfer coefficient and emissivity i.e., $X = \{k, h, \epsilon\}$ and the measurable quantity is the temperature as a function of space and time $T(x, t)$.

3.1 MCMC algorithm for updating prior parameter

The next step in Bayesian inference is to calculate the appropriate statistic (mean/mode/ median/maximum a posterior/and standard deviation) of the unknown parameters using the posterior distribution. However, in inverse heat transfer literatures *mean/maximum a posterior* and *standard deviation* of the unknown parameters are more commonly reported. Monte-Carlo sampling methods are being widely used for this purpose, among which *Markov Chain Monte-Carlo (MCMC)* is more repeatedly used in the literatures. An excellent discussion is presented in [8] as in why to use sampling techniques over the analytical/numerical methods in order to find the *expectations of ppdf*. In the present work, the simple MCMC algorithm is slightly modified to update prior hyper parameters based on the history of likelihood function to overcome the convergence of estimated parameters to the prior mean at high measurement uncertainties. One iteration of the algorithm consist of the following steps,

1. Select a new sample, X^* from the proposal/jumping distribution $q(X^*|X^{i-1})$
2. Find the *ppdf* for new sample, p^* using 2
3. Accept the new sample based on *Metropolis Hasting (MH)* ratio, if $r \leq U$ otherwise reject the sample. Here U is a uniformly distribute random number and

$$r = min\left\{1, \frac{p^* \times q(X^*|X^{i-1})}{p^{i-1} \times q(X^{i-1}|X^*)}\right\}$$

4. Find the parameter vector X, corresponding to *minimum likelihood function* from the chain length of 1 to i-1 iterations and add a normally distributed random number (-1 to 1). Assign it as the new hyper parameter for the prior distribution.
5. Return to step 2 and continue until the number of iterations exceeds the specified limit

It is worth to mention here that in the above algorithm except step 4, all other steps are same as simple MCMC algorithm.

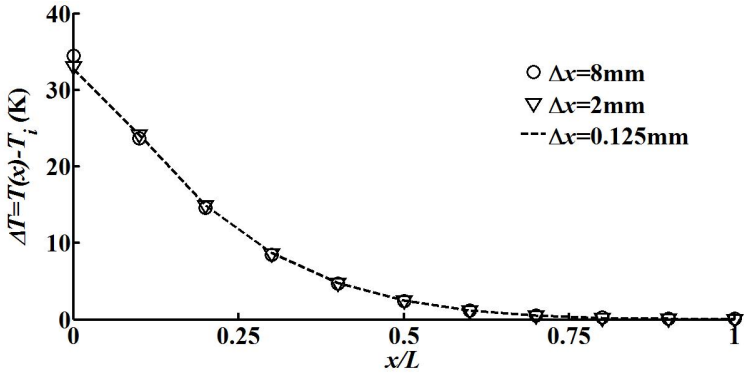


Figure 1: Grid independent study.

4 RESULTS AND DISCUSSION

The solution to the forward problem using FVM is written in Matlab and it is validated with commercial FVM based software package Fluent. A grid independent study is carried out and based on the study, a grid size of $\Delta x = 2$mm is chosen for further studies (refer Fig. 1).

The fin material is assumed to be steel and hence its thermo-physical properties are considered in the analysis for demonstration. It is also assumed that the temperature measurement is carried out in a natural environment; therefore the convective heat transfer coefficient is taken as 10W/m^{2}K. The duration and magnitude of heat input, total time of experiment and number of temperature measurements are decided based on D-optimal test [9]. The results of these preliminary studies are not presented here, since it is not the objective of the present work.

The results of multi-parameter estimation (k=25W/mk, h=10W/m^{2}K and ϵ=0.85) using both simple and present modified MCMC algorithm are presented for an error level of $\pm$0.03K and $\pm$0.3K respectively. The noisy temperature measurements are obtained by adding random errors with zero mean and standard deviation ($\pm$0.03/0.3K) to the solution of forward model.

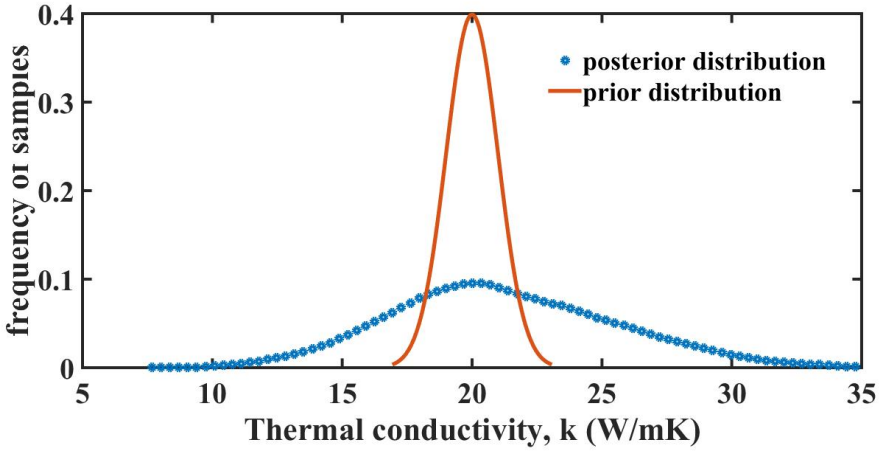


Figure 2: Sampling distribution Vs prior distribution for k (W/mK) obtained using simple MCMC algorithm.

Throughout the study the considered prior mean (μ_p) for k, h and ϵ are 20W/mK, 5W/m^{2}K and 0.6 respectively with standard deviation equal to $0.15\mu_p$. This multi-parameter estimation is especially difficult due to two reasons: (1) the posterior distribution is correlated and (2) due

to high measurement errors the inverse problem becomes more ill-posed. Also the combination of above two makes the estimation highly difficult.

Figure (2-4) shows the sampling distribution for k, h and ϵ respectively at standard deviation $\pm$0.3K obtained using simple MCMC algorithm and from these figures it is clear that the *sampling distribution* for all parameters are very close to prior distribution.

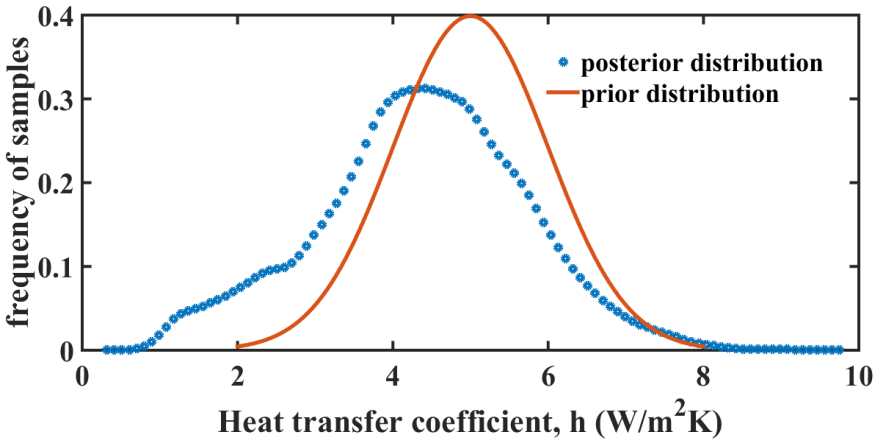


Figure 3: Sampling distribution Vs prior distribution for h (W/m^{2}K) obtained using simple MCMC algorithm.

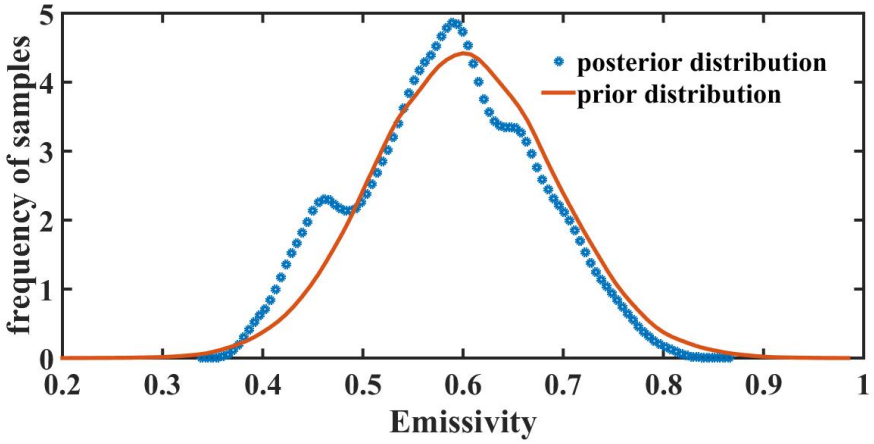


Figure 4: Sampling distribution Vs prior distribution for ϵ obtained using simple MCMC algorithm.

This parameter estimation problem is repeated again using the present modified MCMC algorithm and the resulting sampling distributions for all parameters at standard deviation $\pm$0.3K are shown in Fig. (5-7) and from these figures it is clear that the resulting *sampling distribution* for all parameters are normally distributed with mean close to the actual value for all the parameters. The estimates of the parameters ($\hat{k}$,$\hat{h}$ and $\hat{\epsilon}$) and spread of the samples ($\hat{\sigma}$) are calculated from the sampling distribution that are obtained using both simple and modified MCMC algorithms and are listed in table 1 and 2 respectively. Table 2 clearly shows that the estimates obtained using modified MCMC algorithm are in good agreement with the exact value even at

higher level of measurement error (±0.3K). In contrary, the estimates obtained using simple MCMC are in good agreement with the prior parameter, but far away from the true parameter (refer table 1)

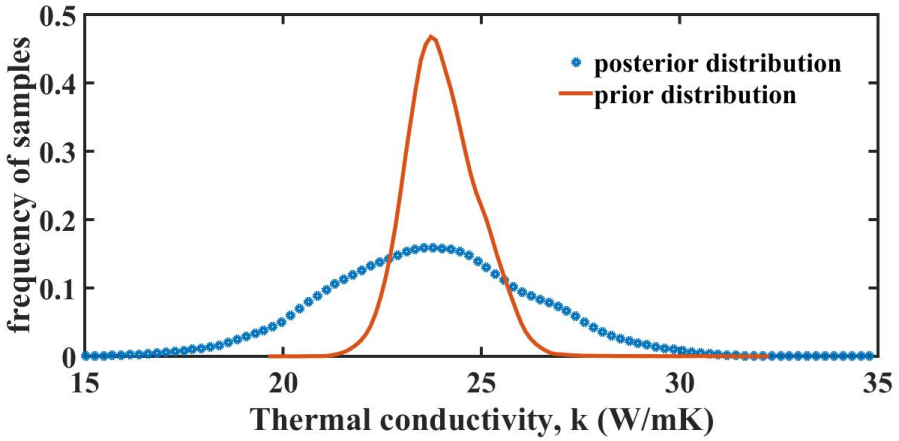


Figure 5: Sampling distribution Vs prior distribution for k (W/mK) obtained using present modified MCMC algorithm.

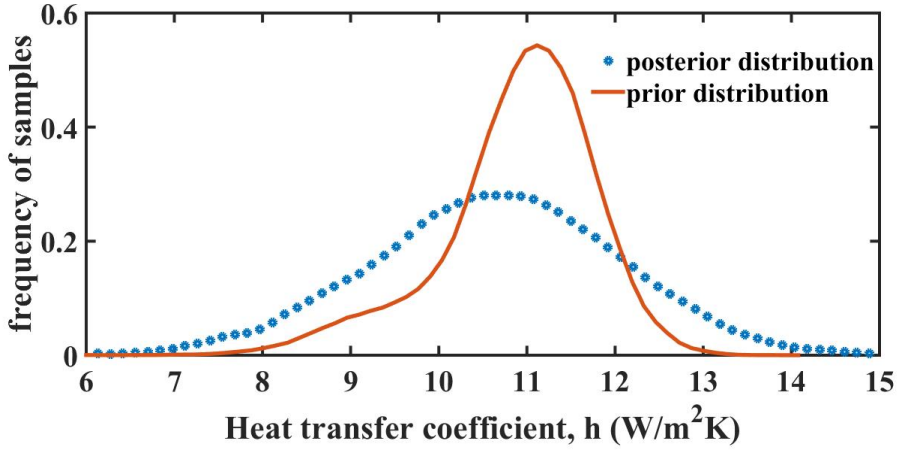


Figure 6: Sampling distribution Vs prior distribution for h (W/m^{2}K) obtained using present modified MCMC algorithm.

	k=25W/mK			h=10W/m^{2}K			ϵ=0.85		
	$\hat{k}$	$\hat{\sigma}_k$	% error	$\hat{h}$	$\hat{\sigma}_h$	% error	$\hat{\epsilon}$	$\hat{\sigma}_\epsilon$	% error
±0.03K	26.78	1.65	7.12	7.54	1.03	24.6	0.811	0.066	4.59
±0.3K	21.25	4.39	15	4.38	1.34	56.2	0.583	0.088	31.41

Table 1: Mean and standard deviation of the parameters calculated from sampling distribution obtained using simple MCMC.

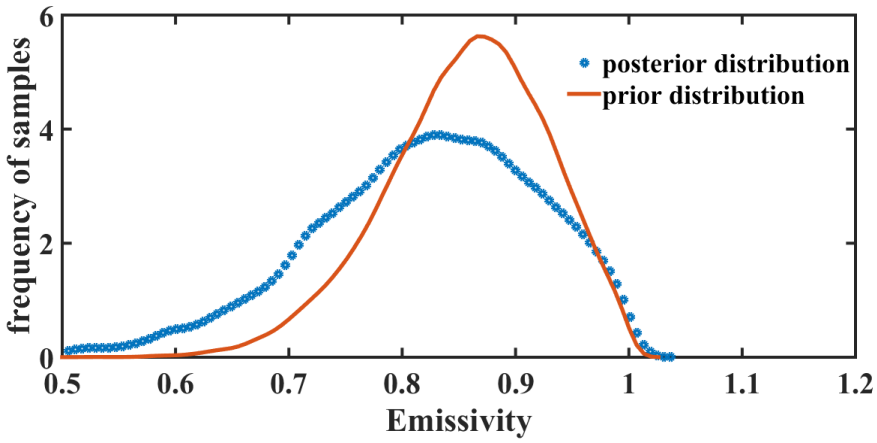


Figure 7: Sampling distribution Vs prior distribution for ϵ obtained using present modified MCMC algorithm.

	k=25W/mK			h=10W/m^{2}K			ϵ=0.85		
	$\hat{k}$	$\hat{\sigma}_k$	% error	$\hat{h}$	$\hat{\sigma}_h$	% error	$\hat{\epsilon}$	$\hat{\sigma}_\epsilon$	% error
$\pm$0.03K	24.92	1.4	0.32	9.38	1.03	6.2	0.862	0.066	1.41
$\pm$0.3K	23.71	2.6	5.16	10.61	1.5	6.1	0.821	0.099	3.41

Table 2: Mean and standard deviation of the parameters calculated from sampling distribution obtained using modified MCMC.

5 CONCLUSIONS

In this work, a modified MCMC algorithm was proposed in which the hyper parameter of the prior distribution was updated in each iteration based on the history of likelihood function. The idea of modifying prior parameter was put forth in order to restrict the sampling distribution moving close to the prior distribution at higher measurement uncertainties.

The proposed modified MCMC algorithm was tested for its reliability by employing it in inverse heat transfer problem, wherein three thermal properties were estimated simultaneously (the posterior for which is correlated). The estimation was carried out for both error and error less temperature measurements. The estimated parameters with modified MCMC were in good agreement with the true parameters, whereas with simple MCMC the estimated parameters were close to the supplied prior information at $\pm$0.3K measurement uncertainty. Since the posterior is correlated, the error associated with the parameter estimation using simple MCMC are more and a maximum deviation of 24.6% occurred in estimating h even at zero error level. Nevertheless, with modified MCMC, all the parameters were estimated with less than 7% deviation from the actual parameter value. Therefore, modified MCMC algorithm could be effectively used in inverse problems to estimate parameters from correlated posterior distribution even at high level of measurement errors.

REFERENCES

[1] Parthasarathy, Siddharth, and C. Balaji, Estimation of parameters in multi-mode heat transfer problems using Bayesian inferenceEffect of noise and a priori. *International Journal of Heat and Mass Transfer*, **51(9)**,2313-2334, 2008.

[2] Mota, Carlos AA, Helcio RB Orlande, Mario Olavo M. De Carvalho, Ville Kolehmainen, and Jari P. Kaipio, Bayesian estimation of temperature-dependent thermophysical properties and transient boundary heat flux. *Heat Transfer Engineering*, **31(7)**,570-580, 2010.

[3] Fudym, O., Orlande, H.R.B., Bamford, M. and Batsale, Bayesian approach for thermal diffusivity mapping from infrared images with spatially random heat pulse heating. *In Journal of Physics: Conference Series, IOP Publishing*, **135(1)**,012042, 2008.

[4] Orlande, H. R., Colao, M. J., and Dulikravich, G. S, Bayesian estimation of the thermal conductivity components of orthotropic solids *5th National Congress of Mechanical Engineering (CONEM 2008)*, Salvador, Bahia, Brazil, August 18-21, 2008.

[5] Patankar, Suhas, *Numerical heat transfer and fluid flow, 1st Edition*. CRC Press, 1980.

[6] Kaipio, Jari, and Erkki Somersalo, *Statistical and computational inverse problems, Vol. 160*. Springer Science & Business Media, 2006.

[7] Beck, James Vere, and Kenneth J. Arnold, *Parameter estimation in engineering and science*. James Beck, 1977.

[8] Gill, Jeff, *Bayesian methods: A social and behavioral sciences approach, Vol. 20*. CRC Press, 2014.

[9] Ozisik, M. Necat, *Inverse heat transfer: fundamentals and applications*. CRC Press, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

CONSTRAINED MULTI-POINT AERODYNAMIC SHAPE OPTIMIZATION OF THE VISCOUS DPW WING THROUGH EVOLUTIONARY PROGRAMMING AND SUPPORT VECTOR MACHINES

Daniel González-Juárez[1,2]**, Esther Andrés-Pérez**[3]**, and Mario J. Martin-Burgos**[1]

[1] Fluid Dynamics Branch, Spanish National Institute for Aerospace Technologies (INTA)
Ctra. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz
e-mail: {gonzalezjd,martinbj}@inta.es

[2] Thermal & Fluid Dynamics Department, Technical University of Cartagena (UPCT)

[3] Engineering Department, Ingeniería de Sistemas para la Defensa de España S.A. (ISDEFE-INTA) & Technical University of Madrid (UPM)
Ctra. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz
Email: eandres@isdefe.es, esther.andres@upm.es

Keywords: Aerodynamic shape design, Surrogate-based global optimization, Evolutionary programming, Computational Fluid Dynamics.

Abstract. *This work presents an application of Surrogate-Based Optimization (SBO) to the multipoint constrained design of the 3D DPW wing [1]in viscous transonic flow conditions. The geometry is parameterized by a control box with 36 design variables. An adaptive sampling technique focused on the optimization problem, the Intelligent Estimation Search with Sequential Learning (IES-LS), is applied. The selected SBO approach is based on the use of Support Vector Machines (SVMs) as the surrogate model for estimating the objective function, in combination with an evolutionary algorithm (EA) to enable the discovery of global optima. The aim of this work is to complement a previous one [2] by adding a study of the capability of this method to obtain an improvement for this multipoint constrained three-dimensional test case.*

1 INTRODUCTION

In the last few years, there has been an increasing interest in the topic of Surrogate-based Optimization (SBO) methods for aerodynamic shape design. This is due to the promising potential of these methods to speed-up the design process by the use of a "low cost" objective function evaluation to reduce the required number of expensive computational fluid dynamics (CFD) simulations. However, the industrial applications of these SBO methods has still to face several challenges, as for instance, the ability of surrogates when handling a high number of design parameters, efficient constraints handling, adequate exploration of the design space, etc.

The aim of this work is to complement a previous one [2] by adding a study of the capability of this method to obtain an improvement for a constrained multipoint three-dimensional viscous test case.

This work is under the scope of the GARTEUR Action Group (AD/AG52) [3], with the objective of providing a comprehensive survey about different surrogate methods for surrogate-based aerodynamic shape optimization, started at the beginning of 2013. Within this AG, research activities are planned over four-year period, with the objective of performing a fair comparison between different surrogate modeling methods applied to the aerodynamic optimization of baseline geometries, sharing the parameterization (volumetric NURBS) and mesh deformation algorithms..

2 PREVIOUS WORK

A physics-based surrogate model was recently applied in [4] to the drag minimization of a NACA0012 airfoil in inviscid transonic flow and RAE2822 airfoil in viscous transonic flow, both using PARSEC parameterization with up to ten design parameters. The drag minimization problem was also addressed by SBO in [5] for the NFL0416 airfoil, parameterized with ten design parameters.

Moreover, a combination of a generic algorithm (GA) and an artificial neural network (ANN) was applied in [6] to the shape optimization of an airfoil, parameterized by a modified PARSEC parameterization involving ten design variables. In [7] a surrogate based on Proper Orthogonal Decomposition (POD) was applied to the aerodynamic shape optimization of an airfoil geometry parameterized by sixteen design variables defined with Class Shape Transformation method (CST). In summary, the ability of SBO methods to manage a high number of design parameters still remains an open challenge and have been studied by several authors in the last few years, as well as the strategies for efficient infill sampling criteria with constraint handling. [7, 8].

Finally, the authors also presented recent works on this topic [2, 9]. This paper is an extension of previous research, here considering the constrained multi-point aerodynamic optimization of the DPW-W1 wing for viscous transonic flow.

3 PROPOSED APPROACH

A surrogate-based global optimization method with an adaptive sampling strategy is used, called 'The intelligent Estimation Search with Sequential Learning (IES-SL)'. Support Vector Machines for regression (SVMr) are combined with Evolutionary Algorithms (EA) in order to perform an efficient adaptive sampling guiding the optimization algorithm towards the most promising regions of the design space. The geometry is parameterized with volumetric Non-Uniform Rational B-Splines which vertical movements are the design variables for this study.

As can be observed, it comprises two steps: First, the algorithm generates an initial database by evaluating a small number of random designs (four in this application study). The initial

surrogate model is then generated using this reduced database. Then, the algorithm searches for the position of the optimum value with the surrogate model to use it as an estimation for the real optimum position [10]. The estimated optimum is evaluated using the CFD solver, obtaining a new pair [design, cost] that will enrich the database. After that, the surrogate is updated by adjusting it to the complete database and the cycle is finished, starting again the search for the new sample. When the maximum number of iterations is reached, the optimum design is obtained as the best parameters on the database. In this way, it is ensured that the design obtained is optimum with respect to the simulator system (CFD solver) and not only to the surrogate model. For more information about the SVMr, EAs and IES-SL readers can consult [5, 6, 11, 12]

4 DEFINITION OF THE OPTIMIZATION PROBLEM

4.1 Baseline geometry: DPW-W1 wing

The public domain transonic DPW-W1 wing (a test case of the Third AIAA Drag Prediction Workshop) was used [1, 10]. Reference quantities for this wing are displayed in the following table:

Sref (wing ref. area)	290322 mm^2
Cref (wing ref. chord)	197.55 mm
Xref*	154.24 mm
b/2 (semi span)	762 mm
AR (aspect ratio, AR=b2/Sref)	8.0

*(relative to the wing root leading edge)

Table 1. DPW reference quantities

The initial geometry (in IGES format) was downloaded from [10]. A set of grids are also available in the website of the 3rd AIAA Workshop on Drag Prediction.

4.2 Parameterization

The DPW geometry is parameterized by a 3D control box (displayed in Figure 1) with 5 control points in direction u, 10 in direction v and 5 in direction w. The parametric u direction corresponds to the y axis, the v direction to the x axis, and the w direction to the z axis.

The design variables are the vertical displacement of those control points set up on the aerodynamic surface. The wing is split in three profile sections and the transition between sections is linear. Each section has 6 active control points for the upper side and other 6 for the lower side, which are independent (the movement of a control point at the upper side does not modify the lower side and vice versa), with a total of 36 design parameters for the whole wing. Authors have previously applied this parameterization technique to other local and global optimization problems [11]. During the optimization performed in this paper, the wing platform will be kept fixed, as well the angle of attack and the torsion.

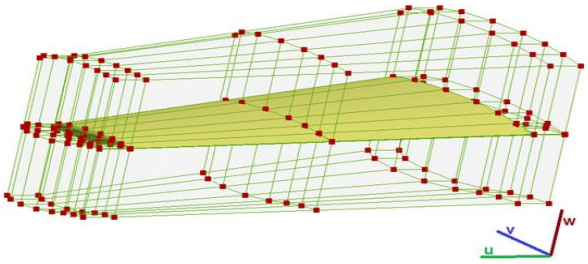


Figure 1. DPW wing parameterization

4.3 Aerodynamic constraints

The following aerodynamic constraints are considered:

1. Prescribed constant lift coefficient (C_L=C_L^0)
2. Minimum pitching moment: C_M>=C_M^0
3. Drag penalty: If constraint in minimum pitching moment is not satisfied, the penalty will be 1 drag count per 0.01 increment in C_M.

4.4 Geometric constraints

Each design variable will be constrained by its minimum and maximum values that will be chosen as the + or – 20% of their original value. Apart from this, other constraints have been defined, according to [1]:

1) Airfoils' maximum thickness constraints:

$$(t/c)_{section} \geq (t/c)^0_{section}$$

where the right term is the maximum thickness for the original wing sections, root, mid-span and tip, which has the value of 13.5%.

2) Beam constraints:

First, two locations (x/c) are fixed to represent the beam constraints:

$$(x/c)_{root,1} = (x/c)_{mid-span,1} = (x/c)_{tip,1} = 0.20$$
$$(x/c)_{root,2} = (x/c)_{mid-span,2} = (x/c)_{tip,2} = 0.75$$

The constraint here is that the thickness value of the optimized wing sections at these locations should be greater or equal than the thickness of the original ones. It is defined with the expressions:

$$(t/c)_{root,1} \geq 12\%, (t/c)_{mid-span,1} \geq 12\%, (t/c)_{tip,1} \geq 12\%$$
$$(t/c)_{root,2} \geq 5.9\%, (t/c)_{mid-span,2} \geq 5.9\%, (t/c)_{tip,2} \geq 5.9$$

4.5 Test cases definition & objective function

The proposed approach is applied to the multipoint aerodynamic shape optimization of a DPW-W1 wing in viscous transonic flow. The specific flow conditions are exposed in Table 2 for both Design Points (DP)

	DP1	**DP2**
M	0.76	0.78
Re	5×10^6	5×10^6
AoA	0	0
Turbulence	SA	SA

Table 2. Test cases definition

The design goal is to achieve a geometry with the minimum drag, while maintaining the specified aerodynamic constraints. Aerodynamic constraints are implemented as penalties in the objective function. The pseudo-code implementations is:

```
lift_penalty=1-(Cl/Cl0);
if (lift_penalty<0) lift_penalty=0;
cm_penalty = (Cm0-Cm)*0.0001/0.01;
if (cm_penalty < 0) cm_penalty = 0;
objective_function=(((Cd+cm_penalty)/Cd0 ) )+5*lift_penalty;
```

4.6 Computational grid

The DPW RANS grid was directly downloaded from [10]. The features of the unstructured grid are:

#points	#surface points	#elements	#surface elements
3770k	152k	9335k	310k

Table 3. Computational grid features

5 NUMERICAL ASSESSMENT

5.1 DPW-W1 optimization results

This section shows the results of the optimization approach exposed in previous sections. Table 4 summarizes the DPW the results of the optimization process. Results exhibit a reduction of 3 & 6 drag counts respectively for each DP which are in the same order that the results obtained in [1]

	DP1			DP2			fobj
	C_L	C_D	C_M	C_L	C_D	C_M	
DPW-W1	0.3632	0.0205	-0.0674	0.3718	0.0222	-0.0692	
Optimized	0.3637	0.0202	-0.0677	0.3712	0.0216	-0.0695	0.9746
Δ	0.0005	-0.0003	-0.0003	0.0006	-0.0006	-0.0003	

Table 4.DPW-W1 optimization results

Figure 2 depicts the baseline and optimized airfoils along wing span.

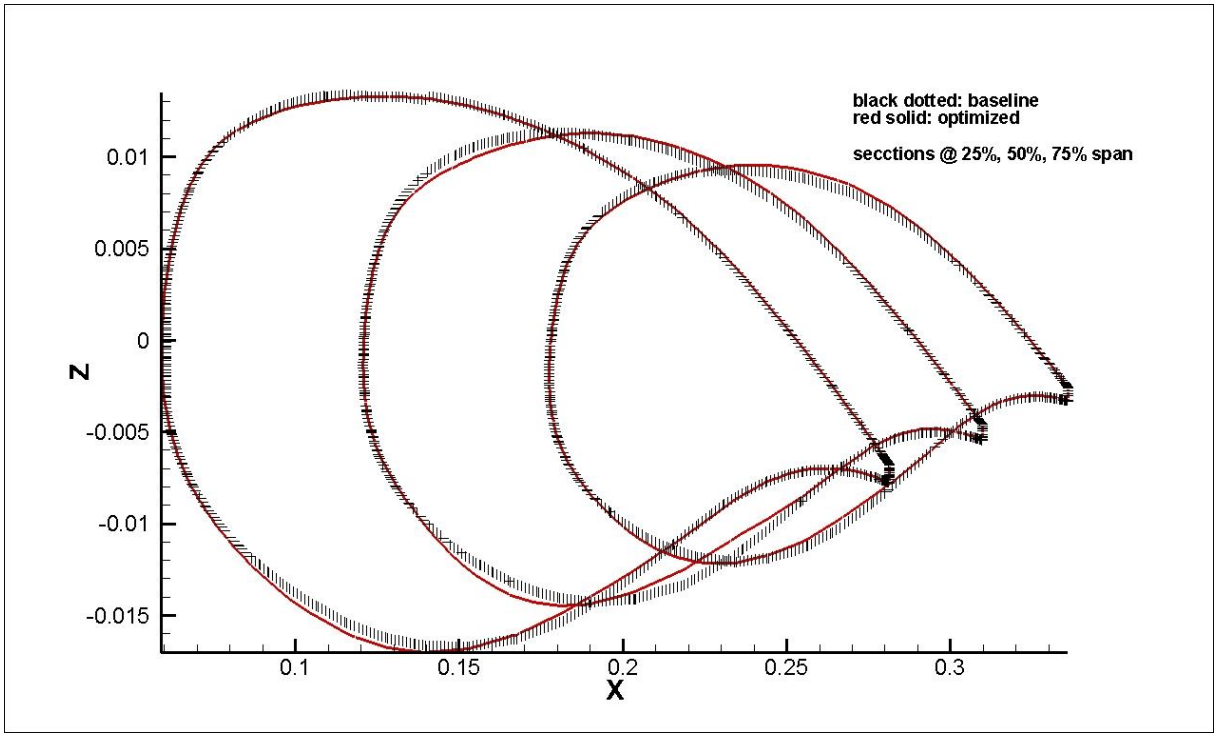


Figure 2. Baseline vs. optimized geometry

Figure 3 andFigure 4 show the C_p distribution @ 25% 50% and 75% of wing span while Figure 5 andFigure 6 contains the C_p contours for both baseline and optimized geometries at each flow conditions.

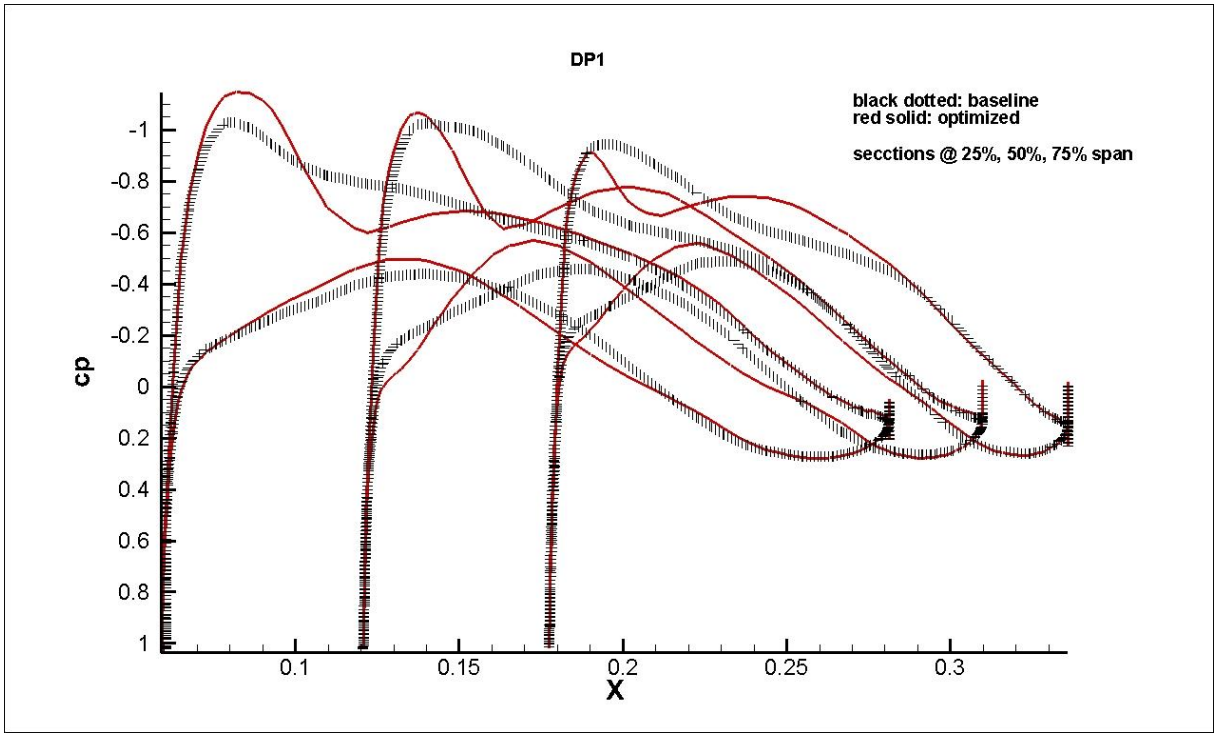


Figure 3. Cp distribution for DP1 along wing span

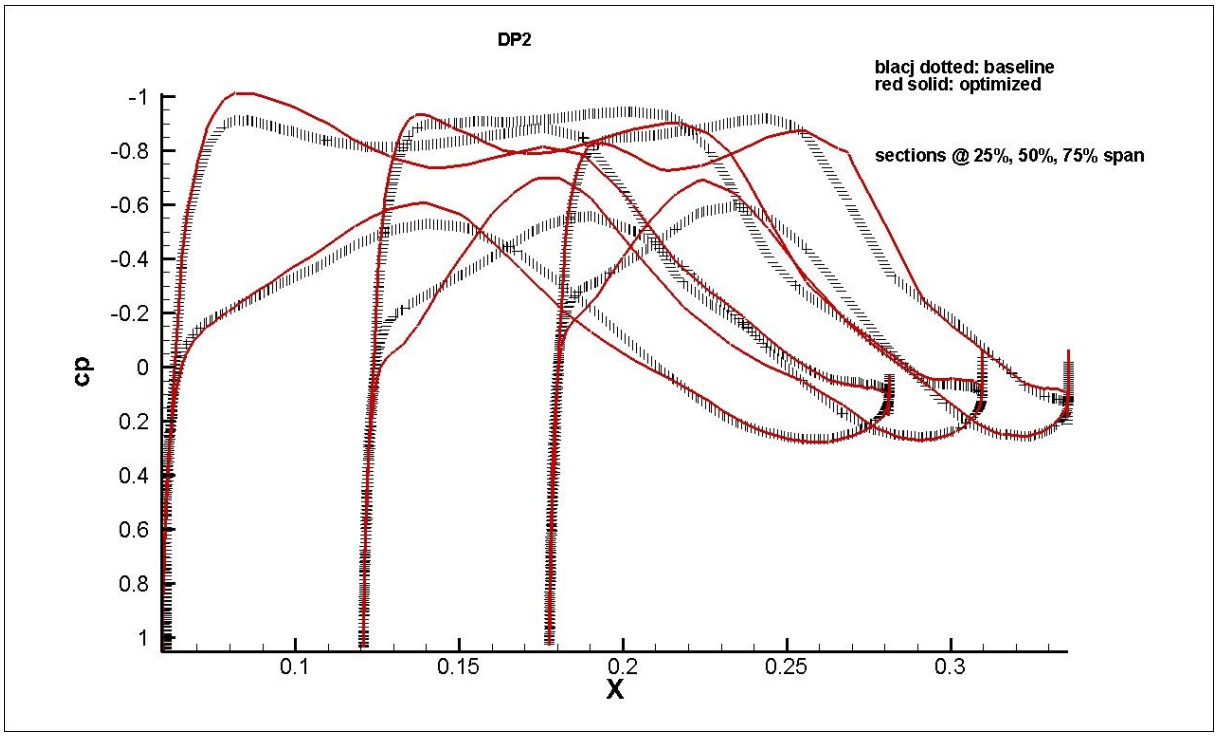


Figure 4. Cp distribution for DP2 along wing span

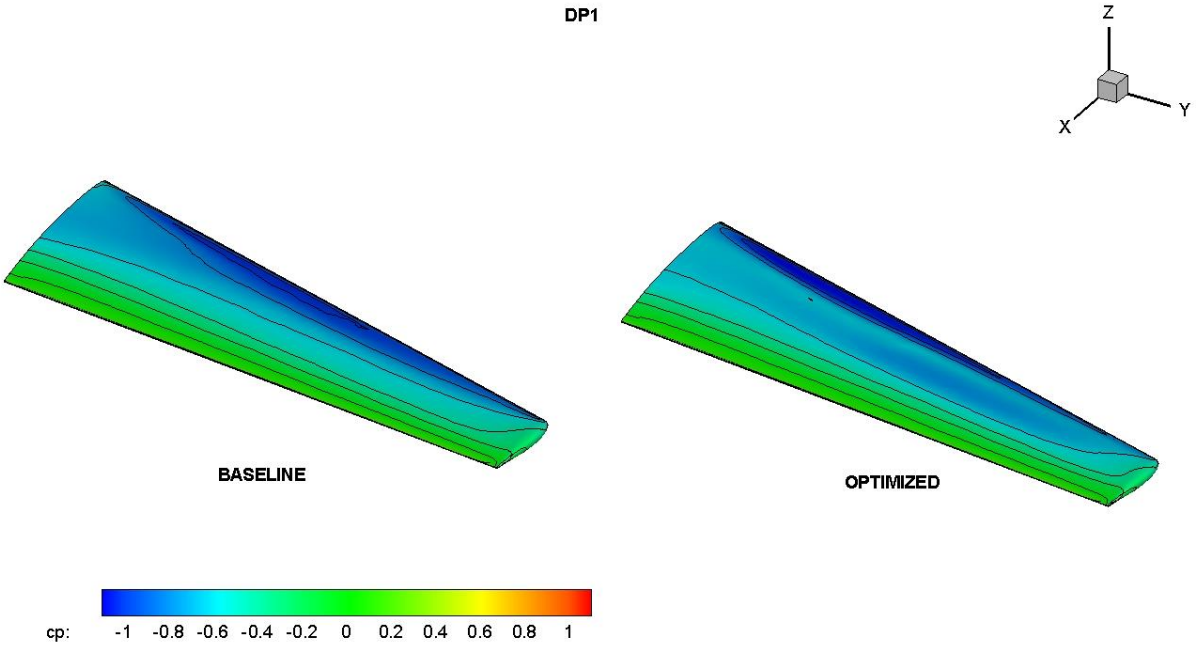


Figure 5. Cp contours for baseline and optimized geometries @ DP1 flow conditions

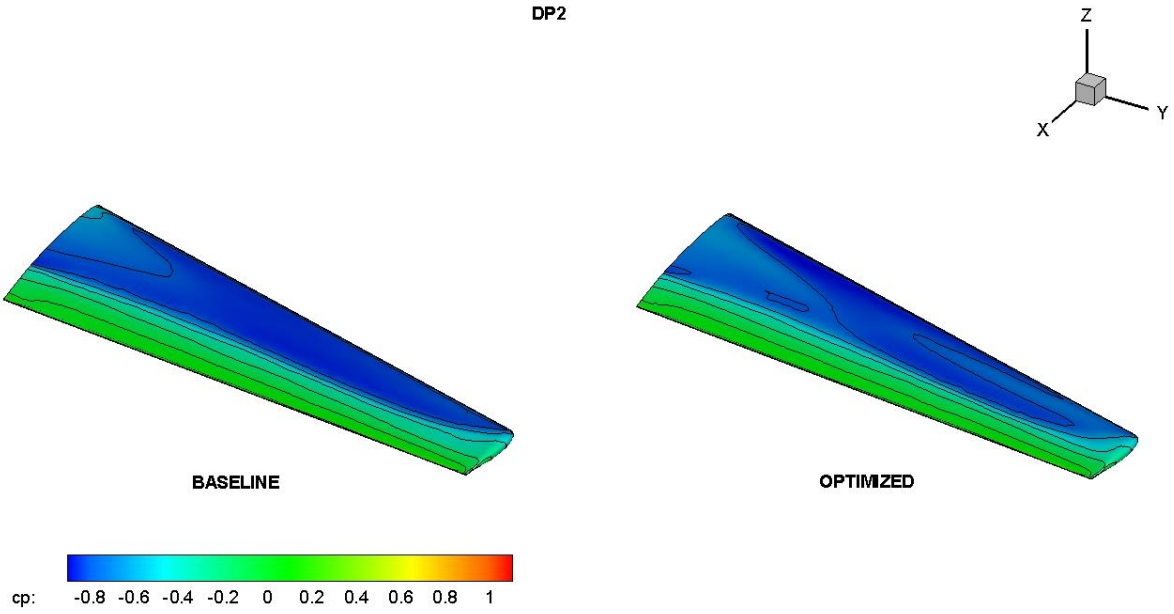


Figure 6. Cp contours for baseline and optimized geometries @ DP2 flow conditions

6 CONCLUSION

This paper presented the application of a global optimization strategy using the Intelligent Estimation Search with Sequential Learning (IES-SL) and the hybridization of EA and SVMr to the multi-point constrained optimization of a three dimensional DPW wing in viscous transonic flow conditions, showing first promising results. Future work will focus on the combination of this approach with traditional gradient-based methods to perform a deep comparison and also an enrichment of the search space in the evolutionary optimization algorithm.

REFERENCES

[1] B. Epstein, A. Jameson, S. Peigin, D. Roman, N. Harrison y J. Vassberg, «Comparative Study of 3D Wing Drag Minimization by Different Optimization Techniques.,» de *46th AIAA Aerospace Sciences Meeting and Exhibit. AIAA paper 2*, Reno, Nevada, 2008.

[2] E. Andrés, D. González, M. J. Burgos y L. Carro, «Constrained Single-Point Aerodynamic Shape Optimization of the DPW-W1 wing through Evolutionary Programming and Support Vector Machines,» de *International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control (EUROGEN 2015)*, Glasgow, 2015.

[3] «Garteur AD/AG52 members 2013. Partial report on Surrogate-based Global Optimization Methos in Aerodynamic Design. http://ag52.blogspot.com.es/».

[4] L. Leifsson, S. Koziel y Y. Tesfahungen, «Aerodynamic Design Optimization: Physics-based Surrogate Approaches for Airfoil and Wing Design,» de *AIAA SciTech*, 2014.

[5] S. Koziel y L. Leifsson, «Multi-Level Surrogate-Based Airfoil Shape Optimization,» de *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, Grapevine (Dallas/Ft. Worth Region), Texas, 2013.

[6] A. Jahangirian y A. Shahrokhi, «Aerodynamic shape optimization using efficent evolutionary algorithms and unstructured CFD solver,» *Computers & FLuids*, vol. 46, pp. 270-276, 2011.

[7] E. Iuliano y D. Quagliarella, «Aerodynamic Shape Optimization via non-intrusive POD-based Surrogate Modeing,» de *IEEE Congress on Evolutionary Computation*, Cancún, Mexico, 2013.

[8] J. Parr, C. Holden, A. Forrester y A. Keane, «Review of Efficent Surrogate Infill Sampling Criteria with Constraint Handling,» de *2nd International Conference on Engineering Optimization*, Lisbon, Portugal, 2010.

[9] D. González, E. Andrés, M. Burgos, L. Carro y S. Salcedo, «Influence of geometry parameterization in aerodynamic shape design of aeronautical configurations by evolutionary algorithms,» de *6th European Conference for Aeronautics and Space Sciences (EUCASS)*, Krakow, Poland, 2015.

[10] E. Andrés, S. Salcedo-Sanz, F. Mongue y A. Pérez-Bellido, «Efficent aerdoynamic design through evolutionary programming and support vector regression algorithms,» *Expert Systems with Applications*, vol. 39, pp. 10700-10708, 2012.

[11] Y. Jin, «Multi-Level Surrogate-Based Airfoil Shape Optimization,» de *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, Grapevine (Dallas/Ft. Worth Region), Texas AIAA2013-0778., 2013.

[12] E. Ortiz-García, S. S. Sanz, Á. M. Pérez-Bellido y J. A. Portilla-Figueras, «Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions,» *Neurocomputing,* 2009.

[13] «AIAA 3rd Drag Prediction Workshop. URL: http://aaac.larc.nasa.gov/tsab/cfdlarc/aiaa-dpw/Workshop3/workshop3.html,» [On line].

[14] M. Burgos, E. Andrés, E. Valero y C. Lozano, «Gradients Calcultion for Arbitrary Parameteizations via Volumetric NURBS: The Control Box Approach,» de *EUCASS*, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SURROGATE-BASED GLOBAL OPTIMIZATION OF A CYLINDER BY THE USE OF EVOLUTIONARY ALGORITHMS, SUPPORT VECTOR MACHINES AND NON UNIFORM B-SPLINES

Esther Andrés-Pérez[1], **Daniel González-Juárez**[2], **Mario J. Martin-Burgos**[2], **Leopoldo Carro-Calvo**[3] **and Sancho Salcedo-Sanz**[3]

[1] Engineering Department, Ingeniería de Sistemas para la Defensa de España S.A. (ISDEFE-INTA) & Technical University of Madrid (UPM)
Ctra. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz
Email: eandres@isdefe.es, esther.andres@upm.es

[2] Fluid Dynamics Branch, Spanish National Institute for Aerospace Technologies (INTA)
Ctra. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz
e-mail: {gonzalezjd,martinbj}@inta.es

[3] University of Alcalá (UAH). Dept. of Signal Theory and Communications
Ctra. Madrid-Barcelona km. 33. 28802 Alcalá de Henares
e-mail: {leopoldo.carro,sancho.salcedo}@uah.es

Keywords: Aerodynamic shape design, Surrogate-based global optimization, Evolutionary programming, Computational Fluid Dynamics.

Abstract. *Nowadays, one of the priorities of the European Commission is to reduce the environmental impact of aviation through the advanced design of novel aircraft configurations. This implies that new methods and tools for aerodynamic shape optimization will have to be developed, allowing aircraft configurations that cannot be obtained with traditional strategies. Evolutionary optimization algorithms have the potential to find a global optimal candidate but on the other hand, they involve a vast number of evaluations, which are CFD runs in aerodynamic analysis. The use of surrogate modeling has been proposed in the literature [1, 2] as a suitable method to speed up the global optimization process. This work presents an application of a surrogate-based global optimization method (SBGO) to the aerodynamic shape optimization of a 2D infinite circular cylinder in laminar flow conditions. The geometry of this test case is parameterized by a Non-Rational B-Splines (NURBS) control box with 10 design variables. An approach based on Support Vector Machines (SVMs) in combination with Evolutionary Algorithms (EA) is applied.*

1 INTRODUCTION

Currently there is a strong need of computational tools for the design of the type of aircraft that will be demanded by the European industry, according to the guidelines stated at the ACARE 2020 [3] and 2050 [4] flight paths. The aeronautical industry agrees that these objectives make necessary the design of an innovative aircraft shape rather than further local improvements in the traditional wing-body-tail configuration. Efficient and accurate shape design optimization tools, able to consider novel concepts through the use of flexible geometry parameterization, are becoming a must for the aeronautical industry. Considering this, aerodynamic shape design and optimization problems based on evolutionary algorithms and surrogate models (also called surrogate-based optimization or SBO) have recently found widespread use in aeronautics, due to the potential to reach optimal configurations that are far away from their baseline geometries, and therefore their ability to enable non-conventional aircraft configurations. In addition, their increasing applicability in aerodynamic shape optimization problems is also due to the promising potential of these methods to speed-up the whole design process by the use of a "low cost" objective function evaluation to reduce the required number of expensive computational fluid dynamics (CFD) simulations.

However, the application of these SBO methods for industrial configurations still requires facing several challenges. The most crucial challenges nowadays are the so-called "curse of dimensionality", the ability of surrogates when handling a high number of design parameters, efficient constraints handling, adequate exploration and exploitation of the design space, and last but not least, how to deal with grid deformations in case of large displacements, which is always the case when trying to achieve novel configurations from the traditional ones.

This work focuses on the application of enhanced methods in aerodynamic shape design optimization to enable novel aircraft configurations. In particular, it aims to demonstrate the feasibility of a combined approach, based on Evolutionary algorithms and Support Vector Machines, to reach optimal configurations that are far away from its baseline geometry. In order to validate this, the optimization approach is applied to the selected baseline geometry, a landing gear master cylinder, resulting on optimal configurations for each of the defined flow conditions. This very simple test case (clean cylinder) has been selected for several reasons: it will allow to validate the potential of the proposed approach to reach non-conventional configurations (those which are far from the initial one), and in addition, it is of interest for an European aircraft manufacturing industry, which is looking for flow optimization in this region. However, in order to further exploit the results in industry, more complex geometries and constraints, including also structural aspects should have to be taken into consideration.

2 PREVIOUS WORK

A physics-based surrogate model was recently applied in [5] to the drag minimization of a NACA0012 airfoil in inviscid transonic flow and RAE2822 airfoil in viscous transonic flow, both using PARSEC parameterization with up to ten design parameters. The drag minimization problem was also addressed by SBO in [6] for the NFL0416 airfoil, parameterized with ten design parameters.

Moreover, a combination of a generic algorithm (GA) and an artificial neural network (ANN) was applied in [7] to the shape optimization of an airfoil, parameterized by a modified PARSEC parameterization involving ten design variables. In [8] a surrogate based on Proper Orthogonal Decomposition (POD) was applied to the aerodynamic shape optimization of an airfoil geometry parameterized by sixteen design variables defined with Class Shape Transformation method (CST). In summary, the ability of SBO methods to manage a high number of design parameters still remains an open challenge and have been studied by several authors in the last

few years, as well as the strategies for efficient infill sampling criteria with constraint handling. [8, 9].

3 PROPOSED APPROACH

A surrogate-based global optimization method with an adaptive sampling strategy is used, called 'The intelligent Estimation Search with Sequential Learning (IES-SL)'. Support Vector Machines for regression (SVMr) are combined with Evolutionary Algorithms (EA) in order to perform an efficient adaptive sampling guiding the optimization algorithm towards the most promising regions of the design space. The geometry is parameterized with volumetric Non-Uniform Rational B-Splines which vertical movements are the design variables for this study.

As can be observed, it comprises two steps: First, the algorithm generates an initial database by evaluating a small number of random designs (four in this application study). The initial surrogate model is then generated using this reduced database. Then, the algorithm searches for the position of the optimum value with the surrogate model to use it as an estimation for the real optimum position [1]. The estimated optimum is evaluated using the CFD solver, obtaining a new pair [design, cost] that will enrich the database. After that, the surrogate is updated by adjusting it to the complete database and the cycle is finished, starting again the search for the new sample. When the maximum number of iterations is reached, the optimum design is obtained as the best parameters on the database. In this way, it is ensured that the design obtained is optimum with respect to the simulator system (CFD solver) and not only to the surrogate model. For more information about the SVMr, EAs and IES-SL readers can consult [6, 7, 10, 11].

4 DEFINITION OF THE OPTIMIZATION PROBLEM

4.1 Geometry parameterization

The cylinder grid is deformed through a volumetric b-spline, as shown in Figure 1. The design variables are the control points located on the upper and lower side, which can freely move in any direction, while the control points located in the middle are kept fixed during optimization. In this method, the original geometry is deformed by the movement of control points in a similar way than the Free Form Deformation technique (FFD) [12], but in contrast to FFD, deformations of the upper-side and lower-side are considered independently one of each other, which provides more flexibility.

The computational surface grid vertices are mapped into the NURBS (Non-Uniform Rational B-Splines) space through the parametric coordinates, which are previously calculated using an appropriate inversion point algorithm [12, 13]). These parametric coordinates are invariant throughout the optimization, allowing to recalculate the spatial coordinates at any time of the process. A second mapping is performed on the cylinder geometry, by means of a discrete uniform rasterization, in order to accurately calculate the volume throughout the optimization.

This geometry mapping is done in parallel, independently of the computational grid and it is used for handling the volume constraints within the optimization process.

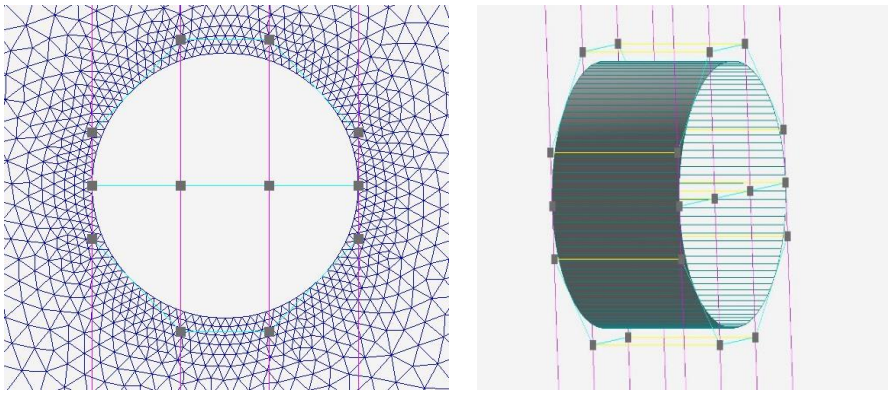


Figure 1. Parameterization of the cylinder. 2D (left) and 3D (right) views.

4.2 Test cases definition

The approach is applied to the aerodynamic shape optimization of a cylinder parameterized as shown in section 4.1, with the problem formulation defined on Table 1. The location of the design parameters on the surface of the test case was previously displayed in Figure 1. A symmetric movement of the upper and lower face control points was imposed. The objective function was to minimize drag while preserving, at least, the 80% of the baseline volume, which was considered the minimum valid volume due to structural requirements. This volume preservation was implemented as a strong penalization of the objective function, which allows exploring the whole design space even if several geometries will finally be discarded by the optimization algorithm. Furthermore, DV's are allowed to move ± 60% of their initial value in both directions, horizontal and vertical.

M	**Re**
0,05	1214
0,1	2490
0,15	3735
0,2	4980
0,25	6226
0,3	7471
0,35	8716
0,4	9961

Table 1. Problem formulation (the Re reference length considered is the cylinder diameter)

The optimization study was performed for a range of Mach numbers from 0.05 and 0.4, meaning a Reynolds below 10^5 (Table 1), ensuring laminar boundary layer separation conditions, as can be observed in Figure 2 [15]. This problem formulation allows using laminar flow conditions, therefore reducing the required computational cost (compared with turbulence RANS modeling).

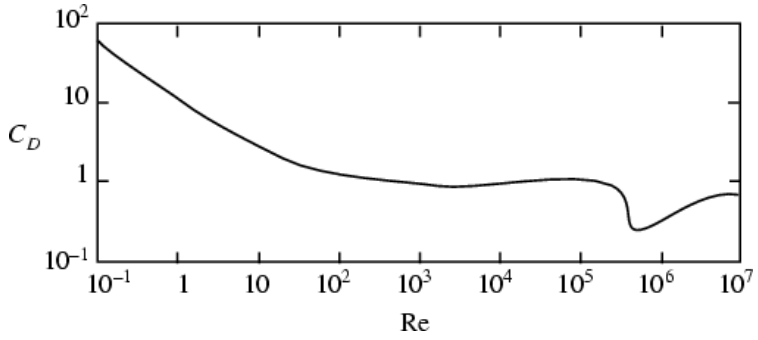


Figure 2. Drag coefficient versus Reynolds number for an infinite circular cylinder [15].

4.3 Drag minimization of a 3D cylinder for different flow conditions

In this section, the approach is applied to the drag minimization and results are displayed in Table 2 & Table 3. In particular, Table 2 shows the optimization results regarding the whole objective function (which includes both drag and volume values considerations), where it can be seen that the total reduction of the objective function was between 73-77% of its original value. Table 3 shows the drag coefficient values of the original and optimized geometries. It can be observed that the drag was minimized between 92-94% of its original value in the baseline geometry (depending on the Mach number considered), while at the same time fulfilling the constraints imposed in the volume.

M	**0.05**	**0,1**	**0,15**	**0,2**	**0,25**	**0,3**	**0,35**	**0,4**
$OF_{baseline}$	0,39248	0,30123	0,26356	0,24341	0,23173	0,22510	0,22124	0,21910
OF_{optim}	0,08918	0,07525	0,06751	0,06385	0,06031	0,05784	0,05740	0,05817
%Improvement	77,28%	75,02%	74,39%	73,77%	73,97%	74,31%	74,05%	73,45%

Table 2. Optimization results (OF means objective function)

M	**0,05**	**0,1**	**0,15**	**0,2**	**0,25**	**0,3**	**0,35**	**0,4**
C-$drag_{baseline}$	0,95672	0,84988	0,80707	0,78708	0,77838	0,77638	0,77909	0,78578
C-$drag_{optim}$	0,06924	0,05842	0,05241	0,04957	0,04682	0,04490	0,04472	0,04570
C-$drag_{optim\_p}$	0,05391	0,04875	0,04507	0,04353	0,04162	0,04033	0,04057	0,04189
C-$drag_{optim\_v}$	0,01532	0,00967	0,00733	0,00604	0,00520	0,00457	0,00415	0,00381
%Improvement	92,76%	93,13%	93,51%	93,70%	93,98%	94,22%	94,26%	94,18%

Table 3. Optimization results (C-drag minimization)

From the mentioned tables, it can be also observed that the gain in the objective function tends to decrease with the Mach number while, on the other hand, the improvement in the drag coefficient tends to increase. This behavior is explained because the optimizer proposes thinner shapes as the Mach number increases, producing a geometry with less drag, but also less volume, which is penalized in its global objective function.

The optimized shapes returned by the optimizer are displayed in Figure 3. For clearness, only the baseline geometry and the optimized geometries for Mach numbers 0.10, 0.20, 0.30

and 0.40 are shown. It can be observed that all the optimized shapes are similar except the one returned for Mach=0.40, where the optimizer returns a geometry with a wider area near the trailing edge, in order to ensure the volume constraint, even when it will affect the drag value.

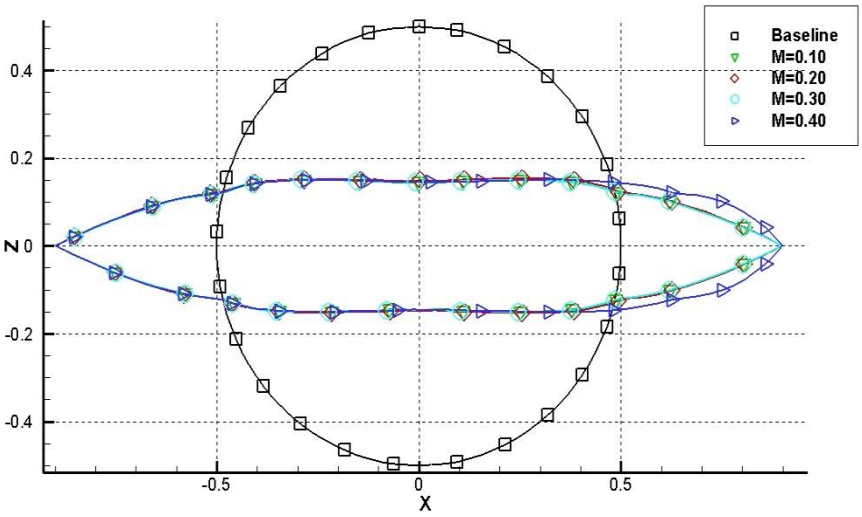


Figure 3. Comparison of baseline and optimal shapes for Mach numbers 0.10, 0.20, 0.30 and 0.40

Finally, **¡Error! No se encuentra el origen de la referencia.** shows the Mach number contours of the original (left) and optimized (right) geometries for each of the Mach numbers considered in the range [0.05, 0.4]. As expected, a pair of vortices (bigger with the Mach number) appear downstream of the baseline geometry. In the optimized shapes, the cross-sectional area has been reduced as much as the geometric and volume constraints allows. This explains the vortices disappearance and the drag reduction, as expected from the aerodynamic point of view, since the objective function was to reduce drag while maintaining the volume. Some asymmetric effects can be seen in the Mach contours of the optimized geometries, which are due to asymmetric volume grid deformation since the geometric surface deformation is propagated to the volume.

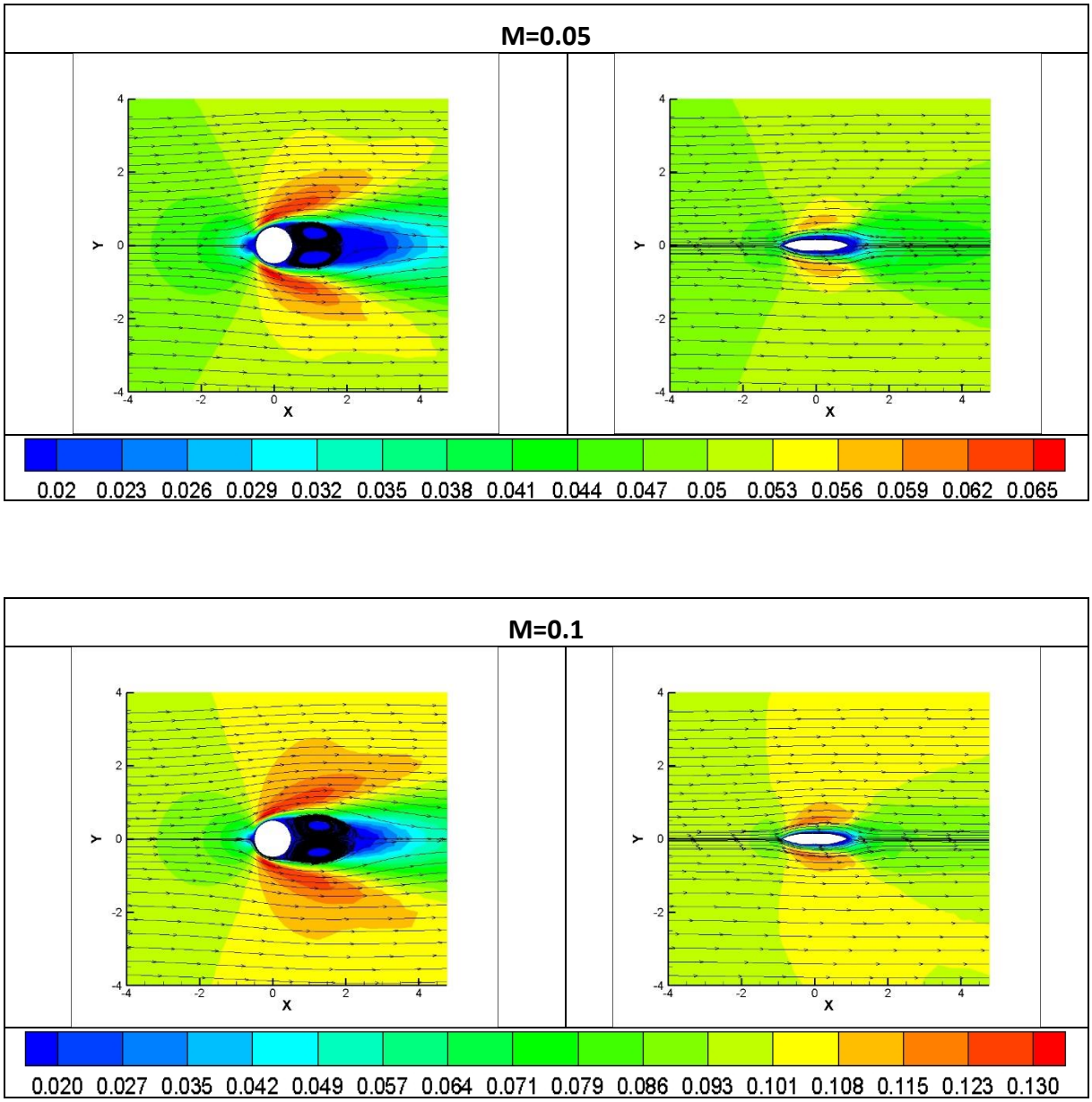


Figure 4. Mach contours and velocity streamlines of the baseline and optimized geometries for Mach numbers 0.05 and 0.1

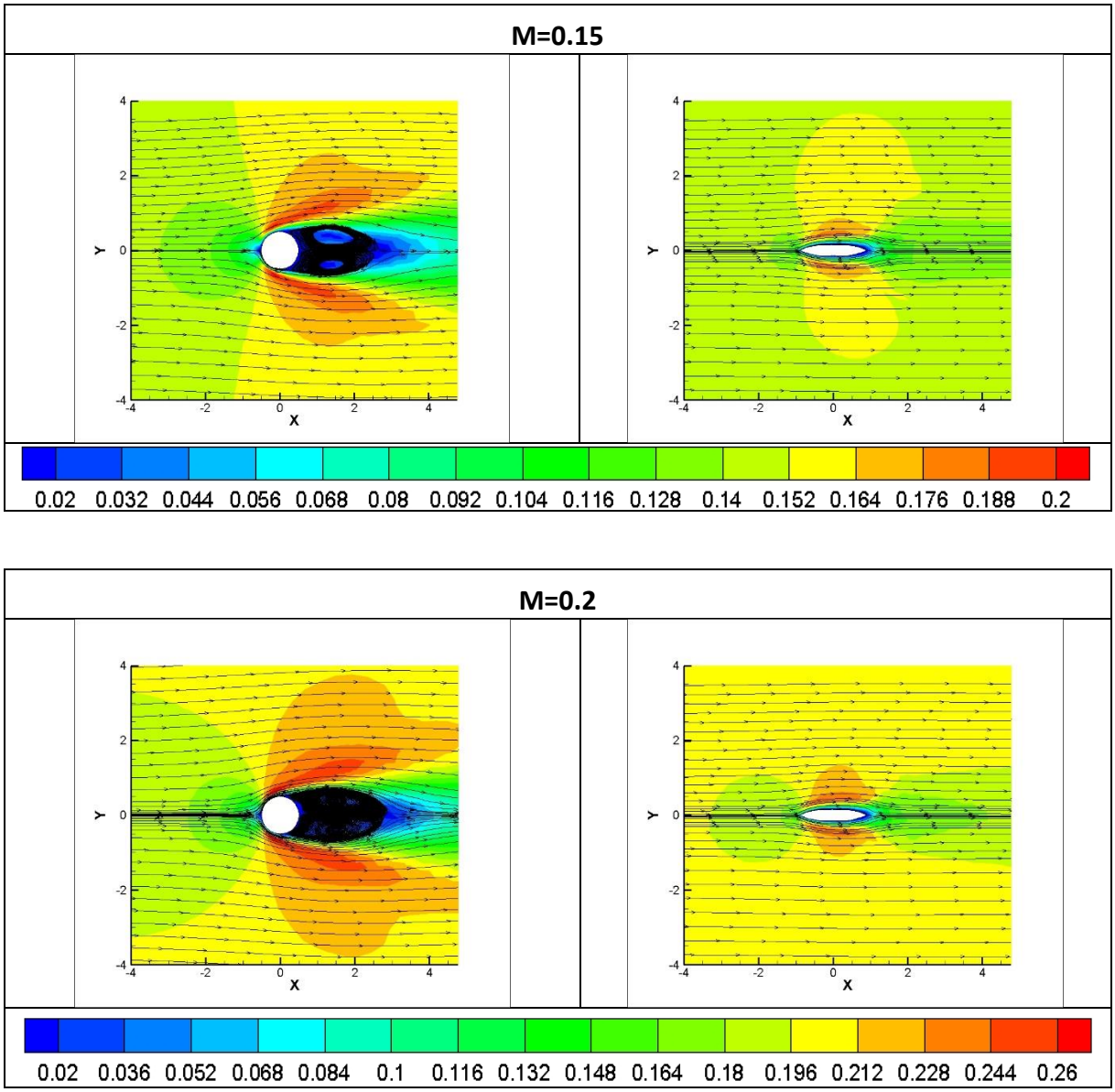


Figure 5. Mach contours and velocity streamlines of the baseline and optimized geometries for Mach numbers 0.15 and 0.2

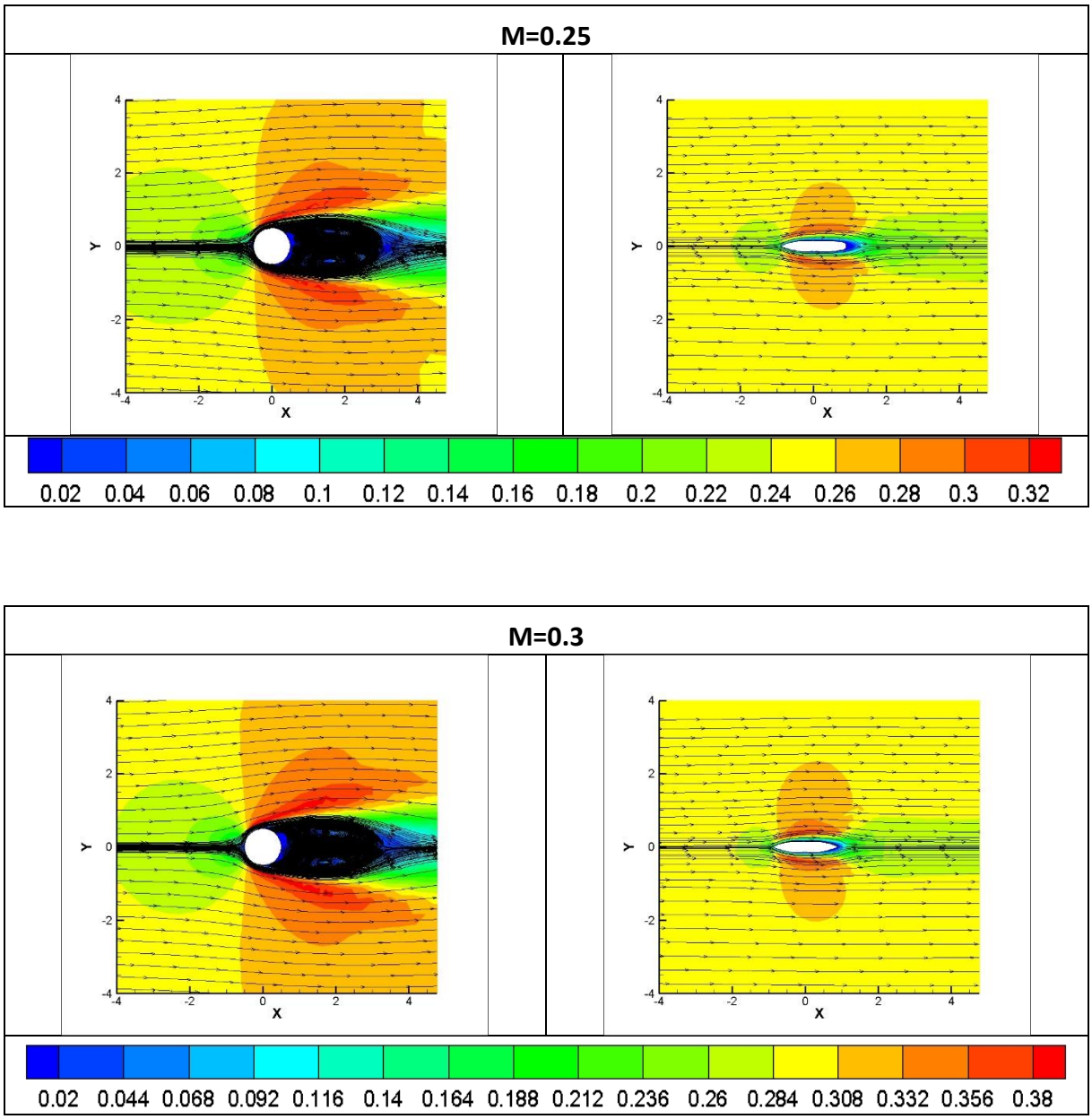
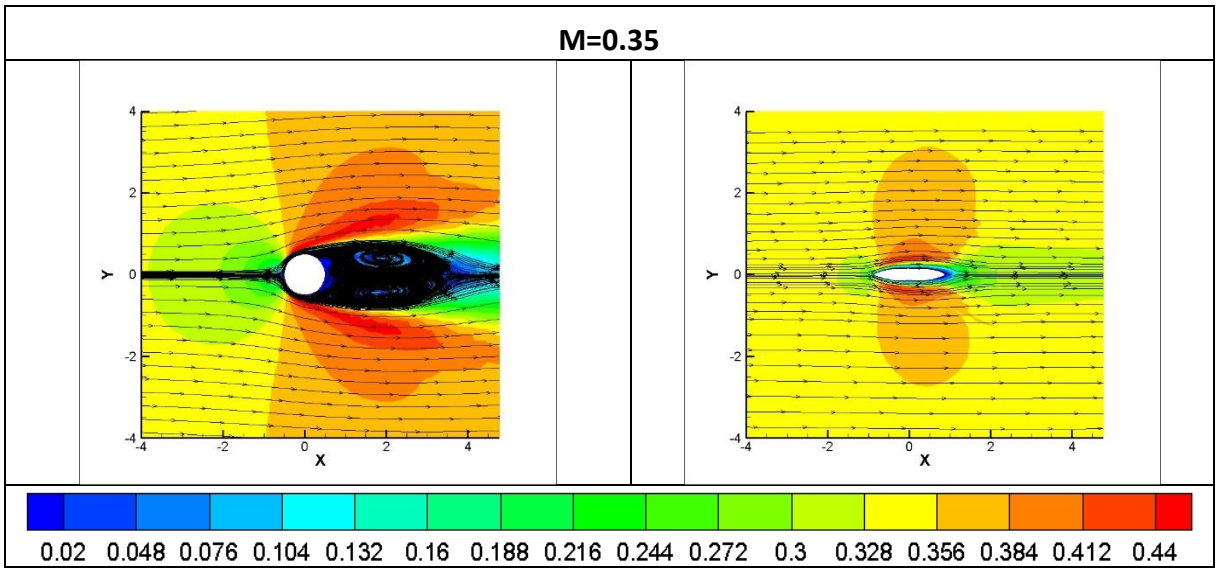


Figure 6. Mach contours and velocity streamlines of the baseline and optimized geometries for Mach numbers 0.25 and 0.3



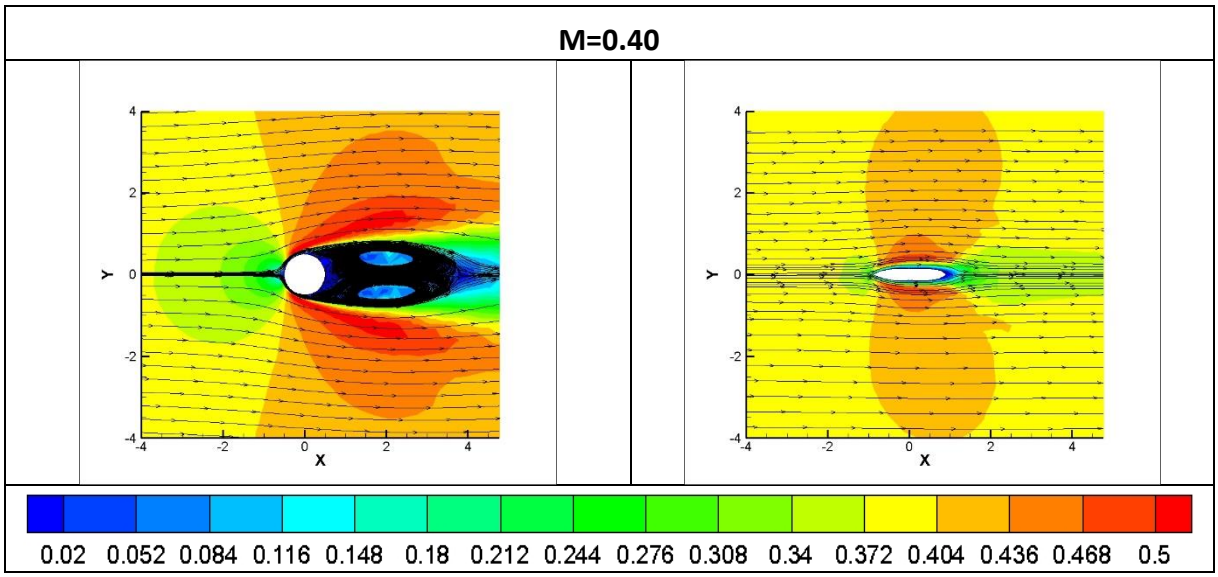


Figure 7. Mach contours and velocity streamlines of the baseline and optimized geometries for Mach numbers 0.35 and 0.4

5 CONCLUSION

This article presented the application of a global optimization strategy using the Intelligent Estimation Search with Sequential Learning (IES-SL) and the hybridization of EA and SVMr to the aerodynamic shape optimization of a clean cylinder representing a simple model of the landing gear master cylinder. The objective of this work was to demonstrate the feasibility of the proposed strategy to reach optimal configurations that are far from the baseline geometry.

This approach allows extensively exploring the design space, without any dependence on an initial solution and expensive CFD computations, since it uses a metamodel (based on SVMr) to estimate the aerodynamic coefficients. At the same time, accuracy is ensured, because in each iteration the result is validated with the CFD tool.

REFERENCES

[1] E. Andrés, S. Salcedo-Sanz, F. Mongue y A. Pérez-Bellido, «Efficent aerdoynamic design through evolutionary programming and support vector regression algorithms,» *Expert Systems with Applications,* vol. 39, pp. 10700-10708, 2012.

[2] A. I. Forrester y A. J. Keane, «Recent advances in surrogate-based optimization,» *Progress in Aerospace Sciences,* vol. 45, nº 1-3, p. 5079, 2009.

[3] ACARE, «Aeronautics and Air Transport: Beyond vision 2020 (towards 2050),» Advisory Council or Aeronautics Research in Europe, 2010.

[4] P. Argüelles y M. Bischoff, «European Aeronautics: A vision for 2020. ACARE».

[5] L. Leifsson, S. Koziel y Y. Tesfahungen, «Aerodynamic Design Optimization: Physics-based Surrogate Approaches for Airfoil and Wing Design,» de *AIAA SciTech*, 2014.

[6] S. Koziel y L. Leifsson, «Multi-Level Surrogate-Based Airfoil Shape Optimization,» de *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, Grapevine (Dallas/Ft. Worth Region), Texas, 2013.

[7] A. Jahangirian y A. Shahrokhi, «Aerodynamic shape optimization using efficent evolutionary algorithms and unstructured CFD solver,» *Computers & FLuids,* vol. 46, pp. 270-276, 2011.

[8] E. Iuliano y D. Quagliarella, «Aerodynamic Shape Optimization via non-intrusive POD-based Surrogate Modeing,» de *IEEE Congress on Evolutionary Computation*, Cancún, Mexico, 2013.

[9] J. Parr, C. Holden, A. Forrester y A. Keane, «Review of Efficent Surrogate Infill Sampling Criteria with Constraint Handling,» de *2nd International Conference on Engineering Optimization*, Lisbon, Portugal, 2010.

[10] Y. Jin, «Multi-Level Surrogate-Based Airfoil Shape Optimization,» de *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, Grapevine (Dallas/Ft. Worth Region), Texas AIAA2013-0778., 2013.

[11] E. Ortiz-García, S. S. Sanz, Á. M. Pérez-Bellido y J. A. Portilla-Figueras, «Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions,» *Neurocomputing,* 2009.

[12] D. Chauhan, C. Praveen y R. Duvigneau, «Wing Shape Optimization Using FFD and Twist Parametrization,» *Aerospace Sciences and Technologies,* pp. 225-230, 2010.

[13] M. Martin, E. Andrés, M. Widhalm, P. Bitrian y C. Lozano, «Non-Uniform Rational B-Splines based aerodynamic shape optimization with DLR TAU code,» *Proceedings of the Institution of Mechanical Engineers, part G, Journal of Aerospace Engineering,* vol. 226, nº 10, pp. 10-13, 2012.

[14] M. Martin, E. Andrés, E. Valero y C. Lozano, «Volumetric B-splines shape parametrization for aerodynamic shape,» *Aerospace Science and technology,* vol. 37, pp. 26-36, 2014.

[15] D. Tritton, Physical FLuid Dynamics 2nd Ed, Oxford University Press, 1998.

[16] B. Epstein, A. Jameson, S. Peigin, D. Roman, N. Harrison y J. Vassberg, «Comparative Study of 3D Wing Drag Minimization by Different Optimization Techniques.,» de *46th AIAA Aerospace Sciences Meeting and Exhibit. AIAA paper 2*, Reno, Nevada, 2008.

[17] E. Andrés, D. González, M. J. Burgos y L. Carro, «Constrained Single-Point Aerodynamic Shape Optimization of the DPW-W1 wing through Evolutionary Programming and Support Vector Machines,» de *International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control (EUROGEN 2015)*, Glasgow, 2015.

[18] «Garteur AD/AG52 members 2013. Partial report on Surrogate-based Global Optimization Methos in Aerodynamic Design. http://ag52.blogspot.com.es/».

[19] D. González, E. Andrés, M. Burgos, L. Carro y S. Salcedo, «Influence of geometry parameterization in aerodynamic shape design of aeronautical configurations by evolutionary algorithms,» de *6th European Conference for Aeronautics and Space Sciences (EUCASS)*, Krakow, Poland, 2015.

[20] «AIAA 3rd Drag Prediction Workshop. URL: http://aaac.larc.nasa.gov/tsab/cfdlarc/aiaa-dpw/Workshop3/workshop3.html,» [En línea].

[21] M. Martin, E. Andrés, E. Valero y C. Lozano, «Gradients Calcultion for Arbitrary Parameteizations via Volumetric NURBS: The Control Box Approach,» de *EUCASS*, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

PERFORMANCE COMPARISON OF KRIGING AND SVR SURROGATE MODELS APPLIED TO THE OBJECTIVE FUNCTION PREDICTION WITHIN AERODYNAMIC SHAPE OPTIMIZATION

D. Viúdez-Moreiras[1], E. Andrés-Pérez[2], D. González-Juárez[1] and M. J. Martin Burgos[1]

[1] Fluid Dynamics Branch, Spanish National Institute for Aerospace Technology (INTA),
Ctra. Ajalvir, km. 4. 28850 Torrejón de Ardoz (Spain)
viudezmd@inta.es, dviudezmoreiras@gmail.com
{gonzalezjd,martinmj}@inta.es

[2] Engineering Department, Ingeniería de Sistemas para la Defensa de España S.A. (ISDEFE-INTA) & Technical University of Madrid (UPM)
Ctra. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz (Spain)
eandres@isdefe.es, esther.andres@upm.es

Keywords: Surrogate Models, Kriging, Support Vector Regression, Aerodynamic coefficients prediction, computational fluid dynamics, surrogate-based global optimization

Abstract. *All optimization methods depend on some form of internal model of the problem space they are exploring. To build such a model when there are many variables can require a large number of analyses to be carried out. Because of these difficulties, it is now common in aerospace design to manage explicitly the building and adaptation of the internal surrogate model used during optimization. However, it is very difficult to know a priori which surrogate model is more suitable for a specific application. As described above, in this paper, the performance of two surrogate models, Kriging and Support Vector Regression is compared in order to choose the most suitable model targeting a future application within an aerodynamic shape optimization process. The selected test case was the DPW wing, from the AIAA Workshop on drag prediction, also used in the GARTEUR AD-AG52 on "Surrogate-based global optimization for aerodynamic shape design."*

1 INTRODUCTION

Surrogate modeling is referred to a group of techniques that make use of previously obtained sampled data in order to build surrogate models, which aims to predict the value of variables at untried points in the design space. These groups of techniques, also known as metamodeling, have been developed from many different disciplines including statistics, mathematics, computer science, and various engineering disciplines.

Different techniques have been studied in the literature, which can be classified in three categories [1] such as, (I) the response surface method (RSM) with optimization of coefficients for a base function, (II) the neural network approximation (NN) and (III) an estimation method using observed values at sampling locations to compute an estimated value at an optional location in a solution space. Although these all can be used practically in industry, each method has different features that have to be taken into account before the application to a particular problem. Several comparisons among those methods have been previously described in [2–5].

The RSM is one of the very effective approaches for modeling with small numbers of design variables and its solution space is not too complex, being successfully used in some optimization problems in engineering [6–9]. However, the RSM usually requires the assumption of the order of the approximated base function because the approximation process is performed using the least-square method for the function coefficients. Therefore, the knowledge of the qualitative trend of the entire design space is required by the designer, which will sometimes be difficult to determine. This problem will be highlighted as the number of design variables increases.

NN has been used for solving difficult modeling problems [10-11]. NN generally minimizes the sum of the approximation errors at sampling locations, so that the accuracy of the approximated value at a sampling location is relatively high. However, NN implies high computational cost incurred for learning stage and the need for the designer to be skilled or experienced in using NN [2].

Estimation methods such as Kriging (KR) [12–15], Radial Basis Functions (RBF) [16] or Support Vector Regression (SVR) [17,18] usually require more sample points in the solution space than RSM or NN within the training stage, in order to perform an accurate estimation [19,20]. However, they allow to build complex high non-linear models [21-24] which it is very difficult to achieve with RSM or too complex with NN. They are then found to be a valuable tool to support a wide scope of activities in modern engineering design such as chemical and materials engineering [25,26], and other fields such as agriculture and ecology [27,28], medicine [29-31] and economy [32].

One of the main model applications can be found within the aerospace field [17]. Thus, the use of long running expensive simulations in design leads to a fundamental problem when trying to compare and contrast various competing options: is very expensive from the computational resources point of view to analyze all combinations of variables in the design space. This problem is particularly highlighted when using optimization schemes. All optimization methods depend on some form of internal model of the problem space they are exploring. To build such a model when there are many variables can require large numbers of analyses to be carried out. Because of these difficulties it is now common in aerospace design to manage explicitly the building and adaptation of the internal model used during optimization (Surrogate Based Optimization, SBO).

Thus, the performance obtained by the surrogate model in SBO schemes is very important in order to minimize the number of iterations in the design process, which implies expensive Computational Fluid Dynamics (CFD) simulations in a high

performance computer (HPC). However, it is very difficult to know a priori which surrogate model is more suitable for a specific application.

This paper presents a comparison between two surrogate techniques, KR and SVR, for the prediction of aerodynamic coefficients and objective functions for different aircraft's wing geometries, in order to have more information for choosing the most suitable surrogate for this kind of application. In order to perform this comparison, both KR and SVR have been applied to the DPW-wing [33, 34], used by the GARTEUR AD/AG-52 Group [35], with the same initial database composed of different geometries. This database is derived with high fidelity CFD simulations for use in the initial stage of the design process. V-fold cross validation [36, 37] has been used to compare both methods.

This paper is structured as follows: Section 2 briefly describes both surrogate model theory, sampling and model validation methodologies used in this research. Section 3 describes the database to be modeled by the surrogate models. Section 4 presents the comparative results between KR and SVR obtained in the application to the wing case described in the previous section. Finally, Section 5 presents the conclusions.

2 SURROGATE MODEL THEORY

This section briefly describes KR and SVR surrogate modelling. Then, the sampling and the model validation methodologies used in this research are stated.

2.1 Kriging (KR)

This section recalls the basic KR algorithm, as an interpolation technique. Full description of KR algorithm and modifications of basic implementation can be found in [1,16,19-21,24].

KR can be seen as a two-step process. First, a regression function f(**x**) is constructed based on the data, and, subsequently, a Gaussian process Z(**x**) is constructed through the residuals [1,38]. Thus, the prediction $\hat{y}(\boldsymbol{x})$ in **x** is derived by:

$$\hat{y}(\boldsymbol{x}) = f(\boldsymbol{x}) + Z(\boldsymbol{x}) \tag{1}$$

where f(**x**) is a regression function and Z is a Gaussian process with mean 0, variance σ^2 and a correlation matrix ψ.

Depending on the form of the regression function Kriging has been prefixed with different names. *Simple Kriging* assumes the regression function to be a known constant, i.e., f(**x**)=0. A more popular version is *Ordinary Kriging*, which assumes a constant but unknown regression function f(**x**), and *Universal Kriging*, which assumes other more complex trend functions such as linear or quadratic polynomials. In general, *Universal Kriging* treats the trend function as a multivariate polynomial:

$$f(x) = \sum_{i=1}^{p} \alpha_i \, b_i(x) \tag{2}$$

where $b_i(\boldsymbol{x})$ are $i = (1 \ldots p)$ basis functions and $\alpha = (\alpha_1 \ldots \alpha_p)$ denotes the coefficients. Therefore, the regression function captures the general trend of the data and the Gaussian Process interpolates the residuals. However, selecting the correct regression function is a difficult problem, hence, the regression function is often chosen constant.

Consider a set of n samples, X = { $\mathbf{x^1}, \dots, \mathbf{x^n}$ } in *d* dimensions and associated function values, **y** = { $y^1, \dots, y^n$ }. Essentially, the regression part is encoded in the n x p model matrix F:

$$F = \begin{pmatrix} b_1(\boldsymbol{x^1}) & \cdots & b_p(\boldsymbol{x^1}) \\ \vdots & \ddots & \vdots \\ b_1(\boldsymbol{x^n}) & \cdots & b_p(\boldsymbol{x^n}) \end{pmatrix}$$

while the stochastic process is mostly defined by the n x n correlation matrix ψ,

$$\psi = \begin{pmatrix} \psi(\boldsymbol{x^1}, \boldsymbol{x^1}) & \cdots & \psi(\boldsymbol{x^1}, \boldsymbol{x^n}) \\ \vdots & \ddots & \vdots \\ \psi(\boldsymbol{x^n}, \boldsymbol{x^1}) & \cdots & \psi(\boldsymbol{x^n}, \boldsymbol{x^n}) \end{pmatrix}$$

where $\psi(\boldsymbol{x^i}, \boldsymbol{x^j})$ is the correlation function. $\psi(\boldsymbol{x^i}, \boldsymbol{x^j})$ is parametrized by a set of hyperparameters $\boldsymbol{\theta}$, which are identified by Maximum Likelihood Estimation (MLE) [39,40]. Subsequently, the prediction mean and prediction variance of KR are derived, respectively, as,

$$\mu(\boldsymbol{x}) = M\alpha + r(\boldsymbol{x})\, \psi^{-1}\, (\mathbf{y} - \mathbf{F}\alpha) \tag{3}$$

$$s^2(\boldsymbol{x}) = \sigma^2\, (\, 1 - r(\boldsymbol{x})\, \psi^{-1} r(\boldsymbol{x})^T + \frac{1-(\, F^T\, \psi^{-1} r(\boldsymbol{x})^T}{F^T\, \psi^{-1}\, F}) \tag{4}$$

where $M = (\, b_1(\boldsymbol{x}) \dots\, b_p(\boldsymbol{x})\,)$ is the model matrix of the predicting point **x**, and $r(\boldsymbol{x}) = (\, \psi(\boldsymbol{x}, \boldsymbol{x^1}) \dots\, \psi(\boldsymbol{x}, \boldsymbol{x^n})\,)$ is an 1 x n vector of correlations between the point **x** and the samples X. The process variance σ^2 is given by:

$$\sigma^2 = \frac{1}{\mathrm{n}}\, (\mathbf{y} - \mathbf{F}\alpha)^T\, \psi^{-1}\, (\mathbf{y} - \mathbf{F}\alpha) \tag{5}$$

and the coefficients of the regression function, α, are determined by Generalized Least Squares (GLS) by:

$$\alpha = (\, F^T\, \psi^{-1}\, F\,)^{-1}\, F^T\, \psi^{-1} \boldsymbol{y} \tag{6}$$

2.2 Support Vector Machines (SVR)

SVR can be solved as a convex optimization problem using kernel theory to face non-linear problems. Thus, SVR consider not only the prediction error but also the generalization of the model [17,41].

SVR consist of training a model with the form $y = w^T \phi(x) + b$ given a set of parameters $C = \{(x_i, y_i)\}, i = 1,2, \dots, l\}$, to minimize a general risk function of the form:

$$R[f] = \frac{1}{2}\left|\left|w\right|\right|^2 + \frac{1}{2} C \sum_{i=1}^{l} L\big(y_i, f(x)\big) \tag{7}$$

where w controls the smoothness of the model, $\phi(x)$ is a function of projection of the inputs space w to the feature space, b is a parameter of bias, x_i is a feature vector of the input space with dimension N, y_i is the output value to be estimated and $L\big(y_i, f(x)\big)$ is the loss function selected. In this study, the $L1$ support vector regression $(L1 - SVR)$ is used,

$$L\big(y_i, f(x)\big) = |y_i + f(x_i)|_\varepsilon \tag{8}$$

In order to train this model, it is necessary to solve the following optimization problem:

$$\min\left(\frac{1}{2}\left|\left|w\right|\right|^2 + \frac{1}{2}C\sum_{i=1}^{l}\xi_i + \xi_i^*\right) \tag{9}$$

subject to:

$$\begin{gathered} \mathrm{y_i} - w^{\mathrm{T}}\,\phi(x) - b \leq\ \varepsilon + \xi_i, i = 1, \dots, l \\ -\mathrm{y_i} + w^{\mathrm{T}}\,\phi(x) + b \leq\ \varepsilon + \xi_i^*, i = 1, \dots, l \\ \xi_i\xi_i^* \geq 0, i = 1, \dots, l \end{gathered} \tag{10}$$

To do this, a dual form is usually applied, obtained from the minimization of the Lagrange function that joins the function to minimize and the constraints:

$$\begin{aligned} \max\Bigg(-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i + \alpha_i^*)\big(\alpha_j + \alpha_j^*\big)K\big(x_i + x_j\big) \\ -\ \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i + \alpha_i^*)\Bigg) \end{aligned} \tag{11}$$

subject to the following constraint:

$$\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) = 0, \alpha_i, \alpha_i^* \in [0, C] \tag{12}$$

In addition to the constraints, also must be taken in account the Karush-Kuhn-Tucker conditions and obtain the bias value. In the dual formulation, the apparition of the kernel function $K\big(x_i, x_j\big)$ must be emphasized, which is equivalent to the scalar product$\big\langle\phi(x_i)\big|\phi(x_j)\big\rangle$. In this case, the kernel function is a Gaussian function:

$$K = \exp(-\gamma \cdot \left|\left|x_i - x_j\right|\right|^2) \tag{13}$$

The final form of the regression model depends on the Lagrange multipliers α_i, α_i^*, following the expression:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{14}$$

In this way, SVR model depends on three parameters, ε, C and γ. (I) ε parameter controls the error margin permitted for the model, as can be seen in equations (9) and (10). (II) C parameter controls the number of outliers allowed on the optimization of the equation (3). Finally, (III) γ parameter determines the Gaussian variance for the kernel. Depending on

the selection of these values, the model can have different performance. To obtain the best SVR performance, search of the most suitable combination of these three parameters must be performed, usually by means of cross validation techniques over the training set, as explained in Section 2.4. To reduce the computational time of this process, different methods have been proposed in the literature to reduce the search space related to these parameters. In this case, it has been applied the one developed in [41], which has proven to require pretty short search times.

2.3 DESIGN OF EXPERIMENTS (DOE) METHODOLOGY

Both KR and SVR surrogate models have been applied to the database generated with the DLR TAU solver [42] for a set of initial wing geometries. Therefore, a suitable Design of Experiments (DoE) technique, that envelopes the design space, is required.

When the initial database is produced by a deterministic computer code, as opposed to a physical experiment or stochastic analysis, a given input will always yield the same output, because there is no measurement error or other random sources of noise. Under these conditions, the DoE need only be space-filling [43,44] so that all regions of the design space Ω as a subset of $\mathbb{R}^{NS}$, being NS the number of independent variables of the design space, are sampled. A commonly used space-filling design is Latin hypercube sampling (LHS) [45], which has been used as sampling methodology in this study in order to perform the comparison between KR and SVR. In LHS, each input parameter is partitioned into N equally spaced sections. Each input parameter is sampled once in each section, resulting in a column vector $\boldsymbol{x}_{\mathrm{i}}$ containing NS different values of the input parameter. The column vectors for each input parameter are arranged side by side into a matrix and the components of the vectors are then randomly reordered. The resulting training set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ is a matrix of size N x NS, known as a Latin hypercube, in which each row corresponds to a different training case defined by the input parameters.

In this particular case, NS columns correspond to the z-axis coordinates of different control points distributed by the wing surface to optimize. N sets of coordinates are derived, in order to obtain the aerodynamic coefficients by means of CFD and, in this way, perform the initial database in order to train in the same conditions both surrogate models.

2.4 METRICS FOR MODEL VALIDATION

In most real applications, only a limited amount of data is available, which leads to the idea of splitting the data: part of data, the *training sample*, is used for training the algorithm, and the remaining data, the *validation sample*, is used for evaluating the performance of the algorithm. The *validation sample* can play the role of "new data".

A single data split yields a validation estimate of the risk, and averaging over several splits yields a cross-validation estimate. V-fold CV (VFCV) was introduced in [46] and now is widely used as a model validation technique within surrogate modelling. Thus, the methodology used in this paper to obtain the model performance in order to obtain the surrogate models comparison is, for a training set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$, the following:

a. Partition the training set X into K independent equal-sized subsets X_k, $k = 1, \dots, K$ such as $X = X_1 \cup X_2 \cup \dots \cup X_K$;
b. for $k = 1, \dots, K$
    i. train the prediction model on $X_{(-k)} = X \backslash X_k$;
    ii. test the prediction model on X_k

iii. $\forall\, \boldsymbol{x}_i \in X_k$ compute the values of $\hat{y}^{-k(i)}(\boldsymbol{x}_i)$,

iv. Derive the quadratic error between the real y_i and the estimated $\hat{y}^{-k(i)}(\boldsymbol{x}_i)$ values:

$$\text{SE}= \left(y_i - \hat{y}^{-k(i)}(\boldsymbol{x}_i)\right)^2 \tag{15}$$

c. Estimate the following metrics:

i. Mean Squared Error of $\hat{y}$

$$MSE(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}^{-k(i)}(\boldsymbol{x}_i)\right)^2 \tag{16}$$

ii. R-squared

$$R^2 = 1 - \frac{\sum_i (y_i - y_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{17}$$

where $\bar{y}$ is the mean value of the true values y_i.

iii. Pearson's correlation coefficient

$$\rho = \frac{N\sum_{i=1}^{N} y_i \hat{y}^{-k(i)}(x_i) - \sum_{i=1}^{N} y_i \sum_{i=1}^{N} \hat{y}^{-k(i)}(x_i)}{\sqrt{N\sum_{i=1}^{N} {y_i}^2 - \left(\sum_{i=1}^{N} y_i\right)^2}\sqrt{N\sum_{i=1}^{N}\left[\hat{y}^{-k(i)}(x_i)\right]^2 - \left[\sum_{i=1}^{N} \hat{y}^{-k(i)}(x_i)\right]^2}} \tag{18}$$

The MSE metric gives an estimate of the expected test error by using the squared error as loss function. It ranges between zero and plus infinity, smaller values indicate smaller errors. The Pearson's correlation coefficient ranges between -1 and 1 and provides the ratio between the covariance of y and $\hat{y}$ and the product of their standard deviations. In other words, it measure the tendency of $\hat{y}$ to vary in function of y. If ρ is close to zero, y and $\hat{y}$ are weakly correlated and, hence, is expected that the prediction model $\hat{y}$ badly reproduces the variation of the function y. On the other hand, if ρ approaches the unity value, a strong correlation between the variables and the two datasets y_i and $\hat{y}^{-k(i)}(x_i)$ is obtained. If ρ is close to -1, anti-correlation exists and is expected that, for positive variation of y, negative variation of $\hat{y}$ is obtained.

3 AERODYNAMIC DATABASE USED FOR SURROGATE COMPARISON

This section describes how the database, to be used for comparison and validation purposes, was generated from the DPW-W1 wing baseline geometry [33], whose reference quantities for this wing are displayed in Table 1.

Reference quantity	Value
S_{ref} (wing reference area)	290,322 mm^2
C_{ref} (wing reference chord)	197.55 mm
X_{ref} (relative to the wing root leading edge)	154.24 mm
b/2 (semi span)	762 mm
AR (aspect ratio, AR=b^2/S_{ref})	8

Table 1: Reference quantities for the DPW wing

As can be seen in Fig. 1, in order to generate the database, the DPW geometry was parameterized by a 3D control box, with 5 control points in direction **u**, 10 in direction **v** and 5 in direction **w**. The parametric **u** direction corresponds to the y-axis, the **v** direction to the x-axis, and the **w** direction to the z-axis. The vertical displacement of those control

points, which correspond to the design space, set up on the aerodynamic surface. The wing is split in three profile sections and the transition between sections is linear. Each section has 6 active control points for the upper side and other 6 for the lower side, which are independent (the movement of a control point at the upper side does not modify the lower side and vice versa), with a total of NS=36 design variables for the whole wing.

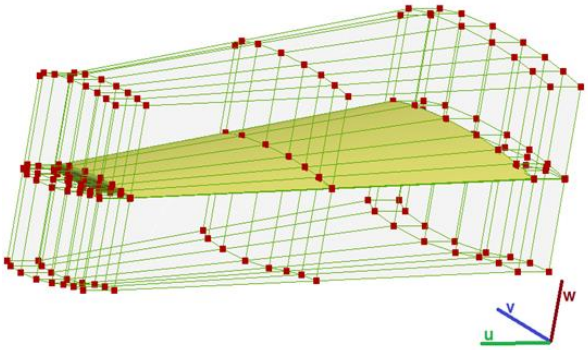


Figure 1: DPW wing parameterization

The initial database X was performed according to the methodology described in previous sections, executing 180 cases for position changes of the 36 design control points by means of DLR TAU solver in a High Performance Computer platform (HPC). Each design variable has been constrained by its minimum and maximum values that will be chosen as +/- 20% of their original value. In addition, other constraints, such as airfoil's maximum thickness and beam constraints have been defined, according to [34]. The flow conditions are Mach = 0.80, angle of attack 0 deg and Reynolds, $5*10^6$.

The design goal for the SBO is to achieve a geometry with the minimum drag, while maintaining the specified aerodynamic constraints:

1. Prescribed constant lift coefficient ($C_L = C_L^0$).
2. Minimum pitching moment: $C_M \geq C_M^0$.
3. Drag penalty: If constraint in minimum pitching moment is not satisfied, the penalty will be 1 drag count per 0.01 increment in C_M.

Therefore, an objective function was derived, based on the lift, drag and pitching moment aerodynamic coefficients obtained by the DLR TAU code, and the previously mentioned constraints [17]. This objective function will be modeled by the surrogate model in order to predict new values within the SBO scheme.

4 EXPERIMENTAL AND COMPARATIVE RESULTS

Surrogate modeling has been applied to the database X. As described in Section 3, this database has N=180 samples and NS=36 columns, with a +/-20% variation of the design variables from the base geometry. Thus, both KR and SVR methodologies have been

applied in the same conditions, validating them with 10-fold CV as described in Section 2.4.

Figure 2 shows the R^2 comparison between KR and SVR. In the KR case, R^2 is 0.8239, 62.7% greater than the value obtained with SVR (0.5064). This difference can be seen to a lesser extent in the Pearson coefficient stated in Table 2, with a 27.6% of variation. In addition, the MSE and the RMSE have a difference of 58.3% and 35.4% respectively.

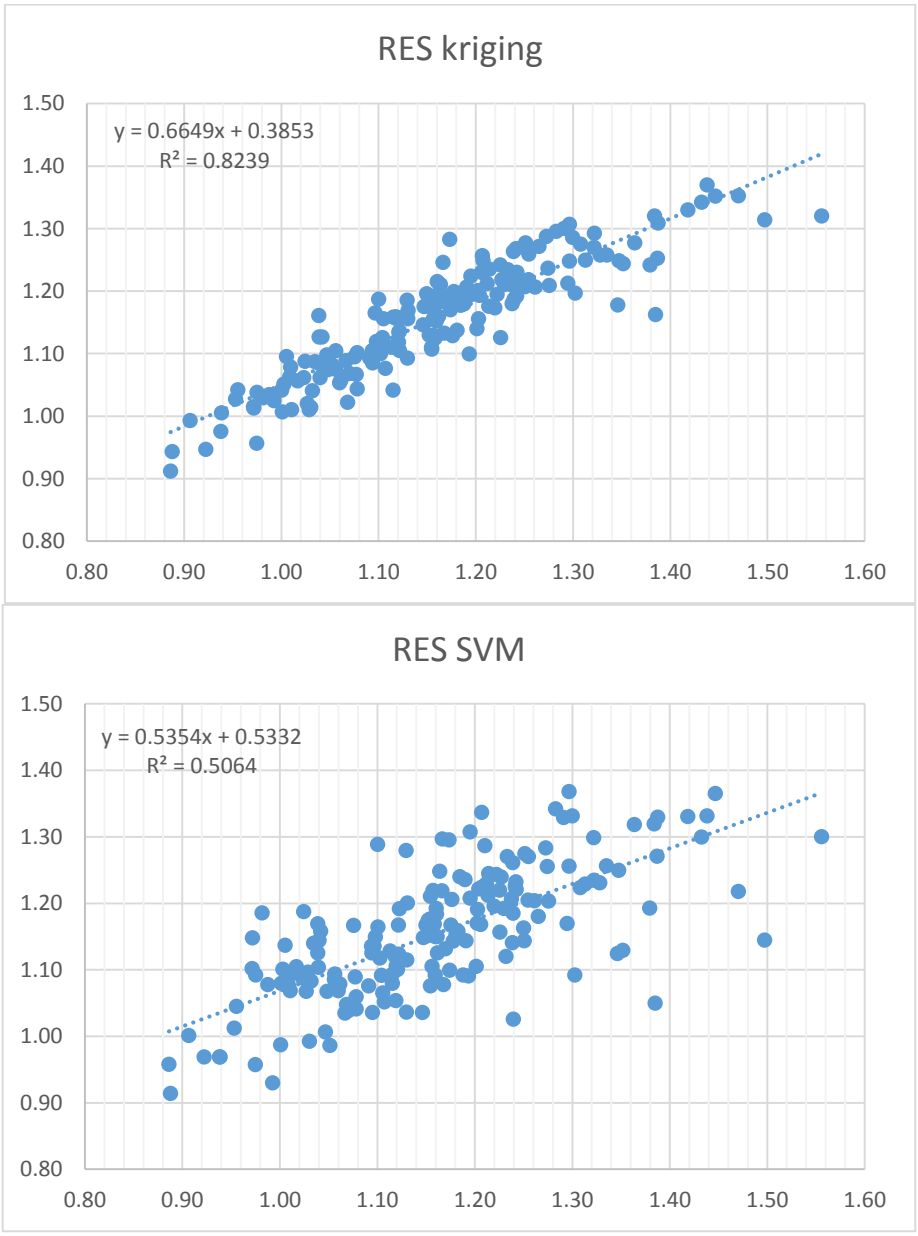


Figure 2: R^2 comparison between KR (above) and SVR (below).

Thus, KR seems to offer a better performance than SVR, applied to the DPW prediction of the objective function. This study allows to select the best surrogate model between SVR and KR based on experimental data, due to it is very difficult to know a priori which surrogate model is more suitable for a specific application.

	SVR	KRIGING	Difference
MSE	7.8093E-03	3.2580E-03	58.3%
RMSE	8.8370E-02	5.7079E-02	35.4%
Pearson	0.7116	0.9077	27.6%
R^2	0.5064	0.8239	62.7%

Table 2: Summary of performance comparison between KR and SVR.

5 CONCLUSIONS AND FUTURE WORKS

This paper describes the comparison between two different surrogate models, Kriging (KR) and Support Vector Regression Machines (SVR), for prediction of the objective function within different aerodynamic configurations, with the aim of aerodynamic optimization. In order to carry out this comparison, both KR and SVR have been applied to the same training database generated with CFD simulations for different geometries. V-fold cross validation has been used to compare both methods, showing the better performance obtained by KR methodology.

Future work will focus on studying the relationship between the model performance and the number of design variables and other parameters with influence, in order to minimize the required CFD simulations for the initial database, and therefore speed up the design process.

REFERENCES

[1] S. Sakata, F. Ashida and M. Zako, An efficient algorithm for Kriging approximation and optimization with large-scale sampling data, Computer methods in applied mechanics and engineering, 2003, pp. 386–404.

[2] W.C. Carpenter, J.F.M. Barthelemy, A Comparison of polynomial Approximations and Artificial Neural Nets as Response Surfaces, A collection of technical papers: the 33rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, AIAA, Dallas, TX, April 13–15, 1992, pp. 2474–2482.

[3] E. Nikolaidis, L. Long, Q. Ling, Neural networks and response surface polynomials for design of vehicle joints, AIAA-98-1777, 1998, pp. 653–662.

[4] N. Papila, W. Shyy, N. Fitz-Coy, R.T. Haftka, Assessment of neural net and polynomial-based techniques for aerodynamic applications, in: AIAA 17th Applied Aerodynamics Conference, AIAA-99-3167, 1999.

[5] R. Jin, W. Chen, T.W. Simpson, Comparative studies of metamodeling techniques under multiple modeling criteria, in: 8th AIAA/NASA/USAF/ISSMO Symposium on Multi-disciplinary Analysis and Optimization, Long Beach, CA, AIAA, 2000, AIAA-2000-4801.

[6] G.E.P. Box, K.B. Wilson. On the experimental attainment of optimum conditions, J. Roy. Statist. Soc. Ser. B 13,1951, 1–45.

[7] T. Kashiwamura, M. Shiratori, Q. Yu, Statistical optimization method, Comput. Aided Optimum Des. Struct. V,1997. 213–227.

[8] I.P. Sobieski, V.M. Manning, I.M. Kroo, Response surface estimation and refinement in collaborative optimization, AIAA-98-4753, 1998, pp. 4753–4770.

[9] S. Hosder, L.T. Watson, B. Grossman, W.H. Mason, H. Kim, R.T. Haftka, S.E. Cox, Polynomial response surface approximations for the multidisciplinary design optimization of a high speed civil transport, Technical Reports of NCSTRL at Virginia Tech CS (TR-01-03), 2001.

[10] S. Thamarai-Selvi, E. Mahendran and S. Rama. Neural Network-Based Interpolation and Extrapolation of Wind Tunnel Test Data Int. J. Comp. Intel. Appl. 08, 225 (2009. DOI: 10.1142/S1469026809002564.

[11] L. Berke, P. Hajela, Application of artificial neural nets in structural mechanics, Struct. Optim. 4,1992, pp. 90–98.

[12] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, Statist. Sci. 4 (4),1989, pp. 409–435.

[13] A. Giunta, L.T. Watson, J. Koehler, A Comparison of approximation modeling techniques: polynomial versus interpolating models, in: 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis & Optimization, St. Louis, MI, September 2–4, AIAA-98-4758-CP, 1998.

[14] D.J. Cappelleri, M.I. Frecker, T.W. Simpson, Design of a PZT bimorph actuator using a metamodel-based approach, J. Mech. Des. Trans. ASME 124,2002, 354–357.

[15] S. Sakata, F. Ashida, M. Zako, Optimization of laminated composites for eigenfrequency problem using K-CSSL method, in: F.- K. Chang (Ed.), Proceedings of the Tenth US–Japan Conference on Composite Materials, DEStech Publishers, 2002, pp. 560–569.

[16] C. Praveen., R. Duvigneau. Radial Basis Functions and Kriging Metamodels for Aerodynamic Optimization. INRIA Technical report 2007.

[17] E. Andres, S. Salcedo-Sanz, F. Monge, A.M. Perez-Bellido. Efficient aerodynamic design through evolutionary programming and support vector regression algorithms. Expert Systems with Applications 39,2012. pp. 10700–10708.

[18] S. M. Clarke, J. H. Griebsch, and T. W. Simpson. Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses. Transactions of ASME, Journal of Mechanical Design, 127(6), 2005, pp. 1077-1087.

[19] T.W. Simpson, T.M. Mauery, J.J. Korte, F. Mistree, Comparison of Response Surface and Kriging Models for Multidisciplinary Design Optimization, 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis & Optimization, St. Louis, MI, September 2–4, AIAA-98-4755, 1998.

[20] S. Sakata, F. Ashida, M. Zako, Structural optimization using Kriging approximation, Comput. Methods Appl. Mech. Engrg. 192 (7–8), 2003, pp. 923–939.

[21] J.P.C. Kleijnen. Kriging metamodeling in simulation: A review. European Journal of Operational Research, 192(3), 2009, pp. 707–716.

[22] J. Sacks, S. B. Schiller, and W. J. Welch. Designs for computer experiments. Technometrics, 31(1):41–47, February 1989.

[23] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. Statistical Science, 4(4):409–423, November 1989.

[24] W. Shao, H. Deng, Y. Ma, and Z. Wei. Extended Gaussian kriging for computer experiments in engineering design. Engineering with Computers, 28(2):161–178, 2012.

[25] E. Içten, Z.K. Nagy, G.V. Reklaitis, Process control of a dropwise additive manufacturing system for pharmaceuticals using polynomial chaos expansion based surrogate model, *Computers & Chemical Engineering*, 83, 221-231,2015.

[26] J. Steuben, J. Micholopoulos, A. Iliopoilos, C. Turner, Inverse characterization of composite materials via surrogate modeling, *Composite Structures*, 132, 694-708, 2015.

[27] X. Wu, Y. Zheng, B. Wu, Y. Tian, F. Han, C. Zheng, Optimizing conjunctive use of surface water and groundwater for irrigation to address human-nature water conflicts: A surrogate modeling approach. *Agricultural Water Management*, 163, 380-392, 2016 (in press).

[28] J.C. Pierson, A. Mortelliti, P.S. Barton, P.W. Lane, D.B. Lindenmayer, Evaluating the effectiveness of overstory cover as a surrogate for bird community diversity and population trends, *Ecological Indicators*, 2015.

[29] M.K. Song, S.E. Ward, J.P. Fine, L.C. Hanson, F.C. Lin, G.A. Hladik, J.B. Hamilton, J.C. Bridgman, Advance Care Planning and End-of-Life Decision Making in Dialysis: A Randomized Controlled Trial Targeting Patients and Their Surrogates, 66 (5), 813-822, 2015.

[30] K. Bollaerts, M. Snock, K. Simons, S. Fierens, A. Poffijn, L.Van Bladel, D. Geraets, P. Gosselin, H. Van Oyen, J. Francart, A. Van Nieuwenhuyse, Thyroid cancer incidence around the Belgian nuclear sites: Surrogate exposure modelling, *Cancer Epidemiology*, 130, 12-26, 2014.

[31] A.M. Osiceanu, L.E. Murao, D. Kollanur, J. Swinnen, A.R. De Vleeschuwer, D.J. Lefebvre, K. De Clerq, J. Neyts, N. Goris, In vitro surrogate models to aid in the development of antivirals for the containment of foot-and-mouth disease outbreaks, *Antiviral Research*, 105, 59-63, 2014.

[32] Y. Yin, P. Shang, Modified cross sample entropy and surrogate data analysis method for financial time series, *Physica A: Statistical Mechanics and its Applications*, 433, 17-25, 2015.

[33] "AIAA 3rd Drag Prediction Workshop. URL: http://aaac.larc.nasa.gov/tsab/cfdlarc/aiaa-dpw/Workshop3/workshop3.html," [Online].

[34] B. Epstein, A. Jameson, S. Peigin, D. Roman, N. Harrison and J. Vassberg, "Comparative Study of 3D Wing Drag Minimization by Different Optimization Techniques.," in 46th AIAA Aerospace Sciences Meeting and Exhibit. AIAA paper 2, Reno, Nevada, 2008.

[35] GARTEUR AD/AG-52 Surrogate Based Global Optimization methods in preliminary Aerodynamic Design. Garteur group, 2015.

[36] Y. Bengio, and Y. Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. J. Mach. Learn. Res., 5, 2004, pp.1089–1105

[37] P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika, 76(3), 1989, pp.503–514.

[38] I. Couckuyt, T. Dhaene, and P. Demeester. ooDACE toolbox: A flexible object-oriented kriging implementation. Journal of Machine Learning Research, 15:3183–3186, 2014.

[39] S.M. Stigler. Statistics on the table: the history of statistical concepts and methods. Harvard University Press, 1999. ISBN 0-674-83601-4.

[40] C. Le, L. Lucien and G.Yang. Asymptotics in statistics: some basic concepts (Second ed.). Springer, 2000. ISBN 0-387-95036-2.

[41] E. Ortiz-García, S. Salcedo Sanz, Á. Pérez-Bellido and J. Portilla-Figueras, "Improving the training time of support vector regression algorithms through novel hyper-parameters searchspace reductions," Neurocomputing, vol. 72, p. 3683–3691, 2009.

[42] DLR Institute of Aerodynamics and Flow Technology. “Technical Documentation of the DLR TAU-Code”. Braunschweig, Germany, 2004.

[43] T.W. Simpson, D. K. J. Lin, and W. Chen, “Sampling Strategies for Computer Experiments: Design and Analysis,” International Journal of Reliability and Applications, Vol. 2, No. 3, 2001, pp. 209–240.

[44] T. W. Simpson, A. J. Booker, D. Ghosh, , A. A. Giunta, , P. N. Koch, , and R. J. Yang, “Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion,” Structural and Multidisciplinary Optimization, Vol. 27, No. 5, July 2004, pp. 302–313. doi:10.1007/s00158-004-0389-9.

[45] M. D. McKay, R. J. Beckman, and W. J. Conover, “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code,” Technometrics, Vol. 21, No. 2, May 1979, pp. 239–245. doi:10.2307/1268522.

[46] S. Geisser. The predictive sample reuse method with applications. J. Amer. Statist. Assoc., 70, 1975. pp. 320–328.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

HEAT TRANSFER OPTIMIZATION OF A RIBBED SURFACE USING SURROGATE-ASSISTED GENETIC ALGORITHMS

Panagiotis Tsirikoglou, Ghader Ghorbaniasl, Simon Abraham, and Chris Lacor

Vrije Universiteit Brussel (VUB), Department of Mechanical Engineering Fluid Mechanics and Thermodynamics Research Group
Pleinlaan 2, 1050, Brussels, Belgium
e-mail: Panagiotis.Tsirikoglou@vub.ac.be, Ghader.Ghorbaniasl@vub.ac.be, Simon.Abraham@ulb.ac.be, Chris.Lacor@vub.ac.be

Keywords: surrogate-assisted optimization, kriging, co-kriging, support vector regression, expected hypervolume improvement

Abstract. *Over the last years, complex design and optimization engineering problems became more and more demanding in computational time. Despite the advances made in computer science, these demands are prohibitive in some cases, in many engineering fields such as fluid mechanics and heat transfer. Aiming to alleviate this, surrogate response models are introduced and coupled with optimization drivers to deliver cheaper and accurate optimization results. The present paper investigates the performance of various surrogate-assisted optimization schemes applied to a heat transfer modeling problem of a ribbed surface. In this study, performance of different surrogate models such as Kriging, Co-Kriging and Support Vector Regression in different evolutionary optimization schemes are assessed. These schemes employ the aforementioned surrogate models coupled with different infill strategies depending on the availability of an uncertainty measure for the prediction. The results show that Co-Kriging model provides accurate results in comparison with the other metamodels while the computational time is reduced by more than 50%. It is illustrated that the combination of multi-fidelity approaches and sophisticated infill strategies can provide accurate predictions at a reduced computational cost.*

1 INTRODUCTION

Design and optimization problems in various engineering fields, are becoming more complex. The growing need of taking into account the highest number of design variables possible, stretches the capabilities of the existing computational systems. In this context, the Curse of Dimensionality is an important issue that has to be addressed. Use of surrogate evaluation models, is a well-known practice to alleviate the computational time of expensive response models. Surrogates can mimic the underlying behaviour of the response function providing a cheap and alternative evaluation tool. The prediction quality of these models relies on their effective training process and the proper selection of the respective training samples. The coupling of those models with optimization schemes, called Surrogate Based Optimization (SBO), results in more computationally efficient strategies.

The seminal study of Jones et al. [1] used Kriging models as a predictor during the optimization cycles. The validation of the surrogate model is crucial during the optimization process to assess the results derived and it is thoroughly discussed at [1]. Emmerich et al. [2] described an SBO scheme for both single and multi-objective optimization problems. The proposed scheme in that study employs Gaussian Random Field Meta-models (GRFM) and a novel procedure of selecting the infill training points from the offspring generated. It should be noted here that this infill strategy requires an uncertainty measure of the surrogate model's prediction. Particularly for the multi-objective problems, the Expected HyperVolume Improvement (EHVI) suggested, uses balanced exploration and exploitation of the metamodel in order to screen out the less promising individuals during the generational process. Thus, the proposed scheme results in a more computationally efficient optimization strategy. Building more on the EHVI concept, Hupkens et al., in studies [3, 4], suggest a faster computational procedure for the calculation of the EHVI scores which replaces the initial Monte-Carlo integration technique. Empirical comparisons showed that this new implementation results in minimum five times faster computations for two or three dimensional objective spaces. It is worth to mention here that the Couckuyt et al. [5] algorithm remains the fastest one for the computation of hypervolume-based Probability of Improvement in higher dimensional objective spaces (more than three dimensions).

The use of SBO schemes for heat transfer modeling problems is of great importance, since the application range is very wide and the numerical response models can be relatively costly. Particularly, the heat transfer optimization of ribbed surfaces is important for the cooling performance of turbine blades. The wide application field of turbomachinery justifies the rigorous research activity, aiming to define the optimal geometrical and operational configuration of these ribbed surfaces. Labbé [6] described the flow topology and heat transfer performance of a ribbed surface with constant-sized ribs using Large Eddy Simulation. Recirculation zones on the top and behind the rib were identified while the enhancement of heat transfer performance is maximum in front of the ribs, as a result of the highly unsteady secondary eddies. The numerical optimization of a ribbed channel was performed in Yang and Chen study [7], where Response Surface Method (RSM) and evolutionary-based optimizer were coupled in order to derive the optimized scenarios. The design variables used in the study were the height and thickness of the ribs and the pitch of the channel while the formulated objective was based on the Nusselt number distribution and the pressure drop. Results showed an apparent increase of the channel's performance particularly for the staggered ribs design scenario. Xie et al. [8] investigated numerically five different configurations of variable sized ribs in a three-dimensional internal cooling channel flow. The different scenarios employ half-sized and same-sized ribs

aiming to define the configuration that results in the maximum Nusselt number distribution and the minimum pressure drop. Despite the fact that no optimizer is used in this study, the numerical investigation provides useful insights for the ribbed surface set up and a detail description of the heat transfer modeling case.

In the present study, a nominal geometrical configuration based on [8] is used to build a SBO test case and investigate its performance using CFD evaluations and different surrogate models. Thus the purpose of this research study is twofold: investigate and assess the performance of the SBO in heat transfer problems and derive further optimized configurations of the ribbed surface under examination. Therefore, the suggested optimization framework will be first described and validated against a well-known analytical test case. Then, the numerical optimization results regarding the heat transfer test case will be presented and thoroughly discussed.

2 A SURROGATE-BASED OPTIMIZATION FRAMEWORK

The implementation of the optimization test cases is based on the development of a modular SBO framework. Each of the implemented modules performs a specific group of tasks. The modular approach enhances the flexibility while it makes easier the addition of new features. The necessary modules are enlisted in Table 1 along with a short description.

Modules	Short Description
Base module:	Co-ordinates the actions of different modules
Parameters module:	Defines the input parameters for the selected test case
Evaluation module:	Implements the proper fitness function for the selected test case
Sampling module:	Performs the initial sampling plan
Optimize module:	Implements the optimization schemes. Contains the optimizer as well
Infill module:	Implements different infill strategies
Surrogate module:	Implements the training, activation and assessment actions of different surrogate models

Table 1: Short description of the SBO framework modules.

The developed framework can function in two basic modes: the offline and the online metamodel training modes. The simpler offline training strategy uses an initial Design of Experiment in order to train the surrogate model and uses it instead of the expensive fitness function during the optimization cycles. The method does not use infill points to update the metamodel. Thus, the quality of the predictions and the final optimization outcome is based on the sampling technique and the initial amount of samples. This primitive strategy is simple to implement and suitable for very low dimensional problems. The more sophisticated and currently used online mode is starting with a Design of Experiment (DoE) which serves as the population of the first generation as well. These design samples are calculated then, employing the exact response. The resulting data is used as input training data for the respective surrogate model. The trained meta-model is activated during the optimization cycles, aiming to calculate the fitness values of the offspring created. An infill strategy, such as EHVI [2], is employed to assess the offspring and to choose the best individual which will be evaluated through the expensive response. Afterwards this individual is added to the training database and the metamodel is updated. In this way the model evolves during the optimization cycles which results into better prediction at the

regions of interest (near the optimum). This strategy is based on a smaller initial DoE and it is depending strongly on the performance of the infill strategy.

The described task-flow of the framework is based on the implementation of specific techniques for each modules. Starting from the sampling strategy, the Optimal Latin Hypercube Design [9] was used due to its improved space-filling properties. A simple genetic algorithm is used to maximize the Morris-Mitchell criterion and derive the optimized sampling plan. The optimizer employed in all the test case is the extensively used Nondominated Sorting Genetic Algorithm Type-II (NSGA-II) [10], which is suitable for multi-objective problems. The respective algorithm was created with the Distributed Evolutionary Algorithms in Python (DEAP) library [11]. The infill strategy used here is the classical Expected Improvement (EI) in case of a single-objective problem and the described EHVI [2] method for the bi-objective problems. A hypervolume-based indicator is used as well in order to be coupled with specific metamodels. The combinations of the surrogates and the infill strategies are explained in the next paragraph. The surrogate model part contains models such as Kriging [12] and its variant Co-Kriging [13] and Support Vector Regression [14]. The implementation of the Co-Kriging model was based on the work of Le Gratiet et al. [15] as it was implemented originally at the Open Multidisciplinary Design Analysis and Optimization (OpenMDAO) [16] framework in Python. The use of Co-Kriging requires the definition of variable levels of fidelity. The model is based on the correlation of those levels which eventually bring advantages in computational efficiency.

Finally it should be mentioned that the different surrogate models are capable to work with specific infill strategies. This is resulting in two different optimization schemes for the online mode. The Kriging and Co-Kriging models are combined with EHVI [2-4] infill criterion which accounts balanced exploration and exploitation of the surrogate models. The SVR metamodel is combined with the hypervolume indicator which in the case of a bi-objective problem translates to the area of the produced Pareto front and purely exploits the surrogate model. The reason behind those choices is that, unlike the SVR models, the Gaussian metamodels can easily provide an uncertainty measure of their prediction. Thus the balance between exploration and exploitation of the surrogate is allowed.

3 TEST CASES DESCRIPTION

3.1 Analytical Test Case

The implementation of different techniques and its coupling is not a trivial task. In order to ensure the proper functioning of those techniques, a validation test case should be implemented. Therefore, the DTLZ2 [17] bi-objective analytical test case is chosen to validate and assess the performance of the multi-objective SBO framework. The mathematical formulation of the test case is described as follows:

$$f_1(\vec{x}) = (1 + g(\vec{x})) \cos x_1 \frac{\pi}{2} \tag{1}$$
$$f_2(\vec{x}) = (1 + g(\vec{x})) \sin x_1 \frac{\pi}{2}$$

where:

$$g(\vec{x}) = \sum_{x_i \in \vec{x}} (x_i - 0.5)^2 \tag{2}$$
$$0 \leq x_i \leq 1, i = 1, 2, \ldots, N_{dimensions}$$

3.2 Heat Transfer Test Case

3.2.1 Design Variables and Computational Domain

The heat transfer test case aims to improve the heat transfer modeling of a surface with variable-size ribs placed in an internal cooling passage. The detailed description of this test case is based on [8], where five different geometric configurations are investigated. The configuration selected to be further optimized is demonstrated in Fig. 1(one-pitch length).

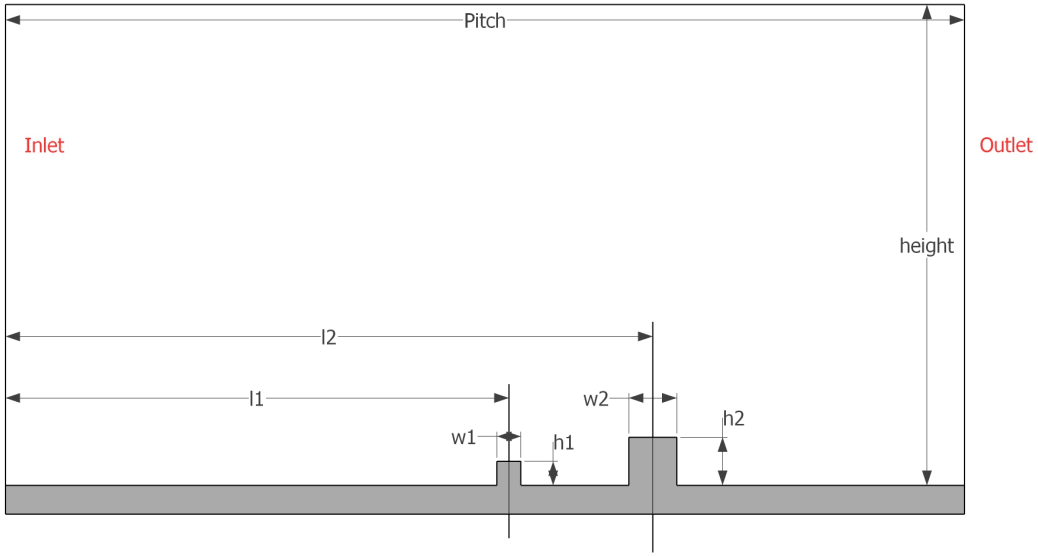


Figure 1: Nominal geometric configuration.

As shown in Fig. 1, the design variables of the test case are the height, width and placement length of the ribs resulting in a six-dimensional design space. The height and the pitch of the ribbed channel are 50 mm and 100 mm in respect, with a hydraulic diameter of 50 mm as well. The computational domain is restricted in one-pitch length in order to decrease the mesh size and accelerate the numerical evaluations. The proper definition of the periodic boundary conditions based on [8] ensures the quality of the CFD results.

3.2.2 Overview of the Numerical Procedure

The CFD evaluations aim to calculate the Nusselt number distribution on the ribbed surface and the pressure drop between the inlet and the outlet of the channel. A two-dimensional, steady and turbulent flow process is assumed to calculate these values. The standard k-ϵ turbulence model with standard wall-function is selected. It should be added that the selected model is not the most suitable for the heat transfer problems. The present study though focuses more on the performance of the SBO scheme. Therefore, the balance between the quality of the results and the simplicity of the model is deemed proper for these purposes. The pressure-velocity coupling is achieved through SIMPLE method, while second order schemes are used for the spatial discretization. Finally the absolute convergence criterion is set at 10^{-6}.

3.2.3 Boundary Conditions

As it was stated previously periodic boundary conditions are used for this problem. The periodicity is defined at the inlet and the outlet of the channel. The mass flow rate is used as a parameter. In order to mimic the heat transfer in an internal cooling passage, constant heat flux of 1000 W/m^2 is applied to the entire walls of the channel, including the ribs. The air enters the passage at a temperature of 300K and a Reynolds number of 20000. The turbulence intensity level is defined as 5% at the inlet.

3.2.4 Definition and Validation of Different Fidelities

As it was mentioned in section 2, the multi-fidelity approach is employed in order to use the Co-Kriging model and benefit from its advantages. Therefore, it is important to describe two different levels of fidelity which serve as a cheap and expensive response. The first cell distance is defined as the basic criterion to derive the alternative levels of fidelity, since the heat transfer phenomena are intense at the boundary layers. Table 2 summarizes the main aspects of the two defined fidelities.

	Lower Fidelity	Higher Fidelity
First Cell Distance [m]	1e-04	1e-05
Mesh Size	15500	32000
Computational Time* [min]	7	30

**The computational time was defined on an 4-cores Intel Xeon E3-1200*

Table 2: Main characteristics of the various fidelities.

In addition, the distributions of Nusselt number on the ribbed surface and the pressure drop along the vertical axis are presented in Fig. 2. It is apparent that the lower fidelity simulations overpredict Nusselt number distribution while the pressure drop is underpredicted.

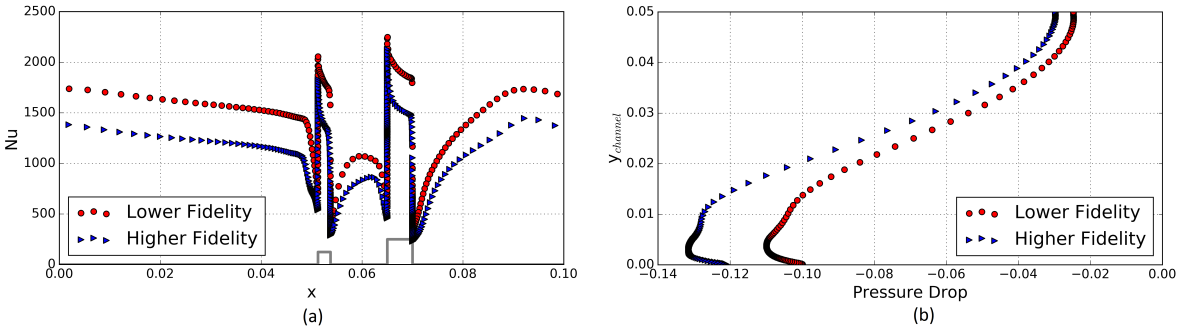


Figure 2: Nusselt number (left) and pressure drop distribution (right) for different fidelities.

In the extent of this definition, the numerical approaches used are validated against the velocity profiles obtained from [8]. An indicative comparison plot is shown in Fig. 3. The overall agreement is good. The underprediction of the velocities is due to the fact that the present study employs a two-dimensional simulation while [8] uses a three-dimensional approach.

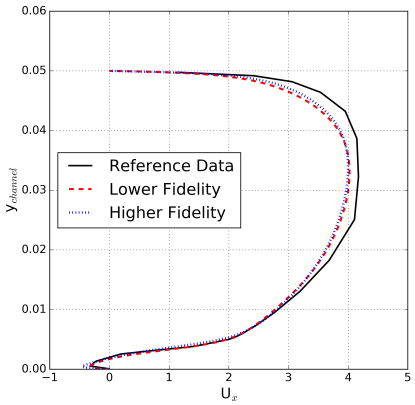


Figure 3: Comparison of derived velocity profiles to the reference CFD data.

4 RESULTS AND DISCUSSION

4.1 Analytical Test Case Results

The DTLZ2, bi-objective analytical test case is used mainly to test and validate the function of the developed framework. The validation is performed using a six-dimensional version of the DTLZ2 problem (same size of the design space as the heat transfer test case), while one surrogate model per objective is used. It is reminded that the results presented are obtained using the combinations of Kriging metamodel with the EHVI strategy and SVR surrogate model with the hypervolume-based indicator. The Co-Kriging model is not used in this test case since the definition of variable levels of fidelity does not lead to any actual computational time differences. The model will be investigated in the heat transfer test case where the different levels of fidelity are properly defined.

A reference Pareto front of the optimized DTLZ2 test case without surrogates employed is obtained. Fig. 4 compares the Pareto fronts obtained by the described surrogate models against the reference one. The results presented are obtained after 20 generations. The Kriging

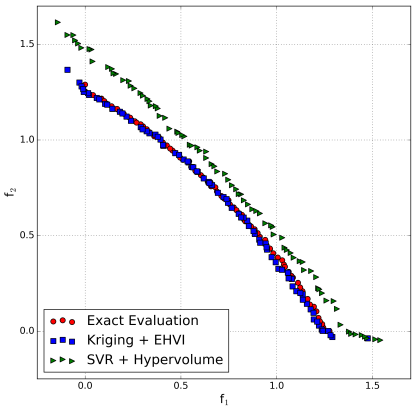


Figure 4: Optimal Pareto fronts obtained with the different optimization schemes.

surrogate model coupled with the EHVI technique managed to derive an optimal Pareto front similar to the reference one by calculating 95% less samples. The SVR model along with the hypervolume-based selection of infill points results in a Pareto front clearly dominated by the reference Pareto. The reason behind this difference is the infill strategy used. The hypervolume-based selection only exploits the surrogate model. Thus, the improvement of the surrogates performance is slower. It is expected that the higher number of samples in the training database will finally increase the prediction quality. Indeed, the Pareto front obtained for SVR models trained with higher number of samples is closer to the reference Pareto as it is presented in Fig. 5. It should be mentioned that in this case the SVR training database is 36% bigger than the Kriging training database.

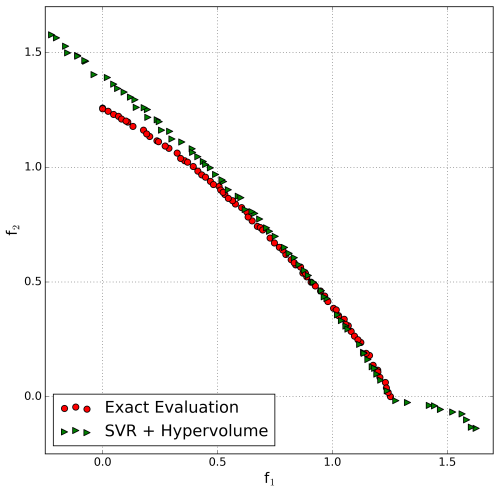


Figure 5: Reference and SVR obtained Pareto fronts.

The implementation of the analytical test case with the developed framework provides a fast way to assess different optimization schemes. Therefore, the comparison between the offline and online training schemes of the surrogate models is described within this test case. The NSGA-II optimizer with a constant number of twenty generations was employed, while the Kriging model is selected for the surrogate evaluations. The online training mode allows the surrogate to start from a small training database and enrich it during the optimization cycles by selecting the most appropriate infill points. On the contrary, during the offline training mode the metamodel is trained a priori, with a constant higher number of samples and then is used as the fitness function. This fact results to the calculation of unnecessary samples which decreases the overall computational efficiency. The superior performance of the online training mode is illustrated at Fig. 6.

The online training mode outperforms the offline one in terms of accuracy for almost the same number of samples. Further increase of the size of the training database, which is translated to more computational time, will improve the performance of the surrogate trained with the offline mode. It is proved that the proper selection of the new points, added in the training database is essential for the performance of the surrogate as a fitness function and for the reduction of the final computational cost. The pre-definition of the training database using solely the initial DoE results in less computationally efficient SBO schemes. The efficiency decreases

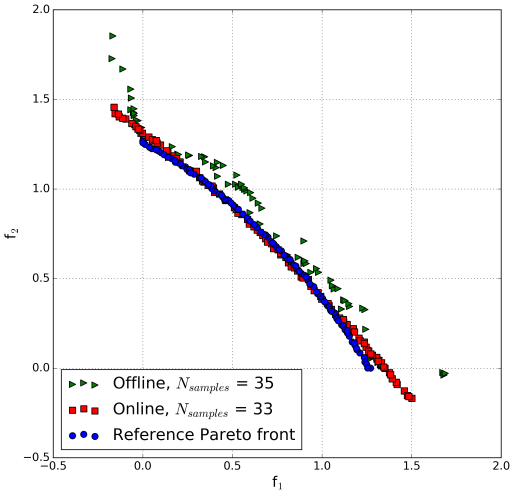


Figure 6: Comparison of online and offline surrogate training modes for the DTLZ2 test case.

more intensively for higher dimensional test cases. Moreover, the prediction accuracy in offline mode depends more strongly on the morphology of the response. Thus, the use of the online training mode is clearly recommended.

4.2 Heat Transfer Test Case Results

The goal of this test case is the heat transfer modeling of a ribbed surface placed in an internal cooling channel. First a reference solution based only on CFD evaluations only is obtained. This solution is considered as the baseline to determine the most efficient scheme among the different SBO approaches. One surrogate model per objective approach is employed in this test case as well. The online training mode of the surrogate is used due to its profound advantages described in the previous section. The definition of various levels of fidelity for this test case (section 3.2.3) allows the meaningful assessment of a multi-fidelity surrogate such as Co-Kriging.

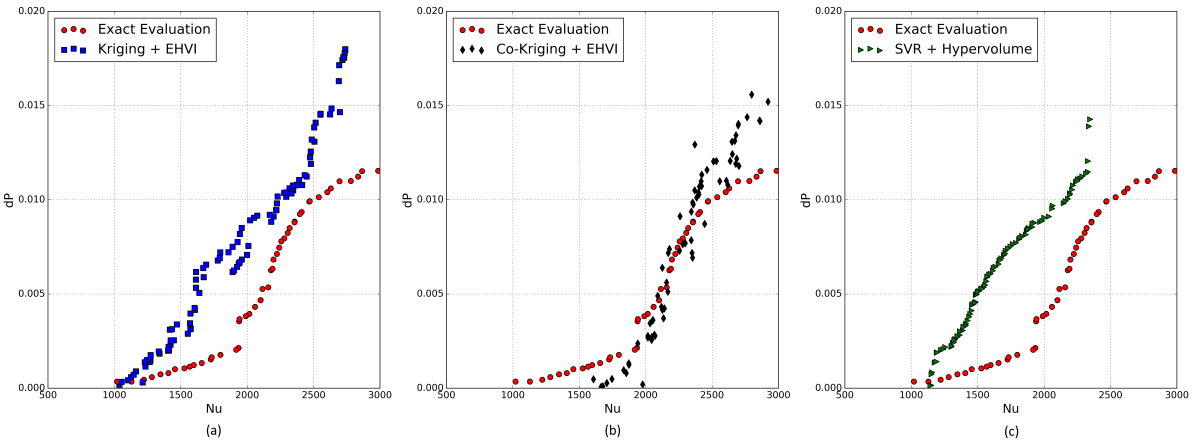


Figure 7: Optimal Pareto fronts obtained with the various surrogate models: a. Kriging, b. Co-Kriging, c. SVR.

In all the cases, optimal Pareto fronts are well formed and tend to follow the same trends as the reference Pareto. The Co-Kriging metamodel manages to derive the optimal Pareto front closer to the reference one. Particularly, the lower part of this Pareto contains individuals which dominate the respective individuals at the reference front. Those configurations indicate further improved designs of the ribbed surface. Among Kriging and SVR surrogate models, the latter present the less dominant Pareto front due to the use of the less efficient hypervolume-based infill strategy. This is justified from the selection process of the infill points as it was described for the analytical test case. The points defined from the DoE and the different infill strategies are presented for each surrogate in Fig. 8.

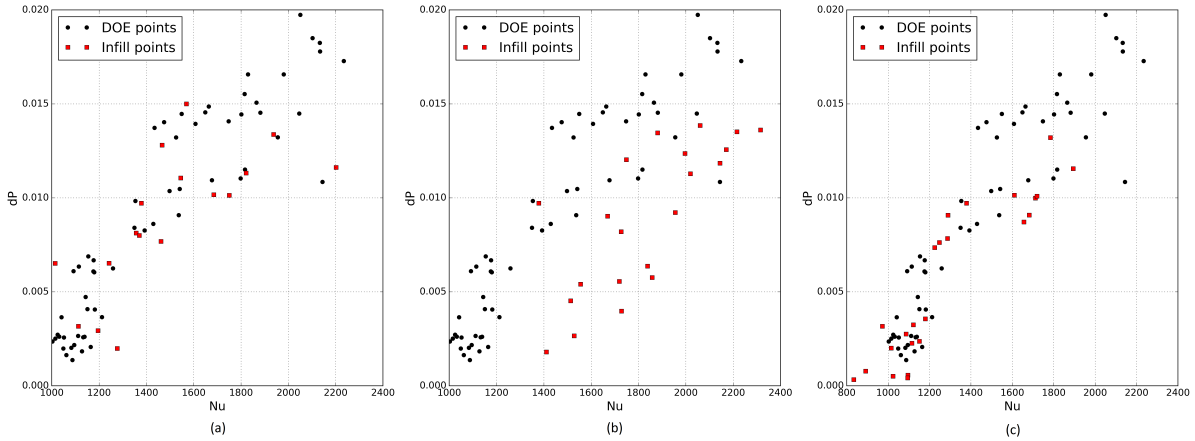


Figure 8: DoE and Infill design sites: a. Kriging. b. Co-Kriging, c. SVR.

The Co-Kriging metamodel coupled with the EHVI technique manages to capture far more infill points placed closer to the reference Pareto (towards bottom-right of Fig. 8b) front than the other two metamodels. Thus, the surrogate model improved its prediction capabilities at the region of interest, at a reduced computational cost due to the multi-fidelity approach. As it was expected the SVR model capture the least points at the region of interest. It is clear that more generations are needed in order to obtain a dominant Pareto front. The additional calculation time though will lead to a further decrease of the computational efficiency.

Finally, the prediction accuracy of the surrogate models are tested. For each optimal Pareto front obtained, 30 points are randomly selected to be evaluated through CFD simulations. Then, the coefficient of determination (R^2) is calculated to determine the quality of the prediction:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3}$$

Table 3 summarizes the prediction performance along with the number of samples in the training database and the respective computational time.

The computational cost of Co-Kriging surrogate is reduced by 57% and 61% compared to the Kriging and SVR model respectively. The huge decrease of computational time is due to the multi-fidelity approach. Among the approximately one hundred samples only twelve of them are calculated using the higher fidelity. The computational efficiency of the Co-Kriging is dramatically increased since the lower fidelity simulations require 75% less time. Thus, the Co-Kriging model coupled with the EHVI method capture the most efficient infill points

	Number of Samples	Computational Time [min]	R^2_{Nu}	R^2_{dP}
Kriging	76	2280	0.8775	0.84915
Co-Kriging	101	983	0.8925	0.8603
SVR	83	2490	0.8205	0.9403

Table 3: Performance metrics of the surrogate models.

by consuming the least time. Moreover the performance of the Kriging and SVR models in terms of computational time is quite close since the size of the final training database is similar. The quality of the infill points though from the optimization standpoint, is reduced for the SVR model. Finally, the generalization error for all the surrogate models is sufficiently good. Further increase of the training database size will improve the quality of the predictions.

5 CONCLUSIONS

The present study described the development of a Surrogate-Based Optimization framework capable of implementing different optimization strategies. The framework then was used to optimize the heat transfer modeling of a ribbed surface of an internal cooling channel.

The proper functioning of the model was validated through a benchmark multi-objective analytical test case. The superior performance of the online surrogate training mode against the corresponding offline mode was proved as well. The results for both test cases showed that the algorithm is able of approximating the real optimal Pareto front particularly for the case of the Gaussian models coupled with the Expected HyperVolume Improvement infill strategy. Furthermore, among the Gaussian models, Co-Kriging showed sufficient good prediction capabilities while the computational time needed was 57% less than the Kriging surrogate model. Thus, the multi-fidelity surrogate models coupled with a sophisticated infill strategy, which allows balanced exploration and exploitation of the metamodel, were considered as the best option in the present study.

Further investigation of the Support Vector Regression (SVR) models regarding the uncertainty measures of their predictions will improve their performance within an optimization framework. This is suggested as a subject of future research since the performance of the SVR models within this study and their capabilities in general are quite promising.

REFERENCES

[1] D. R. Jones, M. Schonlau, W. Welch, Efficient Global Optimization of Black Box Functions. *Journal of Global Optimization*, **13**(4): 455-492, 1998.

[2] M. T. M. Emmerich, K. C. Gianakoglou, B. Naujoks, Single and Multiobjective Evolutionary Optimization Assisted by Gaussian Random Field Metamodels. *IEEE Transactions on Evolutionary Computation*, **10**(4): 421 439, 2006.

[3] I. Hupkens, M. Emmerich, A. Deutz, Faster Computation of Expected Hypervolume Improvement. *arxiv:1408.7114v1[cs.Ds]*,2014.

[4] I. Hupkens, A. Deutz, K. Yang, M. Emmerich, Faster Exact algorithms for Computing Expected Hypervolume Improvement. *In Proc. Part II EMO Conf.*,Guimares, March, 2015.

[5] I. Couckuyt, D. Deschrijver, T. Dhaene, Fast Calculation of Multio-bjective Probability of Improvement and Expected Improvement Criteria for Pareto Optimization. *Journal of Global Optimization*, **60**(3): 575-594, 2014.

[6] O. Labb, Large-eddy-simulation of flow and heat transfer in a ribbed duct. *Computers and Fluids*, **76**, 23-32, 2013.

[7] Y. T. Yang, P. J. Chen, Numerical Optimization of Turbulent Flow and Heat Transfer Characteristics in a Ribbed Channel. *Heat Transfer Engineering*, **36**(3): 290-302, 2015.

[8] G. Xie, S. Zheng, W. Zhang, B. Sundn, A Numerical Study of Flow Structure and Heat Transfer in a Square Channel with Ribs Combined Downstream Half-Size or Same-size Ribs. *Applied Thermal Engineering*, **64**(2): 289-300, 2013.

[9] M. D. Morris, T. J. Mitchell, Exploratory Designs for Computer Experiments. *Journal of Statistical Planning and Inference*, **43**(3): 381-402, 1995.

[10] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, **6**(2): 182-197, 2002.

[11] F. A. Fortin, F. M. De Rainville, M. A. Gardner, M. Parizeau, C. Gagn, DEAP: Evolutionary Algorithms made Easy. *Journal of Machine Learning Research*, **13**, 2171-2175, 2012.

[12] D. G. Krige, A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Engineering Society of South Africa*, **52**(6): 119-139, 1951.

[13] A. I. J. Forrester, A. Sobester, A. J. Keane, Multi-fidelity Optimization via Surrogate Modeling. *Proceeding of the Royal Society A*, **463**(2088): 32513269, 2007.

[14] A. J. Smola, B. Schlkopf, A Tutorial on Support Vector Regression. *Statistics and Computing*, **14**(3): 199-222, 2004.

[15] L. Le Gratiet, J. Garnier, Recursive Co-Kriging Model for Design of Computer Experiments with Multiple Levels of Fidelity. *International Journal for Uncertainty Quantification*, **4**(5): 365-386, 2014.

[16] J. Gray, K. T. Moore, B. A. Naylor, OPENMDAO: An Open Source Framework for Multidisciplinary Analysis and Optimization. *In Proc. AIAA/ISSMO Conf.*, Atlanta, June, 2014.

[17] K. Deb, L. Thiele, M. Laumanns, E. Zitzler, Scalable Multi-objective Optimization Test Problems. *In Proc. CEC 02 Conf.*, Honolulu, May, 2002.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MULTI-FIDELITY EXTENSION TO NON-INTRUSIVE PROPER ORTHOGONAL DECOMPOSITION BASED SURROGATES

Benamara T.[*,1,2,3], **Breitkopf P.**[1], **Lepot I.**[2], **and Sainvitu C.**[2]

[1]Laboratoire Roberval UMR 7337 UTC-CNRS
Université de Technologie de Compiègne, 60200 Compiègne, France
e-mail: {tariq.benamara, piotr.breitkopf}@utc.fr

[2] Cenaero ASBL
Rue des Frères Wright 29, 6041 Gosselies, Belgium
e-mail: {caroline.sainvitu, ingrid.lepot}@cenaero.be

[3] Snecma Villaroche
Rond Point René Ravaud-Réau, 77550 Moissy-Cramayel, France

Keywords: Multi-fidelity, Non-Intrusive POD, Surrogate-Based Optimization.

Abstract. *This paper presents a methodology for building multi-fidelity surrogate models based on Non-Intrusive Proper Orthogonal Decomposition. The proposed strategy aims at fusing multiple fidelity levels of simulation to improve the quality of surrogate models exploited in automated optimization loops of industrial-scale problems. A proof of concept is then given on a mathematical toy example which illustrates the ability of the proposed method to significantly reduce the overall computation cost. A 3D industrial study of a 1.5 stage booster is then presented to address the scaling capability of the proposed methodology.*

1 INTRODUCTION

Despite huge advances in the high performance computing domain, the integration of high-fidelity computations, taking up to thousands of CPU hours, into automated optimization loops still stands as a stumbling block for engineers. The design process being intrinsically multi-scale, multi-fidelity and multi-disciplinary [1]; the earlier the different scales, fidelities and disciplines are all integrated in the process, the better the technical solution can be [2, 3].

In this context, multi-fidelity surrogate modeling proposes a promising way to reduce the computational cost of industrial optimization [4]. The "data-fusion", also referred to as corrective approach in scaling methods [4, Section 6.1], enables the improvement of surrogate models [5] at a given computational cost. The most documented multi-fidelity model is the "co-Kriging" proposed by Kennedy and O'Hagan [6]. This technique has been widely applied to Multidisciplinary Design Optimization (MDO) [2, 5–11] and allows the enhancement of sparse high-fidelity information with cheaper low-fidelity data as well as the gradient of the modeled function [12, 13].

In the meantime, "Reduced-Order Models" (ROM) such as Proper Orthogonal Decomposition technique (POD) [14] have become very popular and their ability to focus on physical phenomena with highest impact on the performances targeted by the designer [15, 16] fosters their integration into new optimization strategies such as the "Non-Intrusive POD" (NI-POD) [17].

The aim of this work is to propose a methodology leveraging multi-fidelity computations during the training phase of an improved NI-POD model. This strategy can be seen as a re-interpretation of the constrained POD concept [15, 18] dedicated to multi-fidelity surrogate modeling. The methodology efficiency will be compared to classical NI-POD models as well as scalar mono- and multi-fidelity coKriging models.

The paper is organized as follows. Section 2 will present the NI-POD concept and the proposed extension. In Section 3, a proof of concept will be given on an academic example. Section 4 introduces the industrial application to illustrate the scalability of the proposed methodology. Finally, conclusions and perspectives are drawn in Section 5.

2 NON-INTRUSIVE POD

2.1 Classical NI-POD

POD, also known as "Karhunen-Loève expansion", or Principal Component Analysis (PCA) was introduced in the context of turbulence by Lumley [19] and allows to approximate high-dimensional numerical solution's output space by low-dimensional representation. The NI-POD methodology consists in building the orthonormal basis which optimally spans training database simulation's space and interpolating its coordinates in the basis via data-fitting methods [17].

The POD basis is computed by "snapshots" method [20] based on a Singular Value Decomposition (SVD) of M centered snapshots of the physical solution corresponding to M distinct positions $\boldsymbol{\Theta}$ in a p-dimensional design space $\mathcal{D} \subset \mathbb{K}^p$.

Let's consider a solver giving the physical response of any configuration in the design space $\mathcal{D}$,

$$\begin{aligned} \bar{\mathbf{s}} \ : \ \mathcal{D} &\rightarrow \mathbb{R}^n \\ \boldsymbol{\theta} &\rightarrow \bar{\mathbf{s}}(\boldsymbol{\theta}). \end{aligned} \tag{2.1}$$

The centered data are obtained by substracting the mean response over the training database $\mathbf{s}(\boldsymbol{\theta}) = \bar{\mathbf{s}}(\boldsymbol{\theta}) - \dfrac{1}{M}\sum_{i=1}^{M} \bar{\mathbf{s}}^{(i)}$. Following the formulation introduced in [21], the POD procedure gives the best projector $\mathcal{P}(\mathbf{s})$ on the orthonormal basis $\boldsymbol{\Phi}$ contained in the set of snapshots $\mathbf{S}$, in the sense of the Frobenius norm $\|A\|_F^2 = tr(A^\top A)$,

$$\boldsymbol{\Phi} = \underset{\boldsymbol{\Phi}}{argmin} \left(\|\mathbf{S} - \mathcal{P}(\mathbf{S})\|_F^2 \right), \text{ with } \boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \mathbf{I}, \tag{2.2}$$

and,

$$\mathcal{P}(\mathbf{s}) = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \mathbf{s}. \tag{2.3}$$

The POD basis $\boldsymbol{\Phi}$ lies in a $\mathbb{R}^{M\times n}$ space where $M \ll n$. This ROM procedure is usually associated with a truncating step that removes the less energetic modes yielding a POD basis $\boldsymbol{\Phi}^m$ in $\mathbb{R}^{m\times n}$, where $m < M$ and characterized by the error $\epsilon(m) = 1 - \dfrac{\sum_{i=1}^{m} \boldsymbol{\lambda}_i}{\sum_{j=1}^{M} \boldsymbol{\lambda}_j}$, where $\boldsymbol{\lambda}$ is the vector of monotically decreasing eigenvalues associated with the basis $\boldsymbol{\Phi}$. In the coming sections, we consider a full basis of size M, for the sake of simplicity.

Denoting $\boldsymbol{\alpha} = \boldsymbol{\Phi}^\top \mathbf{s}(\boldsymbol{\theta})$ the coordinate of a given configuration $\boldsymbol{\theta}$ in the POD basis, the NI-POD procedure approximates the vector $\boldsymbol{\alpha} \in \mathbb{R}^M$ over the whole design space $\mathcal{D}$. Assuming a good predictive behaviour of the built surrogate model, the NI-POD procedure enables the engineer for predicting the full vectorial response at any location in $\mathcal{D}$.

$$\tilde{\mathbf{s}}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \tag{2.4}$$

where $\forall i \in [\![1, M]\!]$, $\tilde{\boldsymbol{\alpha}}_i = \widetilde{\alpha_i}(\boldsymbol{\theta}) \backsimeq \alpha_i(\boldsymbol{\theta})$. It is important to keep in mind that the type of the surrogate model chosen to approximate the coefficients is critical. Depending on the sampling method, the variations in the simulation output vector over the design space and on the number of training samples, different techniques could lead to either good or bad approximation. One can find a non-exhaustive list of potential models in Forrester et al. [5, Section I.2].

2.2 Proposed multi-fidelity extension to NI-POD based surrogate modeling:

Given the computational cost of industrial simulations, the number of known locations in the design spaces is usually relatively low reducing the representativeness of the POD basis $\boldsymbol{\Phi}$ over the whole design space $\mathcal{D}$. The integration of multi-fidelity information tends to address this limitation by improving the basis and the interpolation of the coefficients surrogate models thanks to cheaper low-fidelity solutions the engineer can compute more intensively. Adapting the constrained POD modeling method introduced by Xiao et al. [15, 18], we propose to append the POD basis with low-fidelity modes projected on the complementary space spanned

by the high-fidelity data available. In the meantime, we improve the predictive capabilities of the coefficients surrogate models by increasing the number of constraints they are submitted to. Indeed, while the sparse high-fidelity projection coefficients are interpolated by the surrogate models, the higher density of known low-fidelity projection coefficients gives information about the global trend the surrogate models should follow.

Let's consider an approximated solution associated with the simulation output of a complex system characterized by its position in a p-dimensional design space $\mathcal{D}$. Assuming the two levels of fidelity at hand are lying in spaces with matching dimensions, we can denote the low- and high-fidelity solutions respectively $\mathbf{s}_L$ and $\mathbf{s}_H$ both in $\mathbb{R}^n$,

$$\begin{aligned} \mathbf{s} \ &: \mathcal{D} \rightarrow \mathbb{R}^n \\ &\boldsymbol{\theta} \rightarrow \mathbf{s}_L(\boldsymbol{\theta}) \\ &\boldsymbol{\theta} \rightarrow \mathbf{s}_H(\boldsymbol{\theta}). \end{aligned} \tag{2.5}$$

We compute the low- and high-fidelity databases $\mathbf{S}_L$ and $\mathbf{S}_H$ assuming the solutions $\mathbf{s}$ are already centered as presented in Section 2.1 on two Design of Experiments (DoE) $\boldsymbol{\Theta}_L$ and $\boldsymbol{\Theta}_H$.

$$\begin{aligned} \mathbf{S}_L(\boldsymbol{\Theta}_L) &= \begin{bmatrix} s_1^L(\boldsymbol{\theta}^{(1)}) & & s_1^L(\boldsymbol{\theta}^{(M_L)}) \\ \vdots & \cdots & \vdots \\ s_n^L(\boldsymbol{\theta}^{(1)}) & & s_n^L(\boldsymbol{\theta}^{(M_L)}) \end{bmatrix} \quad , \ \forall \boldsymbol{\theta}^{(i)} \in \boldsymbol{\Theta}_L \\ \mathbf{S}_H(\boldsymbol{\Theta}_H) &= \begin{bmatrix} s_1^H(\boldsymbol{\theta}^{(1)}) & & s_1^H(\boldsymbol{\theta}^{(M_H)}) \\ \vdots & \cdots & \vdots \\ s_n^H(\boldsymbol{\theta}^{(1)}) & & s_n^H(\boldsymbol{\theta}^{(M_H)}) \end{bmatrix} \quad , \ \forall \boldsymbol{\theta}^{(i)} \in \boldsymbol{\Theta}_H \end{aligned} \tag{2.6}$$

Usually, the number of low-fidelity experiments will be higher $M_L > M_H$, and no nested sampling is here required to build the DoE's $\boldsymbol{\Theta}_L$ and $\boldsymbol{\Theta}_H$. One should notice that all the presented notions are straightforwardly extendable to more than 2 different levels of fidelity.

2.2.1 Multi-fidelity basis

We first seek the best definition of the space spanned by the high-fidelity simulation output via a QR decomposition,

$$\mathbf{S}_H = [\mathbf{Q}_1 \ \mathbf{Q}_2] \, [\mathbf{R}_1 \ \mathbf{0}]^\top. \tag{2.7}$$

$\mathbf{Q}_1$ is an orthonormal basis spanning the high-fidelity image of the high-fidelity DoE while $\mathbf{Q}_2$ is a non-unique description of the complementary space of the space spanned by the basis $\mathbf{Q}_1$.

As the two simulation output spaces are both lying in $\mathbb{R}^n$, we can project the low-fidelity database $\mathbf{S}_L$ onto $\mathbf{Q}_2$ and perform an SVD to find the best projector for the low-fidelity information in the complementary space of the high-fidelity information

$$\mathbf{V} \boldsymbol{\lambda} \mathbf{U}^\top = \mathbf{Q}_2^\top \mathbf{S}_H \ . \tag{2.8}$$

We define the multi-fidelity POD basis $\mathbf{\Psi}$ in Eq. 2.9 as the best projector for the low-fidelity information with a cancelling constraint on the high-fidelity error

$$\mathbf{\Psi} = [\mathbf{Q}_1 \;\; \mathbf{Q}_2\mathbf{V}]\,, \tag{2.9}$$

with $\mathbf{\Psi} \in \mathbb{R}^{(M_L+M_H)\times n}$.

2.2.2 Multi-fidelity coefficients surrogate model

Once the multi-fidelity POD basis $\mathbf{\Psi}$ is created, the final steps of the classical NI-POD procedure 2.1 are repeated. Given two databases, the information used for building the surrogate model is tagged by its level of fidelity. We first project the low- and high-fidelity databases $\mathbf{S}_L$ and $\mathbf{S}_H$ on the basis $\mathbf{\Psi}$. This step enables the construction of data-fitting models on every real-valued coordinate of the vectors of coefficients $\boldsymbol{\alpha}$.

The surrogate modeling techniques dealing with multi-fidelity databases are of two kinds. The category of nested sampling based models does not restrain the high-fidelity database to any subset of the low-fidelity database. This category being the less used in the literature, we will focus on multi-fidelity surrogate models based on nested DoE's such as classical corrective models. We first project the two nested databases $\mathbf{S}_L$ and $\mathbf{S}_H$ onto $\mathbf{\Psi}$ and use the low-fidelity coefficients of projection to create a low-fidelity Radial Basis Function Network (RBFN) model $\tilde{\boldsymbol{\alpha}}_L(\boldsymbol{\theta}),\ \forall \boldsymbol{\theta} \in \mathcal{D}$. Then we build another RBFN model of the difference between low- and high-fidelity coefficients and define the coefficients approximation as the sum of the two previous models. The proposed model is an additive scaling model defined in Eq. 2.10 where the models $\tilde{\boldsymbol{\delta}}_{H-L}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\alpha}}_L(\boldsymbol{\theta})$ are one hidden-layer artificial neural networks

$$\tilde{\boldsymbol{\alpha}}_H(\boldsymbol{\theta}) = \tilde{\boldsymbol{\delta}}_{H-L}(\boldsymbol{\theta}) + \tilde{\boldsymbol{\alpha}}_L(\boldsymbol{\theta}) \;\; \forall \boldsymbol{\theta} \in \mathcal{D}\,. \tag{2.10}$$

3 VALIDATION ON MATHEMATICAL FUNCTIONS

In this section, the developed multi-fidelity NI-POD model is used to approximate vectorial analytical functions with different training sets. Various mathematical functions are referenced in the literature for the validation of optimization methodologies, but none, to the authors knowledge, exhibits a referenced low-fidelity definition and a multi-dimensional output space. We decided to derive our analytical test case from the multi-fidelity examples in [7] and [22, Section 3.5] as detailed in the coming section.

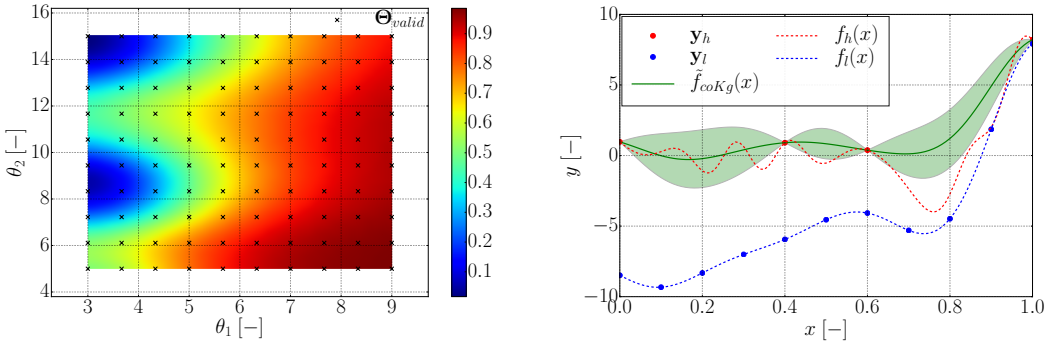
3.1 Mathematical definition

We assume two solvers associated with the functional solutions in Eq. 3.2. A 2-dimensional "design" space $\mathcal{D}$ is created by parameterizing the low- and high-fidelity solutions defining a 1-dimensional function

$$\begin{aligned} f : \mathcal{D} &\to \mathcal{F} \subset \mathcal{C}^\infty[0,1] \\ \boldsymbol{\theta} &\to f_{l_{\boldsymbol{\theta}}}(x) \\ \boldsymbol{\theta} &\to f_{h_{\boldsymbol{\theta}}}(x) \end{aligned} \tag{3.1}$$

with

$$
\begin{aligned}
f_{l_{\boldsymbol{\theta}}}(x) &= 0.5\left((\theta_1 x - 2)^2 \cdot \sin(12x - 4)\right) + \theta_2(x - 0.5) + 5, \\
f_{h_{\boldsymbol{\theta}}}(x) &= 0.5(\theta_1 x - 2)^2 \cdot \sin(12x - 4) + \sin(\theta_2 \cos(5x)).
\end{aligned}
\tag{3.2}
$$

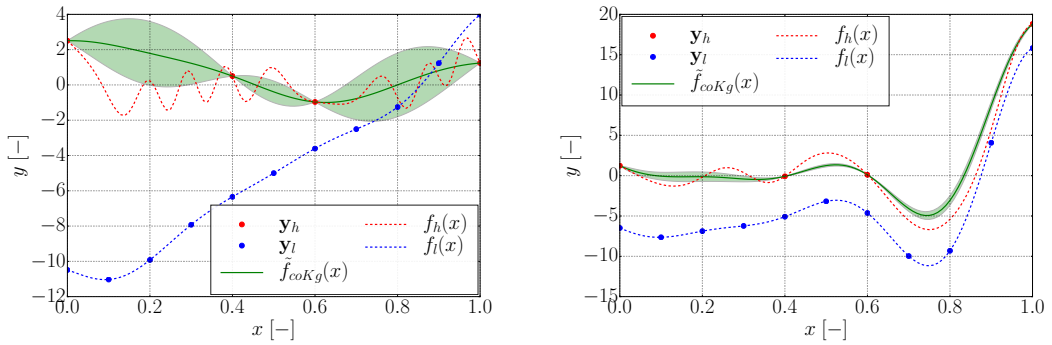


(a) Evolution of the Pearson's correlation coefficient over $\mathcal{D}$ (with bi-cubic interpolation)

(b) Low- and high-fidelity functions at reference location ($\theta_1 = 6$, $\theta_2 = 10$)

Figure 3.1: Correlation between low- and high-fidelity functions

Two vectorial solutions $\mathbf{s}_L$ and $\mathbf{s}_H$ are extracted from the functions $f_{l_{\boldsymbol{\theta}}}(x)$ and $f_{h_{\boldsymbol{\theta}}}(x)$ by sampling the $[0, 1]$-interval with 100 points evenly spaced. A first 10-level full factorial DoE $\boldsymbol{\Theta}_{valid}$ can show the correlation between low- and high-fidelity levels represented by its Pearson's correlation coefficient [23] painted on Figure 3.1a. We can distinguish different regions in the design space, some with good agreement (Fig. 3.2b) between low- and high-fidelity solutions and others where low- and high-fidelity are completely decorrelated (Fig. 3.2a), while the reference point (Fig. 3.1b) inspired by [7] and [22, Section 3.5] presents an average correlation level. As illustrated on Figure 3.2, the $2 - \sigma$ area (green shaded) of the coKriging model $\tilde{f}_{coKg}(x)$ is highly dependent on the low- to high-fidelity correlation intensity.



(a) Uncorrelated low- and high-fidelity functions ($\theta_1 = 4$, $\theta_2 = 14$)

(b) Strongly correlated low- and high-fidelity functions ($\theta_1 = 8$, $\theta_2 = 6$)

Figure 3.2: Range of correlations between low- and high-fidelity functions

3.2 Design of Experiments

Multiple nested Latin Hypercube DoE's ($\boldsymbol{\Theta}_H \subset \boldsymbol{\Theta}_L$) are created to assess the predictive performance of the proposed multi-fidelity model. Figure 3.3 shows a possible repartition of training points for a $2 + 10$ samples multi-fidelity DoE in the design space. The next section will present the results obtained with the proposed methodology in comparison with classical NI-POD models.
As the definition of a Latin Hypercube Sampling (LHS) is not unique, a statistical analysis of the impact of the training set on the model performances will be given for models trained on 100 different LHS for each couple (M_L, M_H).

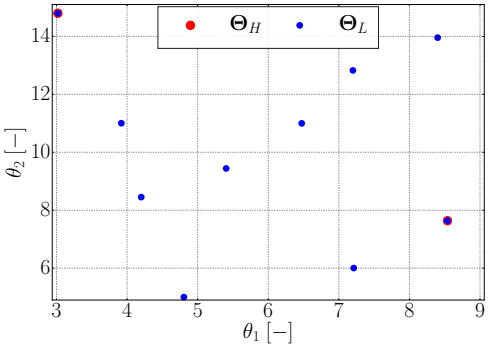


Figure 3.3: Multi-fidelity DoE in the design space $\mathcal{D}$ with $(M_L, M_H) = (10, 2)$

3.3 Results and classical NI-POD comparison

The performances of the proposed approach have to be evaluated in two phases. First, we want to assess the impact of the low-fidelity information gathered all over the design space on the quality of the multi-fidelity POD basis $\boldsymbol{\Psi}$ with respect to two representative bases:

- $\boldsymbol{\Phi}$: POD basis trained from $\mathbf{S}_H(\boldsymbol{\Theta}_H)$
- $\boldsymbol{\Phi}_\forall$: POD basis trained from $\mathbf{S}_H(\boldsymbol{\Theta}_L)$

The POD basis $\boldsymbol{\Phi}$ can be seen as the computational cost equivalent basis assuming the low-fidelity solution is almost free whereas the POD basis $\boldsymbol{\Phi}_\forall$ is the "best" basis we would have if we could afford intensive use of the high-fidelity solver.

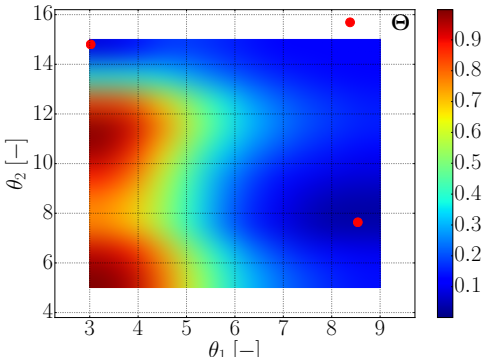
3.3.1 Comparison of POD basis quality

Considering the DoE illustrated on Figure 3.3, the three basis $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Phi}_\forall$ are built and the absolute L_2-error of the POD reconstruction of the solution $\mathbf{S}_H(\boldsymbol{\Theta}_{valid})$ is shown on Figure 3.4. The represented errors are expressed as following for any sample point in the design space $\mathcal{D}$ computed with the high-fidelity solver:

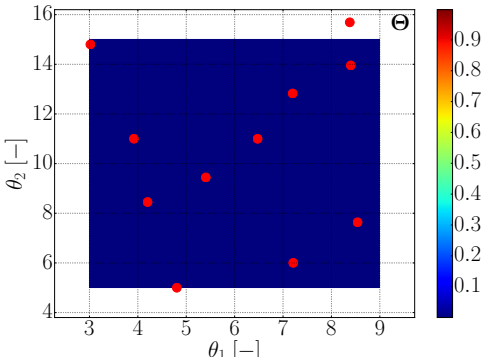
- $\epsilon_{\boldsymbol{\Phi}}(\boldsymbol{\theta}) = \sqrt{\frac{\|\mathbf{s}_H(\boldsymbol{\theta}) - \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \mathbf{s}_H(\boldsymbol{\theta})\|^2}{\|\mathbf{s}_H(\boldsymbol{\theta})\|^2}}$
- $\epsilon_{\boldsymbol{\Phi}_\forall}(\boldsymbol{\theta}) = \sqrt{\frac{\|\mathbf{s}_H(\boldsymbol{\theta}) - \boldsymbol{\Phi}_\forall\boldsymbol{\Phi}_\forall^\top \mathbf{s}_H(\boldsymbol{\theta})\|^2}{\|\mathbf{s}_H(\boldsymbol{\theta})\|^2}}$
- $\epsilon_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \sqrt{\frac{\|\mathbf{s}_H(\boldsymbol{\theta}) - \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \mathbf{s}_H(\boldsymbol{\theta})\|^2}{\|\mathbf{s}_H(\boldsymbol{\theta})\|^2}}$

One can notice that the information introduced into the multi-fidelity basis construction process globally reduces the absolute L_2-error in the design space (Fig. 3.4c). This is due to the enrichment of the basis with orthogonal modes taken from the low-fidelity solutions supposed to be relatively close to the high-fidelity solution. Adding these modes in the complementary of the high-fidelity output space cancels any contamination of the basis by incoherent information from the low-fidelity solutions. We can also see (Fig. 3.4b) that for this 2-dimensional design space, 10 samples are sufficient to drop drastically the reconstruction error.

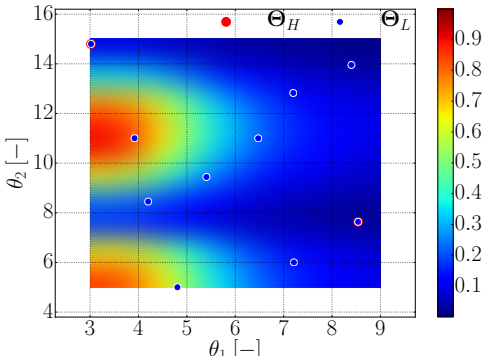
The proposed multi-fidelity POD basis tends to reduce the reconstruction error in both badly and strongly correlated low- and high-fidelity levels areas. This reduction of reconstruction error has to be confirmed after the surrogate modeling of the POD coefficients step.



(a) L_2-error of the $\mathbf{\Phi}$-projector : $\epsilon_{\mathbf{\Phi}}(\boldsymbol{\theta})$



(b) L_2-error of the $\mathbf{\Phi}_\forall$-projector : $\epsilon_{\mathbf{\Phi}_\forall}(\boldsymbol{\theta})$



(c) L_2-error of the $\mathbf{\Psi}$-projector : $\epsilon_{\mathbf{\Psi}}(\boldsymbol{\theta})$

Figure 3.4: Absolute L_2-error of the POD reconstruction over $\mathcal{D}$

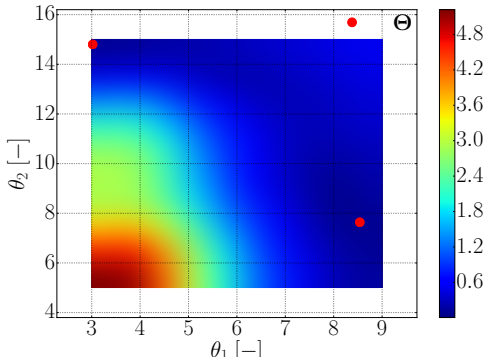
3.3.2 Comparison of NI-POD models quality

Once the POD basis is built, the training points are projected onto the basis and surrogate models are created to fit the coefficients databases. The chosen surrogate models are RBFN with tuned kernel and widths. For the proposed methodology, the surrogate model described on Equation 2.10 is used for each dimension of the vector of coefficients $\boldsymbol{\alpha}$.

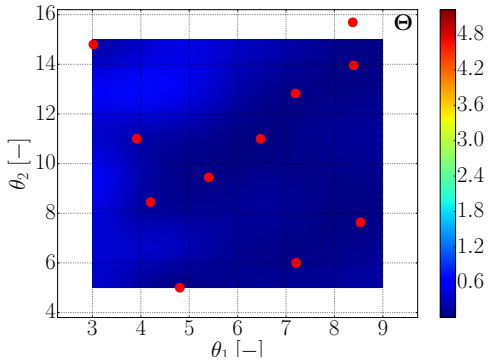
Figure 3.5 presents the prediction error, hereafter detailed, of the three NI-POD models compared:

- $\mathbf{\Phi}\tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta})$: NI-POD model trained on $\mathbf{S}_H(\mathbf{\Theta}_H)$,
  with scalar mono-fidelity RBFN for $\tilde{\alpha}_i \; \forall i \in [\![1, M_H]\!]$
- $\mathbf{\Phi}_\forall\tilde{\boldsymbol{\alpha}}_\forall(\boldsymbol{\theta})$: NI-POD model trained on $\mathbf{S}_H(\mathbf{\Theta}_L)$,
  with scalar mono-fidelity RBFN for $\tilde{\alpha}_i \; \forall i \in [\![1, M_L]\!]$
- $\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}^+_{H,L}(\boldsymbol{\theta})$: NI-POD model trained on $\mathbf{S}_H(\mathbf{\Theta}_H) \cup \mathbf{S}_L(\mathbf{\Theta}_L)$,
  with scalar multi-fidelity RBFN (Eq. 2.10) for $\tilde{\alpha}_i \; \forall i \in [\![1, M_H + M_L]\!]$.

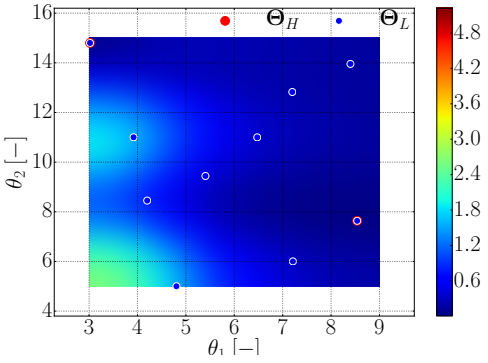
We can see on Figure 3.5b that for this test case, the 10 samples are sufficient to build a very good approximation of the POD coefficients as well. The maximum error observed in the lower-left corner of the design space on Figure 3.5a appears reduced on the Figure 3.5c as low-fidelity information is added to the model in regions where the low- to high-fidelity correlation coefficient is higher than 0.4 (Fig. 3.1a).



(a) L_2-error of the $\mathbf{\Phi}\tilde{\boldsymbol{\alpha}}$-predictor : $\epsilon_{\mathbf{\Phi}\tilde{\boldsymbol{\alpha}}}(\boldsymbol{\theta})$



(b) L_2-error of the $\mathbf{\Phi}_\forall\tilde{\boldsymbol{\alpha}}_\forall$-predictor : $\epsilon_{\mathbf{\Phi}_\forall\tilde{\boldsymbol{\alpha}}_\forall}(\boldsymbol{\theta})$



(c) L_2-error of the $\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}^+_{H,L}$-predictor : $\epsilon_{\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}^+_{H,L}}(\boldsymbol{\theta})$

Figure 3.5: Absolute L_2-error of the NI-POD prediction over $\mathcal{D}$

It is also important to keep in mind that the projection of the low-fidelity information on the complementary of the high-fidelity output space tends to reduce the impact of the POD coefficients associated with low-fidelity related modes as their contribution at high-fidelity location is cancelled. This behaviour is responsible for the non-pollution of the multi-fidelity NI-POD model in the top-left corner where both low- and high-fidelity are available and bad correlation coefficient is observed on Figure 3.1a.

3.3.3 Statistical analysis

To validate the impact of the proposed methodology in terms of computational cost, we propose to estimate the mean absolute L_2-error $\bar{\epsilon}$ over the entire design space depending on the size of the DoE's $\boldsymbol{\Theta}_H$ and $\boldsymbol{\Theta}_L$ and on the NI-POD based surrogate modeling technique. The construction of any DoE given couple of sizes (M_H, M_L) being non-unique, we build 100 independent DoE's for every couple $(M_H, M_L) \in [\![2]\!] \times [\![2, 28]\!]$. Figure 3.6 shows the decreasing trend of $\bar{\epsilon}_{\boldsymbol{\Phi}_\forall \tilde{\boldsymbol{\alpha}}_\forall}$ while $\bar{\epsilon}_{\boldsymbol{\Phi}\tilde{\boldsymbol{\alpha}}}$ remains statistically stable meaning that the 100-DoE's samples are sufficient to determine coherent conclusions for the study case at hand. We can also observe that the mean error associated with the proposed surrogate model $\bar{\epsilon}_{\boldsymbol{\Psi}\tilde{\alpha}^+_{h,l}}$ starts decreasing with added low-fidelity training points before stabilizing after $M_L > 8$. Once all the high-fidelity information lying in low-fidelity solutions is added to the POD basis, the computation of new low-fidelity experiments becomes useless and new high-fidelity computations are required to reduce further the global error of the surrogate model.

Statistical errors have also been compared for higher number of high-fidelity training snapshots and led to the same conclusions. These results are not illustrated because of too fast drop in the mean error due to the fact that the design space $\mathcal{D}$ is low-dimensional in this case ($p = 2$).

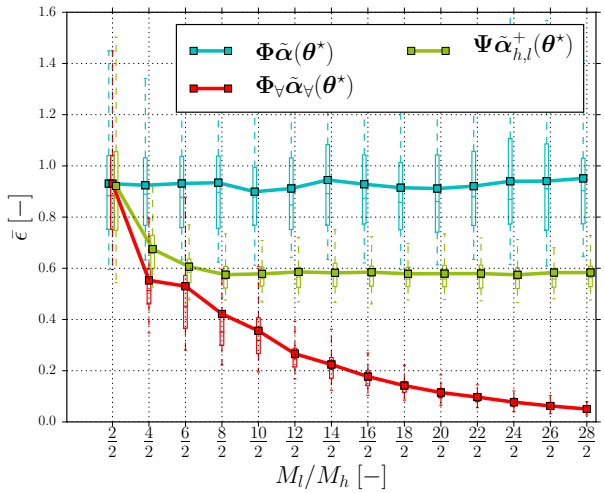


Figure 3.6: Evolution of the mean error $\bar{\epsilon}$ over $\mathcal{D}$ depending on the modeling technique

To investigate further the performance of the proposed modeling technique, we intend to test its efficiency on an industrial application presented hereafter.

4 INDUSTRIAL APPLICATION AND RESULTS

Aerodynamic design of Low-Pressure Compressors (LPC) is of major interest for jet engine manufacturers. The economical, technical and environmental constraints applied to turbofan designs force engineers to make a tradeoff between stage loading and both component efficiency and stability. Based on previous work presented by Lepot et al. [24], we propose to improve classical NI-POD surrogate models by adding low-fidelity information from coarse mesh based simulations.

4.1 Test case presentation

The test case presented in [24] consists in a surrogate-assisted optimization of a 1.5 stage booster based on $3D$ RANS simulations. The computational domain contains the inlet vane, the first rotating row and its following stator as illustrated on Figure 4.1. The deformed geometries are restrained to the rotating row, where the blade stacking as well as the hub axisymmetric profiling are parameterized respectively by 2 and 17 parameters. The geometric deformations are performed by either an in-house blade shape modeler for the stacking modifications or *CATIA v5R21* for the hub profiling. 3-level multigrid meshes are generated by *Autogrid v9r1.1* and $3D$ RANS simulations are computed by *elsA v3.3p1* [25]. Both high- and low-fidelity computations are conducted with $k-\varepsilon$ turbulence model without any near-wall treatment.

High-fidelity simulations are based on a structured mesh of about 5 million points preserving a y+ value below 1 along blades and endwalls. The convergence is accelerated thanks to a 2-level multigrid V-cycling leading to an overall run time around 2 hours for both design and near stall configurations on 32 Intel Xeon E5-2680v3 computation cores of Cenaero's cluster. Low-fidelity computations are performed on the mesh χ_{lf}, obtained by isotropic coarsening of the high-fidelity mesh χ_{hf}, leading to execution time around 15 minutes.

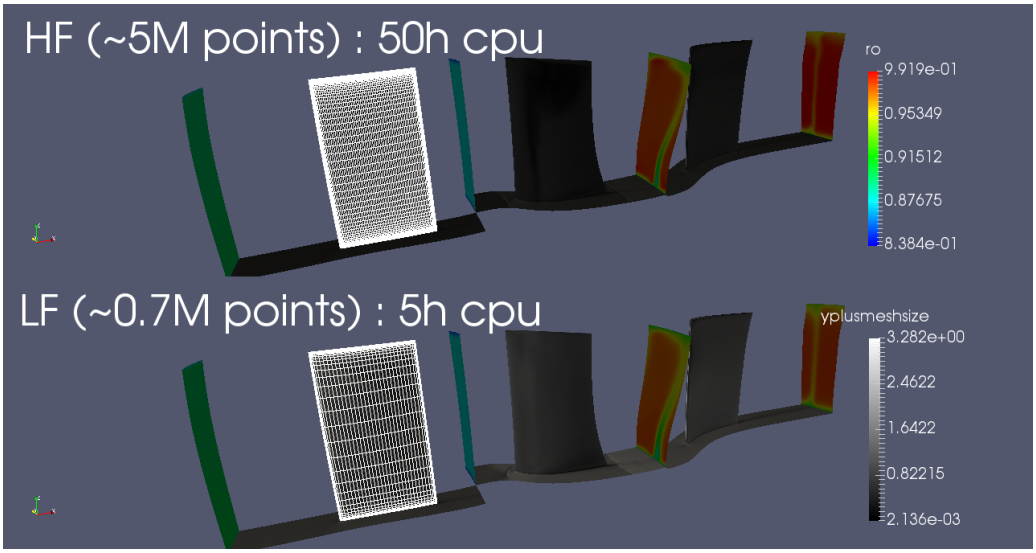


Figure 4.1: Low- and high-fidelity computational domains with isocontour of surface mesh coefficient y^+ and ρ distribution on mixing planes

4.2 Snapshot definition

We chose to approximate the conservative variables of the $3D$ computation in 4 sections defined by the inlet, outlet and inter-rows mixing planes. The definition of every mixing plane is

orthogonal to the revolution axis of the engine which avoid important mesh interpolation efforts.

As explained in Section 2.2, the low- and high-fidelity solutions lie in spaces with matching dimensions. The high-fidelity solution is therefore considered only at low-fidelity mesh point locations ("coarsening" step) to train the multi-fidelity NI-POD model and to compute scalar indicators such as isentropic efficiency. Figure 4.2 shows the impact of "coarsening" the high-fidelity solution on the computation of isentropic efficiency at design point η. We can see that considering high-fidelity solution only on the low-fidelity mesh χ_{lf} has a negligible impact on the isentropic efficiency. Indeed, the efficiency computed from the "coarsened" high-fidelity solution $\eta_{hf}[\chi_{lf}]$ (green points) is very close to the one integrated from the computed high-fidelity solution η_{hf} (red line). On the contrary, the isentropic efficiency integrated from low-fidelity solutions η_{lf} (blue points) presents non-negligible error with respect to the perfect match line (red line).

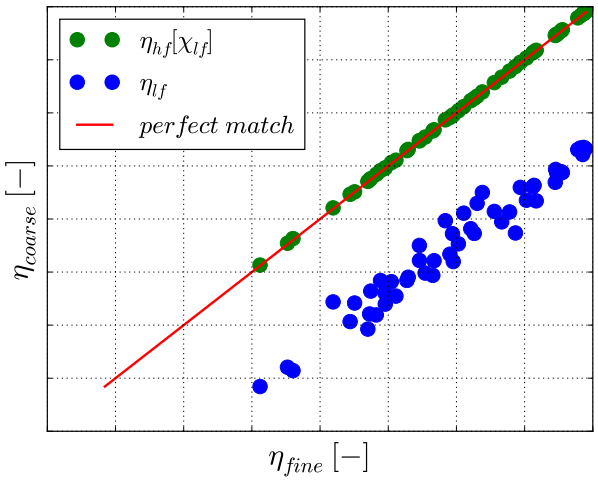


Figure 4.2: Low- to high-fidelity correlation of the efficiency for the 1.5 stage compressor at design configuration

4.3 Design of Experiments and multi-fidelity NI-POD

Given the overall computational cost of high-fidelity solutions, we decided to perform a 100-points DoE to train the proposed multi-fidelity NI-POD models as well as the POD and data-fitting based models used for validation purposes. Both the low- and high-fidelity solutions of the 100 training points are computed enabling a pseudo-statistical study of the impact of high-fidelity training samples on models performances. The success rate of the overall computation chain is about 60% removing all *CATIA* regeneration failures and all meshes with minimal skewness angle lower than 3°.

The presented results only concern the design point performance approximations. By randomly selecting high-fidelity solutions for a part of the training DoE we seek the impact of a priori sampling on the surrogate model quality. Different models are compared at predicting the high-fidelity design point efficiency computed for a 200-points validation DoE:

- Reconstruction capabilities assessment :

1. POD model built on the high-fidelity training set only $\mathbf{\Phi}\mathbf{\Phi}^\top$s
2. POD model built on a high-fidelity training set computationally equivalent to the multi-fidelity database $\mathbf{\Phi}_{ce}\mathbf{\Phi}_{ce}^\top$s
3. multi-fidelity POD model built on the high-fidelity training set enhanced by all low-fidelity training solutions $\mathbf{\Psi}\mathbf{\Psi}^\top$s

- Prediction capabilities assessment :

1. NI-POD model built on the high-fidelity training set only $\mathbf{\Phi}\tilde{\boldsymbol{\alpha}}$
2. NI-POD model built on a high-fidelity training set computationally equivalent to the multi-fidelity database $\mathbf{\Phi}_{ce}\widetilde{\boldsymbol{\alpha}_{ce}}$
3. multi-fidelity NI-POD model built on the high-fidelity training set enhanced by all low-fidelity training solutions $\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}^+$
4. mono-fidelity tuned radial basis function network built on the high-fidelity training set $\tilde{\eta}$
5. mono-fidelity tuned radial basis function network built on a high-fidelity training set computationally equivalent to the multi-fidelity database $\tilde{\eta}_{ce}$

4.4 Results

Figure 4.3 shows a comparison of the reconstruction capabilities of all POD-based models. One can see that the reconstruction error of all 3 models ($\mathbf{\Phi}\mathbf{\Phi}^\top$s, $\mathbf{\Phi}_{ce}\mathbf{\Phi}_{ce}^\top$s, and $\mathbf{\Psi}\mathbf{\Psi}^\top$s) is very low as the circles are close to the perfect match line (in red). This means that the POD basis contains already ($M_H \approx p$, where $\mathcal{D}$ is p-dimensional) all the modal information needed to reproduce the conservative variables distributions of the 110 successfully computed samples in the validation DoE. Note that the reconstruction capability of a POD based model is totally useless for a "real world" application as projecting a solution on the POD basis first requires the expensive computation it is derived from.

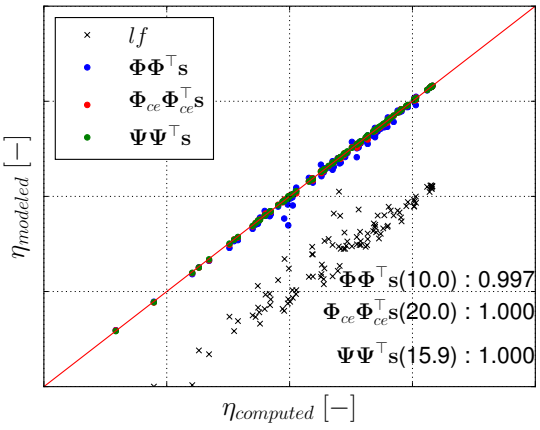
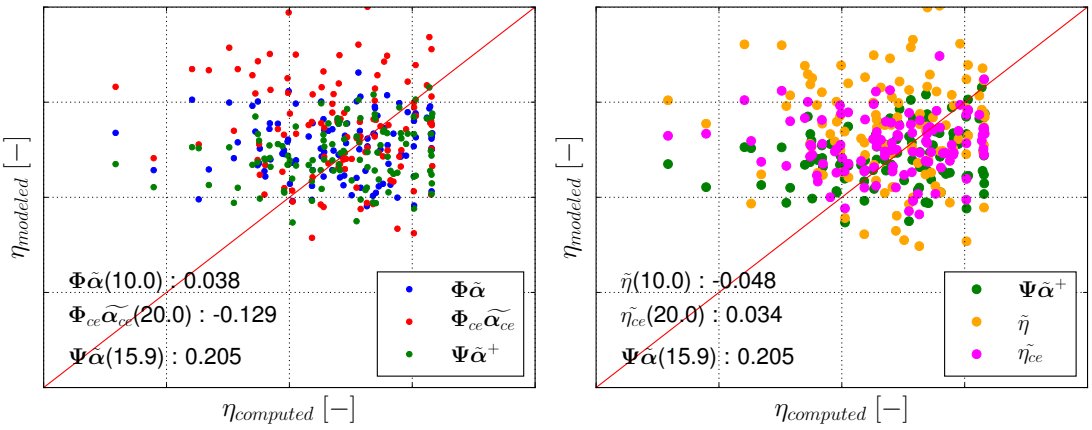


Figure 4.3: Computed- to modeled-efficiency correlation for POD-based models reconstruction

On the contrary, the predictive capabilities of NI-POD based models present more differences depending on the chosen approach (see Fig. 4.4a). The integration of low-fidelity points for

the training step increases the quality (correlation coefficients on Fig 4.4) of the proposed model with both respect to classical NI-POD models and mono-fidelity scalar RBF-based models (Fig. 4.4b). We can see on Fig. 4.4a that the green points (associated to the isentropic efficiency prediction of the proposed model) are closer on average to the perfect match line than the blue and red points (associated to the isentropic efficiency prediction of classical NI-POD models). The same observation can be made for mono-fidelity RBFN based models (see Fig. 4.4b).



(a) Computed- to modeled-efficiency correlation for NI-POD based models prediction

(b) Computed- to modeled-efficiency correlation for radial basis function based models prediction

Figure 4.4: Comparison of predictive capabilities of NI-POD and radial basis functions based models

As mentioned earlier, we try to study the impact of the initial high-fidelity training set $\mathbf{\Theta}_H$ on the models quality. Table 1 shows that the proposed methodology is responsible on average for an increase of the model quality and tends to limit the impact of the initial DoE.

		Pearson's coefficient (μ)	Pearson's coefficient (σ)	CPU cost (μ)
POD	$\mathbf{\Phi\Phi}^\top\mathbf{s}$	0.9913	0.0311	10.7
	$\mathbf{\Phi}_{ce}\mathbf{\Phi}_{ce}^\top\mathbf{s}$	0.9996	0.0008	17.7
	$\mathbf{\Psi\Psi}^\top\mathbf{s}$	1.0000	0.0000	16.6
NIPOD	$\mathbf{\Phi}\tilde{\boldsymbol{\alpha}}$	0.0586	0.1611	10.7
	$\mathbf{\Phi}_{ce}\widetilde{\boldsymbol{\alpha}_{ce}}$	0.0803	0.1558	17.7
	$\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}^+$	0.1816	0.0588	16.6
RBF	$\tilde{\eta}$	0.0707	0.1573	10.7
	$\tilde{\eta}_{ce}$	0.0763	0.1491	17.7

Table 1: Mean and standard deviation of Pearson's correlation coefficients for different models trained with 50 different $\mathbf{\Theta}_H$ of given mean costs

As the selection of high-fidelity training points is non-uniformly random, the overall cost of training points computation is not constant. The mean value of training CPU costs is given in

Table 1, and shows that despite its higher mean training cost, the "computationally equivalent" high-fidelity trained NI-POD model presents lower quality that the proposed multi-fidelity model. This is due to higher coverage of the design space with low-fidelity samples improving globally the model quality.

5 CONCLUSIONS AND PERSPECTIVES

One can consider non intrusive-POD models as better integrators of physics than classical data-fitting based models. In the context of surrogate-assisted optimization, this leads to better insight in the relation between design parameters and performance indicators and similarly between parameters and the global system behaviour. Based on this assumption, the current paper has presented an extension of non-intrusive POD to multi-fidelity design. The proposed methodology leverages relatively high density of low-fidelity solutions in the design space to enhance both the POD basis and the surrogate models of coefficients of classical NI-POD models. First implementation of the presented technique has been tested on an analytical test case inspired by coKriging related articles [7, 22] and an industrial test case of a 1.5-stage low-pressure compressor [24].

The proposed model efficiency has been compared to classical NI-POD models and mono-fidelity RBFN-based models. The multi-fidelity NI-POD model significantly improved both the POD basis representativeness and the NI-POD predictive capabilities on the analytical test case. The results showed the contribution of low-fidelity samples in the prediction error of the model. At the same time, the potential limit in error reduction due to the lack of correlation between low- and high-fidelity solutions in some regions of the design space has been revealed.

The prediction error reduction, obtained from added low-fidelity information, has also been verified on the industrial scale problem in Section 4. The models performances have been compared on the isentropic efficiency at design point showing again the impact of densely spaced low-fidelity information. Despite relatively low correlations observed, the results showed a real potential of the proposed method to reach the objective of starting an optimization with DoE samples as large as the problem dimensionality ($M_H \approx p$).

Some issues remain to be addressed concerning the building methodology itself as well as some questions remain opened on the industrial application presented. The question of truncation is one of the cornerstones of NI-POD based approximation and is totally opened when dealing with multi-fidelity POD models. The impact of truncation has to be questionned especially for problems where only a few modes are needed but highly non-linear behaviour is observed for projection coefficients. The impact of snapshot definition is also to be investigated. In the presented test case, the stacking deformation incited us to post-process conservative variables only on mixing planes. Because of relatively important distances between rows and diffusion, the physical complexity of wake and secondary flows can vanish and be missed in the integrated snapshots. The extension of the proposed study to other performance indicators and off-design configurations would allow more precise comparison with the results highlighted in [24].

In terms of general perspectives, a comparison with Kriging or coKriging models could be considered in future work where none is presented in this study because of the hyper-parameters optimization cost in high-dimensional design space. Last but not least, the integration of the proposed methodology in an automated optimization loop has also to be performed and would

check the validity of the given observations for higher correlation coefficients needed for a successfull surrogate-assisted optimization.

Acknowledgements

The present work was partly founded by the *Association Nationale de la Recherche et Technologie*. The authors would like to thank *Snecma* and *Techspace Aero* from the *SAFRAN Group* for their support and permission to publish this study and especially Ir. Stephane Hiernaux and Ir. Jean Coussirou for their technical support in this research project.

References

[1] E. Tromme, O. Brüls, J. Emonds-Alt, M. Bruyneel, G. Virlez, and P. Duysinx. "Discussion on the optimization problem formulation of flexible components in multibody systems". In: *Structural and Multidisciplinary Optimization* 48.6 (2013), pp. 1189–1206.

[2] A. March and K. Willcox. "Constrained multifidelity optimization using model calibration". In: *Structural and Multidisciplinary Optimization* 46.1 (May 2012), pp. 93–109.

[3] A. J. Keane. "Wing optimization using design of experiment, response surface, and data fusion methods". en. In: *Journal of Aircraft* 40 (July 2003), pp. 741–750.

[4] A. J. Keane and P. B. Nair. *Computational Approaches for Aerospace Design: The Prusuit of Excellence*. John Wiley & Sons, 2005.

[5] A. I. J. Forrester, A. Sóbester, and A. J. Keane. *Engineering Design via Surrogate Modelling : A Practical Guide*. John Wiley & Sons, 2008.

[6] M. C. Kennedy and A. O'Hagan. "Predicting the output from a complex computer code when fast approximations are available". In: *Biometrika* 87.1 (Jan. 2000), pp. 1–13.

[7] A. I. J. Forrester and A. J. Keane. "Recent advances in surrogate-based optimization". In: *Progress in Aerospace Sciences* 45.1-3 (Jan. 2009), pp. 50–79.

[8] Y. Kuya, K. Takeda, X. Zhang, and A. I. J. Forrester. "Multifidelity surrogate modeling of experimental and computational aerodynamic data sets". In: *AIAA Journal* 49.2 (Feb. 2011), pp. 289–298.

[9] D. J. Toal and A. J. Keane. "Efficient multipoint aerodynamic design optimization via cokriging". en. In: *Journal of Aircraft* 48.5 (Sept. 2011), pp. 1685–1695.

[10] L. Huang, Z. Gao, and D. Zhang. "Research on multi-fidelity aerodynamic optimization methods". en. In: *Chinese Journal of Aeronautics* 26.2 (Apr. 2013), pp. 279–286.

[11] Z.-H. Han, S. Görtz, and R. Hain. "A Variable-Fidelity Modeling Method for Aero-Loads Prediction". English. In: *New Results in Numerical and Experimental Fluid Mechanics VII*. Ed. by A. Dillmann, G. Heller, M. Klaas, H.-P. Kreplin, W. Nitsche, and W. Schröder. Vol. 112. Notes on Numerical Fluid Mechanics and Multidisciplinary Design. Springer Berlin Heidelberg, 2010, pp. 17–25.

[12] Z.-H. Han, R. Zimmermann, and S. Görtz. "A new cokriging method for variable-fidelity surrogate modeling of aerodynamic data". In: *Proc. 48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*. AIAA, Jan. 2010.

[13] Z.-H. Han, S. Görtz, and R. Zimmermann. "Improving variable-fidelity surrogate modeling via gradient-enhanced Kriging and a generalized hybrid bridge function". In: *Aerospace Science and Technology* 25.1 (Mar. 2013), pp. 177–189.

[14] P. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Second. Cambridge Books Online. Cambridge University Press, 2012.

[15] M. Xiao, P. Breitkopf, R. F. Coelho, P. Villon, and W. Zhang. "Proper orthogonal decomposition with high number of linear constraints for aerodynamical shape optimization". In: *Applied Mathematics and Computation* 247.0 (2014), pp. 1096–1112.

[16] R. Filomeno Coelho, P. Breitkopf, and C. Knopf-Lenoir. "Model reduction for multidisciplinary optimization - application to a 2D wing". In: *Structural and Multidisciplinary Optimization* 37.1 (2008), pp. 29–48.

[17] M. Guénot, I. Lepot, C. Sainvitu, J. Goblet, and R. Filomeno Coelho. "Adaptive sampling strategies for non-intrusive POD-based surrogates". In: *Engineering Computations* 30.4 (Jan. 2013), pp. 521–547.

[18] M. Xiao, P. Breitkopf, R. F. Coelho, C. Knopf-Lenoir, P. Villon, and W. Zhang. "Constrained Proper Orthogonal Decomposition based on QR-factorization for aerodynamical shape optimization". In: *Applied Mathematics and Computation* 223.0 (2013), pp. 254–263.

[19] J. L. Lumley. "The structure of inhomogeneous turbulent flows". In: *Atmospheric turbulence and radio propagation*. Ed. by A. M. Yaglom and V. I. Tatarski. Nauka, 1967, pp. 166–178.

[20] L. Sirovich. *Turbulence and the Dynamics of Coherent Structures, Part1: Coherent Structures*. Vol. 45. Quarterly of Applied Mathematics 3. Brown University, Division of Applied Mathematics, Oct. 1987, pp. 561–571.

[21] B. Raghavan, L. Xia, P. Breitkopf, and P. Villon. "Towards simultaneous reduction of both input and output spaces for interactive simulation-based structural design". In: *Computer Methods in Applied Mechanics and Engineering* 265 (July 2013), pp. 174–185.

[22] L. Le Gratiet. "Multi-fidelity Gaussian process regression for computer experiments". PhD thesis. Université Paris-Diderot-Paris VII, Oct. 2013, 306pp.

[23] K. Pearson. "Mathematical Contributions to the theory of evolution. III. Regression, Heredity, and Panmixia". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 187 (1896), pp. 253–318.

[24] I. Lepot, T. Mengistu, S. Hiernaux, and O. De Vriendt. "Highly loaded LPC blade and non axisymmetric hub profiling optimization for enhanced efficiency and stability". In: *ASME Turbo Expo 2011: Turbine Technical Conference and Exposition*. Vol. 7. Jan. 2011, pp. 285–295.

[25] L. Cambier, S. Heib, and S. Plot. "The Onera elsA CFD software : input from research and feedback from industry". In: *Mechanics & Industry* 14.3 (Apr. 2013), pp. 159–174.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

EVOLUTIONARY AERODYNAMIC SHAPE OPTIMIZATION THROUGH THE RBF4AERO PLATFORM

Dimitrios H. Kapsoulis[1], Varvara G. Asouti[1], Kyriakos C. Giannakoglou[1], Stefano Porziani[2], Emiliano Costa[2], Corrado Groth[3], Ubaldo Cella[3] and Marco E. Biancolini[3]

[1]National Technical University of Athens (NTUA), School of Mech. Eng.,
Parallel CFD & Optimization Unit, Athens, Greece
e-mail: jim.kapsoulis@gmail.com, vasouti@mail.ntua.gr, kgianna@central.ntua.gr

[2] D'Appolonia S.p.A.
Viale Cesare Pavese 305, 00144 Rome, Italy
e-mail: {stefano.porziani, emiliano.costa}@dappolonia.it

[3] University of Rome Tor Vergata, Dept. of Enterprise Engineering
Via Politecnico 1, 00133, Rome Italy
e-mail: biancolini@ing.uniroma2.it

Keywords: Aerodynamic Optimization, Metamodels, Evolutionary Algorithms, Radial Basis Functions, Mesh Morphing.

Abstract. *This paper demonstrates a way of solving industrial aerodynamic shape optimization problems using the RBF4AERO platform, developed in the framework of the EU–funded RBF4AERO project. The platform combines optimization algorithms (stochastic and gradient–based ones), a mesh morphing tool based on Radial Basis Functions (RBFs) and various evaluation tools (CFD, CSD, etc). In this paper, the Evolutionary Algorithms (EAs) based tool assisted by metamodels trained on a sampling performed during the Design of Experiment phase is used along with a CFD evaluation tool. The use of Response Surface Models (RSM) significantly reduces the number of CFD runs required to reach the optimal solution(s), while the use of RBF–based mesh morphing avoids re–meshing prior to each CFD–based evaluation. In optimization problems, the platform starts by selecting a number of individuals to undergo CFD–based evaluations. The latter constitute the training patterns for the RSM which is, then, used as a low-cost evaluation tool within the EA–based optimization. "Optimal" solutions found by the EA–based search exclusively based on the trained metamodel are then re-evaluated by means of the CFD tool. The database of already evaluated individuals is updated, the RSM is re–trained and the EA–based optimization is repeated. The optimization terminates when convergence criteria related to the RSM prediction accuracy are met or the computational budget is exhausted. The optimization of an ultra-light aircraft and that of the DrivAer car model for minimum drag are showcased. In all cases presented, the simpleFoam solver of OpenFOAM is used to evaluate candidate solutions.*

1 INTRODUCTION

During the last decades, Evolutionary Algorithms (EAs) have widely been applied to shape optimization problems. Though EAs can easily accommodate any evaluation software as a black–box tool, they ask for a great number of evaluations to reach the optimal solution(s) which often makes their utilization in optimization problems prohibitively expensive, especially in industrial problems with excessive computational cost per evaluation. A known remedy to this problem is the use of surrogate evaluation models or metamodels, leading to the so–called metamodel–assisted EAs (MAEAs). Metamodels replace the exact evaluation tool, thus reducing the total number of evaluations required to reach the optimal solution(s).

MAEAs can be classified to on–line and off–line trained ones, based on whether the metamodel training takes place during the evolution or not. On MAEAs with on–line trained metamodels, the EA relies upon the interleaving usage of the problem–specific (to be considered as the high–fidelity) evaluation tool and the metamodel. The entire EA population can be evaluated with any of these tools, [1, 2], by switching either periodically or based on some criteria. Another option is to pre–evaluate all the generation members on the metamodel and use the high–fidelity tool only for the most promising, according to the metamodel, individuals out of them [3, 4, 5, 6]. Metamodels used can be of either local or global support. On the other side, off–line trained metamodels usually rely on a single global metamodel and a sampling technique, often referred to as Design of Experiments (DoE), [7] which defines the appropriate training patterns. Once the metamodel has been trained, this becomes the only evaluation tool during the EA-based search.

In the RBF4AERO platform, developed in the framework of the EU–funded RBF4AERO project, the optimization algorithm relies on off–line trained metamodels. In particular, a polynomial-based Response Surface Model (RSM), trained on appropriately selected samples in the design space, is implemented.

Another important feature of the platform is the mesh morphing tool. Generally, mesh morphing techniques are based on shape parameterization, which can parameterize the surface along with the surrounding nodes of the interior mesh. These techniques allow the interior of the computational mesh to be deformed, avoiding, thus, costly re-meshing. Methods like Radial Basis Functions (RBFs) [8, 9] and volumetric B-splines or NURBS [10] can provide the required parameterization and mesh deformation. The RBF4AERO platform makes use of a morphing tool implementing RBFs, [11]. A number of parameters controlling the positions of groups of RBF control points are used as design variables. Though the RBF4AERO platform may accommodate several CFD tools, for the applications presented in this paper the simpleFoam solver of the OpenFOAM is used, since we are dealing with low speed applications.

In what follows, the basic features of the platform are described and, then, this is used for the re–design/shape optimization of an ultra–light aircraft and a car model.

2 OPTIMIZATION THROUGH THE RBF4AERO PLATFORM

The EA–based optimization algorithm of the RBF4AERO platform is summarized below:

1. Define the RBF shape modifications that require a specific set–up and a steerable parametric mesh.
2. Sample the design space with a DoE technique.
3. Evaluate the above samples on the high–fidelity (herein CFD) tool, after appropriately morphing a pre–existing computational mesh using the RBF–based morpher tool. Store

the so–computed objective function value(s) in the database (DB), paired with the corresponding values of the design variables.

4. Train the RSM using the DB entries as training patterns.
5. Perform EA–based optimization, exclusively based on evaluations on the RSM.
6. Re–evaluate ”optimal” solution(s) resulting from step 5 on the exact evaluation tool and update DB.
7. Return to step 4 until convergence.

In the re–evaluation phase, step 6, there is the possibility to perform additional high–fidelity evaluations, if one area of the design space is not explored properly. Below, the basic features of each step are further described. Note that the optimization algorithm settings, i.e. DoE, RSM and EA parameters are user–defined through the RBF4AERO platform Graphical User Interface.

2.1 Design space sampling

The RBF4AERO platform offers three DoE [12, 13] options to sample the design space, namely the full factorial, partial factorial and randomized designs. Either option decides how many and which candidate solutions will be evaluated on the CFD tool. To perform DoE, each design variable is discretized to a number of user–defined levels. In case of a full factorial design, all possible combinations of design variables' levels are considered. Full factorial designs may result to high CPU cost since all possible combinations must then be evaluated on the computationally expensive tool. The partial (or fractional) factorial design extracts a sub–group from the corresponding full factorial design after cutting some user–selected “less important” variables off. To obtain the partial factorial, a full factorial design is firstly defined using only the “important” variables and the levels of the cut–off variables results from combinations of the other variables' levels. The partial factorial is a better compromise in terms of CPU cost but requires a very good knowledge of the problem in hand in order to correctly decide the design variables to cut–off. Finally, in the randomized design, a user–defined number of designs must be generated. To do so the algorithm computes all possible partial factorial designs and keeps the one with number of members less or equal to the user–defined number of individuals. If the number of collected members is less than the user–defined one, the remaining design vectors are selected from the remaining members of the corresponding full factorial design at random.

2.2 Individuals mesh morphing and evaluation

The individuals determined by the DoE are evaluated on the CFD tool. In this paper, all evaluations are carried out using the steady state solver of OpenFOAM. Instead of re–meshing the computational domain for each and every geometry change during the DoE, an RBF–based morpher undertakes the morphing of a baseline mesh before delivering it to the evaluation manager for the CFD run. In all cases, the baseline mesh is generated using the snappyHexMesh tool of OpenFOAM.

RBFs are mathematical functions able to interpolate data defined at discrete points only (source points) in an n-dimensional environment. The RBF function has the following form

$$s(\mathbf{x}) = \sum_{k=1}^{K} \gamma_k \, \phi(||\mathbf{x} - \mathbf{x}_k||) + h(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ the position vector of a mesh node and γ_k the coefficients of the polynomial, fitted imposing the known values at source points $\mathbf{x}_k$. The interpolation quality and its behavior depends on the chosen ϕ function, being of either global or compact support. In 3D mesh morphing, the bi–harmonic function $\phi(r) = r$ is adopted for its smoothing abilities. The solution of the RBF mathematical problem consists of the computation of the scalar coefficients of a linear system of order equal to the number of considered source points (RBF centers). Once the RBF system coefficients have been computed, the displacement of an arbitrary mesh node can be expressed as a function of the distance-based contributions from the RBF centers.

The RBF method has several advantages that make it very attractive for mesh morphing. Being a mesh–less method, where only grid points are moved regardless their connectivity, its parallelization is quite straightforward. Over and above, it is able to exactly prescribe known deformations onto the surface mesh. This can be achieved by using all the mesh nodes as RBF centers with prescribed displacements, including the simple zero field to denote a surface which is left untouched by the morphing action.

The industrial implementation of the RBF mesh morphing poses two challenges, (a) the numerical complexity related to the solution of the RBF problem for a large number of centers and (b) the definition of suitable paradigms to effectively control shapes using RBF. The RBF Morph software, [11], included in the RBF4AERO platform addresses both challenges as it comes with a fast RBF solver capable to fit large dataset (hundreds of thousands of RBF points can be fitted in a few minutes) and with a suite of modeling tools that allows the user to easily set–up each shape modification.

2.3 Response surface model training

The surrogate evaluation model (or metamodel) implemented in the platform is a Response Surface Model [7] based on polynomial functions. The mathematical expression of the response is

$$\hat{F}(\vec{x}) = b_0 + \sum_{i=1}^{N} \sum_{j=1}^{P_i} b_{ij} x_i^j + \sum_{j=0}^{M} a_j \prod_{i=1}^{N} x_i^{I_i} \tag{2}$$

where $\hat{F}$ is the approximate objective function value, N is the number of design variables, x_i is the ith design variable, M the number of interactions, I_i the power that the ith design variable is raised to and P_i the maximum power for each variable. The RSM coefficients b_{ij}, a_j are computed during the training phase. Interactions, [14], denote the relationship among design variables and are mathematically expressed as multiplications among the related/interacting variables. In case that the number of training patterns exceeds the number of coefficients to be computed, the least–squares method is used during the RSM training. The number of equations constructed by the least–squares method is equal to the number of coefficients. As a result, the cost of the training depends on the number of used coefficients.

The P_i and I_i values in equation 2 can either be selected by the user or defined automatically by minimizing the RSM's error. Note that a different RSM, i.e. with a different configuration (maximum powers, interactions, etc) is trained for each objective function and constraint.

2.4 Optimization using evolutionary algorithms

Once the RSM has been trained, a (μ, λ)EA, with μ parents and λ offspring, undertakes the optimization by exclusively using the RSM as the evaluation tool. A real coded EA with tournament selection for the parents population and simulated binary crossover scheme is im-

plemented. Note that the computational cost of this phase is negligible since the previously trained RSM is used to approximately evaluate all candidate solutions. Upon the termination of the evolution, the "optimal" solution(s) resulting from EA are re–evaluated on the high-fidelity tool and added to the DB.

2.5 Convergence/Termination criteria

The optimization platform includes three convergence criteria, checked upon completion of each optimization cycle. The first convergence criterion is related to the CPU clock time of the optimization (computational budget) and is defined as the maximum number of high-fidelity evaluations that can be performed with the provided budget. This limits the size of the initial sampling and the number of optimization cycles to be performed. The second convergence criterion is related to the RSM prediction accuracy. The optimal solution is considered as found, if the RSM error is very small and its prediction practically replicates the objective function value which results from the problem–specific evaluation tool. The third convergence criterion checks if the "optimal" solution found by the EA is not improving during a user-defined number of evaluations.

3 OPTIMIZATION OF AN ULTRA–LIGHT AIRCRAFT

The first case is concerned with the re–design/optimization of an ultra–light aircraft [15] aiming at the minimization of its drag coefficient. The aircraft geometry was provided by Pipistrel, a light aircraft manufacturer, partner in the RBF4AERO project. In particular, the re–design focuses on the wing root–body junction by defining two boxes. The larger one is used to limit the morphing action and remains fixed thanks to the proper spacing of RBF points. The smaller box, which includes the whole wing, is allowed to move along all three directions, leading to 3 design variables in total (the degrees of freedom are the displacements of the control box in the x, y and z axes). The mesh lying in the space between the small box (which is allowed to change) and the large one (still), including part of the fuselage and wing surfaces, is deformed by the RBF model. A close–up view of the wing root–body junction and the morphing boxes is shown in figure 1.

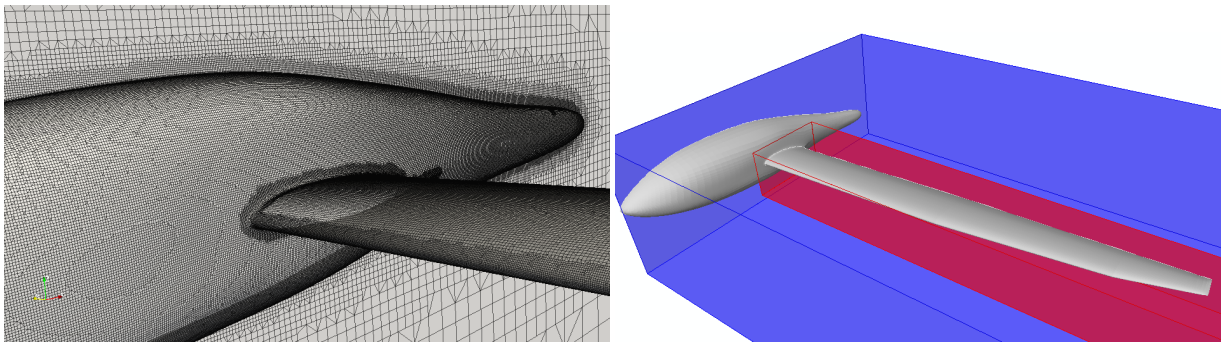


Figure 1: Optimization of an ultra–light aircraft. Close–up view of the wing root–body junction on a surface mesh (left) and RBF control boxes (right).

The flow conditions are $M_\infty = 0.08$, flow angle 10° and $Re = 10^6$ (based on the wing chord). Each candidate solution is evaluated on the OpenFOAM incompressible solver (simpleFoam) coupled with the Spalart–Allmaras turbulence model with wall functions. The computational mesh around the aircraft is unstructured and consists of about 4.7M cells.

The optimization starts by evaluating, on the aforementioned CFD tool, 45 samples which are used for training a sixth degree ($P_i = 6$ in equation 2) RSM. Then a $(\mu, \lambda) = (15, 30)$ EA undertakes the optimization with a termination criterion of 500 evaluations on the RSM. The "optimal" solution resulting from the EA is re–evaluated on the CFD tool and the RSM is trained anew. Note that each re–training results in different degree and coefficients in equation 2, since the algorithm automatically defines the parameters needed (see section 2.3). This procedure is repeated ten times before meeting the convergence criteria of the optimization procedure (herein, RSM error criterion is firstly met), resulting into a total cost of 55 evaluations on the CFD tool. The convergence history of the optimization is shown in figure 2. Note that the horizontal axis starts from 45 since the optimization started after evaluating 45 samples during the DoE phase.

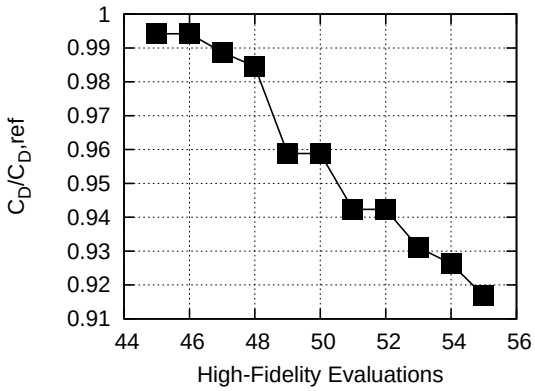


Figure 2: Optimization of an ultra–light aircraft. Convergence history of the optimization.

The optimized geometry yields drag coefficient which is lower by 9% compared to the reference one. The displacement of the junction towards the rear and bottom part of the fuselage (figure 3) is responsible for the observed objective function reduction. A by–product of this optimization is that the lift of the optimized geometry becomes higher, without being included in the objective function. Figures 3, 4 and 5 compare the reference and the optimized aircraft geometries. In all figures, the pressure field on the aircraft surface is plotted.

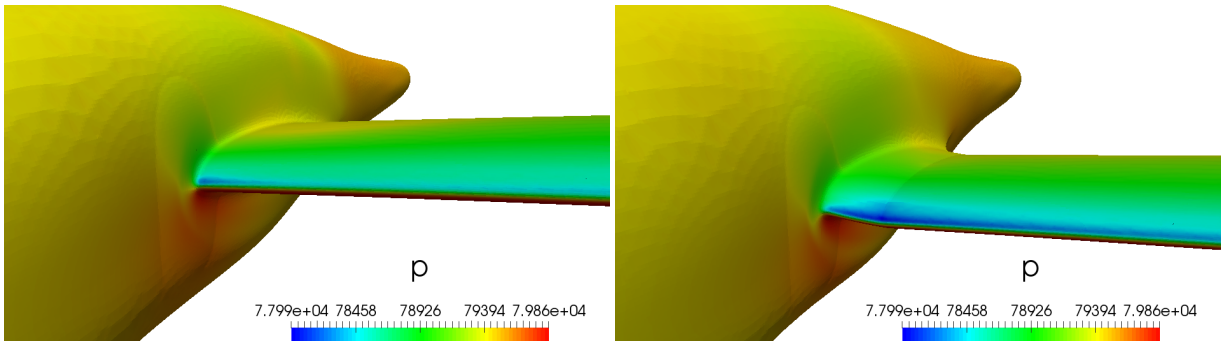


Figure 3: Optimization of an ultra–light aircraft. Comparison of the reference (left) and optimized (right) aircraft geometries. Close–up front view of the wing root–body junction.

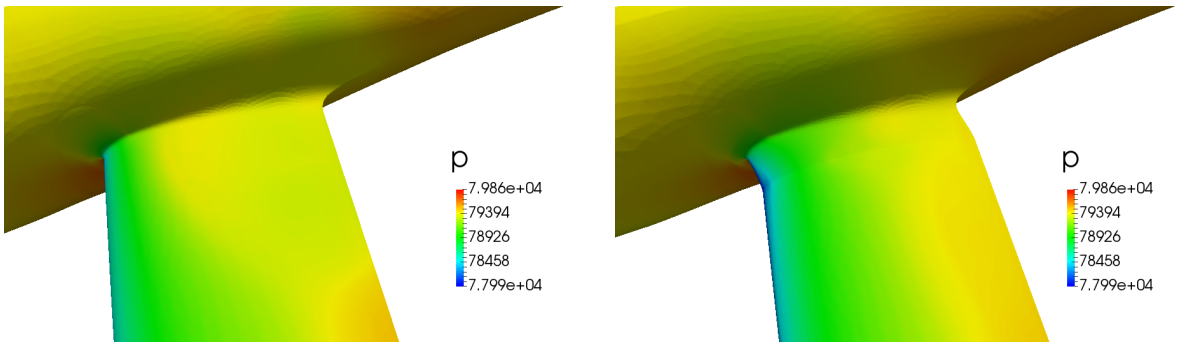


Figure 4: Optimization of an ultra–light aircraft. Comparison of the reference (left) and optimized (right) aircraft geometries. Upper view.

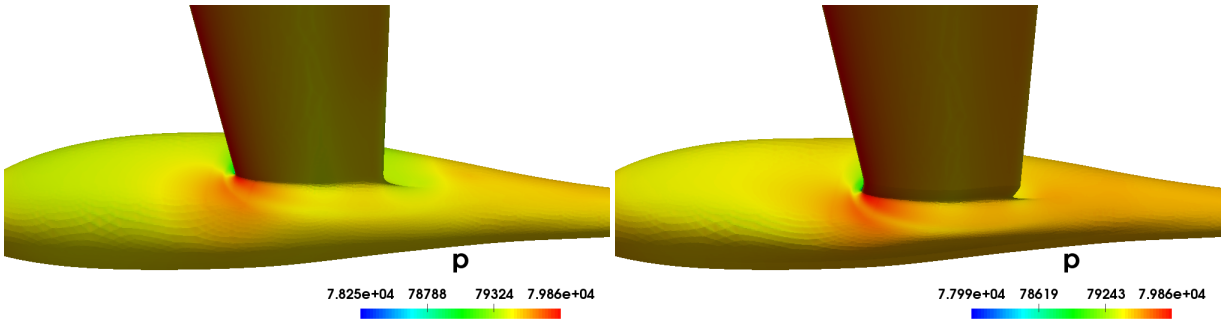


Figure 5: Optimization of an ultra–light aircraft. Comparison of the reference (left) and optimized (right) aircraft geometries. Bottom view.

4 OPTIMIZATION OF THE DRIVAER CAR MODEL

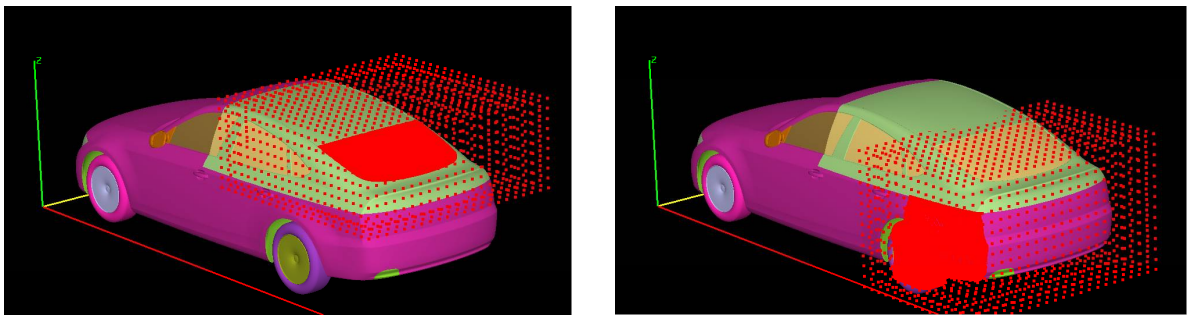
The second case is dealing with the drag minimization of a specific configuration of the DrivAer car model, a generic car model developed at the Institute of Aerodynamics and Fluid Mechanics of TU München, [16], to facilitate aerodynamic investigations of passenger vehicles. Herein, the fast–back configuration with a smooth underbody, mirrors and stationary wheels is used [17]. For this configuration, the generated unstructured mesh consists of about 3.8M cells.

Six design variables related the shape deformation of the mirror, the rear window, the front and back underbody, the car's distance from the road and the boat tail are used. The RBF set–up for two out of the six design variables is shown in figure 6. The evaluation tool is the steady state solver of OpenFOAM with the Spalart–Allmaras turbulence model with wall functions. In some candidate solutions, for which the flow becomes unsteady, the objective function value results from averaging results of the 100 last iterations.

The optimization starts with 20 samples, resulting from a randomized design which are then used to train the initial RSM. A $(25, 50)$EA with 500 evaluations on the RSM is used as described before. A total of 10 optimization cycles was needed to meet the convergence criteria. The convergence history is depicted in figure 7.

After 30 high-fidelity evaluations (including those in the DoE phase), the best solution yields 7% reduction in mean drag. Figures 8 and 9 compare the reference and the optimized car geometries. In all figures, the pressure over the DrivAer model surface is plotted.

Since, the shape optimization of this car model with the same design variables has previously been presented by the same team in [17], using a gradient–based optimization algorithm and the continuous adjoint method for the computation of the objective function gradient, it is interest-



(a) Rear window. (b) Boat tail.

Figure 6: DrivAer car shape optimization. Preview of the RBF source points, before applying morphing, for two out of the six design variables.

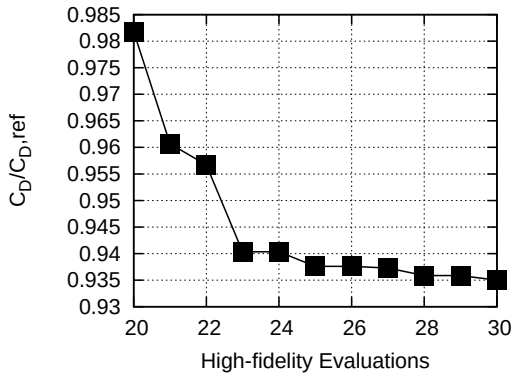
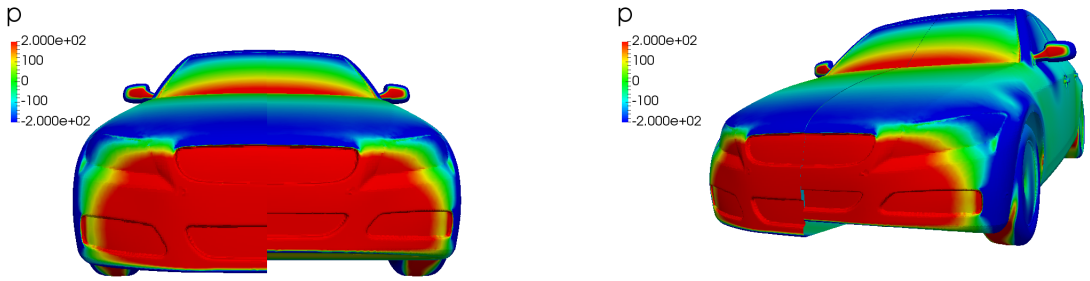


Figure 7: DrivAer car shape optimization. Convergence history of the optimization.



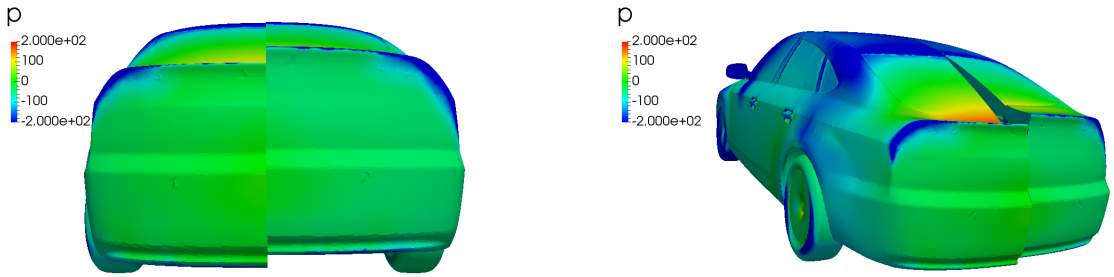
(a) Front view. (b) Side front view.

Figure 8: DrivAer car shape optimization. Comparison of the reference (left) and the optimized (right) car geometries.

ing to compare the resulting optimal solutions. Both methods yield almost similar results at the same computational cost and reduce the objective function value by about 7%. The resulting optimal geometries are similar and small differences can be identified in the front bumper and the spoiler, see figure 10.

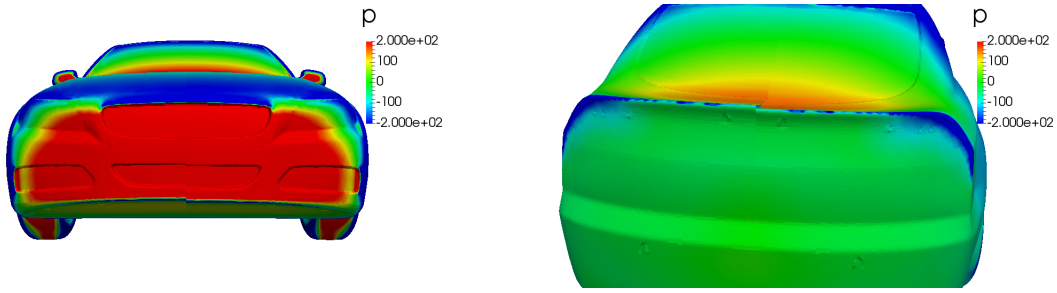
5 CONCLUSIONS

This paper presented EA–based features of the RBF4AERO optimization platform along with real–world applications. The gradient–based features are presented in a companion paper, [18]. The use of RSM as a surrogate evaluation model during the EA–based search allows for



(a) Rear view. (b) Side rear view.

Figure 9: DrivAer car shape optimization. Comparison of the reference (right) and the optimized (left) car geometries.



(a) Front view. EA (left), adjoint (right) (b) Rear view. Adjoint (left), EA (right).

Figure 10: DrivAer car shape optimization. Comparison of optimized geometries resulted from the EA assisted by the RSM and a gradient–based optimization in conjunction with the adjoint method, as presented by the same group in [17].

reduced optimization time. Over and above, the RBF–based mesh morphing reduces further the wall clock time per evaluation to be performed on the high–fidelity (CFD) tool. The overall optimization algorithm described is fully automated and quite user–friendly thanks to the graphic user interface of the RBF4AERO platform.

ACKNOWLEDGMENT

This work was financially supported by the RBF4AERO Project, funded by the European Union 7th Framework Programme (FP7-AAT, 2007 - 2013) under Grant Agreement no. 605396.

REFERENCES

[1] Y. Jin, M. Olhofer, and B. Sendhoff. A framework for evolutionary optimization with approximate fitness functions. *IEEE Transactions on Evolutionary Computation*, 6(5):481–494, 2002.

[2] D. Büche, N. Schraudolph, and P. Koumoutsakos. Accelerating evolutionary algorithms with Gaussian process fitness function models. *IEEE Transactions on Systems, Man, and Cybernetics*, 35(2):183–194, 2005.

[3] M.K. Karakasis, A.P. Giotis, and K.C. Giannakoglou. Inexact information aided, low-cost, distributed genetic algorithms for aerodynamic shape optimization. *International Journal for Numerical Methods in Fluids*, 43(10-11):1149–1166, 2003.

[4] M. Emmerich, K.C. Giannakoglou, and B. Naujoks. Single- and multi-objective evolutionary optimization assisted by Gaussian random field metamodels. *IEEE Transactions on Evolutionary Computation*, 10(4):421–439, 2006.

[5] M.K. Karakasis and K.C. Giannakoglou. On the use of metamodel-assisted, multi-objective evolutionary algorithms. *Engineering Optimization*, 38(8):941–957, 2006.

[6] J. Branke and C. Schmidt. Faster convergence by means of fitness estimation. *Soft Computing*, 9(1):13–20, 2005.

[7] R.H. Myers and D.C. Montgomery. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 2nd Ed.* New York, 2002.

[8] M.E. Biancolini, I. Viola, and M. Riotte. Sails trim optimisation using CFD and RBF mesh morphing. *Computers & Fluids*, (36):1119–1136, 2007.

[9] S. Jakobsson and O. Amoignon. Mesh deformation using radial basis functions for gradient–based aerodynamic shape optimization. *Computers & Fluids*, (36):1119–1136, 2007.

[10] M.J. Martin, E. Andres, C. Lozano, and E. Valero. Volumetric b-splines shape parametrization for aerodynamic shape design. *Aerospace Science and Technology*, (37):26–36, 2014.

[11] M.E. Biancolini. Mesh morphing and smoothing by means of radial basis functions (RBF): A practical example using Fluent and RBF Morph. *Handbook of Research on Computational Science and Engineering: Theory and Practice (vol. 2)*, pages 347–380, 2012.

[12] V.V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

[13] D.C. Montgomery. *Design and Analysis of Experiments, 5th Ed.* John Willey & Sons, Inc., New York, 2008.

[14] J. F. Kenney and E. S. Keeping. Linear regression and correlation. In *Mathematics of Statistics Pt. 1 3rd ed.*, chapter 15, pages 252–285. Princeton, NJ: Van Nostrand, 1962.

[15] M.E. Biancolini, E. Costa, U. Cella, C. Groth, G. Veble, and M. Andrejasic. Glider fuselage-wing junction optimization using CFD and RBF mesh morphing. *Aircraft Engineering and Aerospace Technology (to appear), DOI: 10.1108/AEAT-12-2014-0211.R1*.

[16] A. Heft, T. Indinger, and N. Adams. Experimental and numerical investigation of the DrivAer model. In *ASME 2012, Symposium on Issues and Perspectives in Automotive Flows*, pages 41–51, Puerto Rico, USA, July 8-12 2012.

[17] E.M. Papoutsis-Kiachagias, S. Porziani, C. Groth, M.E. Biancolini, E. Costa, and K.C Giannakoglou. Aerodynamic optimization of car shapes using the continuous adjoint method and an rbf morpher. In *EUROGEN 2015*, Glasgow, UK, September 14-16 2015.

[18] E. M. Papoutsis-Kiachagias, M. Andrejašič, S. Porziani, C. Groth, D. Eržen, M.E Biancolini, E. Costa, and K.C. Giannakoglou. Combining an RBF-based morpher with continuous adjoint for low-speed aeronautical optimization applications. In *ECCOMAS*, Crete, Greece, June 5-10 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THE RBF4AERO BENCHMARK TECHNOLOGY PLATFORM

Massimo Bernaschi[1], Alessandro Sabellico[1], Giorgio Urso[1], Emiliano Costa[2], Stefano Porziani[2], Fabrizio Lagasco[2], Corrado Groth[3], Ubaldo Cella[3], Marco Evangelos Biancolini[3], Dimitrios H. Kapsoulis[4], Varvara G. Asouti[4], Kyriakos C. Giannakoglou[4]

[1]National Research Council of Italy - Institute of Applied Computing
via dei Taurini, 19, 00185, Rome
{m.bernaschi, a.sabellico, g.urso}@iac.cnr.it

[2] D'Appolonia S.p.A.
Viale C. Pavese, 305, 00144, Rome
{emiliano.costa, stefano.porziani, fabrizio.lagasco}@dappolonia.it

[3] University of Rome Tor Vergata
Dept. of Enterprise Engineering
Via Politecnico 1, 00133, Rome
corrado.groth@uniroma2.it, ubaldo.cella@designmethods.aero, biancolini@ing.uniroma2.it

[4] National Technical University of Athens, School of Mech. Eng.
Parallel CFD & Optimization Unit
Athens, Greece
jim.kapsoulis@gmail.com, vasouti@mail.ntua.gr, kgianna@central.ntua.gr

Keywords: Optimization, Evolutionary Algorithms, Adjoint Coupling, Mesh Morphing, Radial Basis Functions, Benchmark Technology, Multiphisics

Abstract. *This paper presents the RBF4AERO benchmark technology platform, developed in the framework of the EU-funded RBF4AERO project. The platform enables the so-called Benchmark Management System (BMS) used for benchmark submission and results reporting. The BMS is deployed using three modules, namely the Graphical User Interface (GUI), the Workflow Manager (WM) and the Benchmarking Database System (BDS) which cooperate during the whole optimization benchmark life-cycle. The GUI is the only component which interacts with the end-user. It enables the optimization benchmark submission, along with the progress, results and computational platform resources monitoring. The configuration of the Optimization (OT) and the Morpher Tool (MT) is a pre-requisite for the optimization benchmark submission. In an optimization scenario the WM, which is practically the controller of the system, queries the OT in order to get a table of samples and gives back the results of the simulator (for instance a CFD tool). The evaluated individuals serve as training patterns of a Response Surface Model (RSM) which is, then, used for an Evolutionary Algorithms based optimization. The resulting 'optimal' solution(s) are delivered back to the WM for re-evaluation on the CFD tool. For each evaluation on the CFD tool, when a new geometrical shape is required, the computational grid is morphed using the MT based on radial basis functions.*

1 Introduction

The final goal of the RBF4AERO Project `http://www.rbf4aero.eu` is the development of a Benchmark Technology, namely a dedicated numerical platform and strategy capable to allow aeronautical design engineers to build up the novel optimization environment by using their own numerical models and computing platforms, and achieve the results of multi-objective and multi-disciplinary optimization studies in shorter time with respect to current practices. Besides, the RBF4AERO numerical platform enables to solve other relevant aircraft design studies such as fluid-structure interaction (FSI) and icing growth, and proposes a challenging CFD optimization technique that foresees the adjoint-morphing coupling. The basic idea of the optimization strategy is to make the CFD model parametric through an innovative shape optimization environment based on a high performance meshless morphing technique. The technique is founded on Radial Basis Functions (RBF) approach which offers a number of distinct advantages over the more traditional optimization approaches with no need to face with typical limiting trade-off constraints of speed, accuracy and extent, here the speed is related to the time required by the overall optimization process, the accuracy is achieved using large CFD meshes, and the extent is related to the number of different configurations fully calculated during the optimization process (see Figure 1).

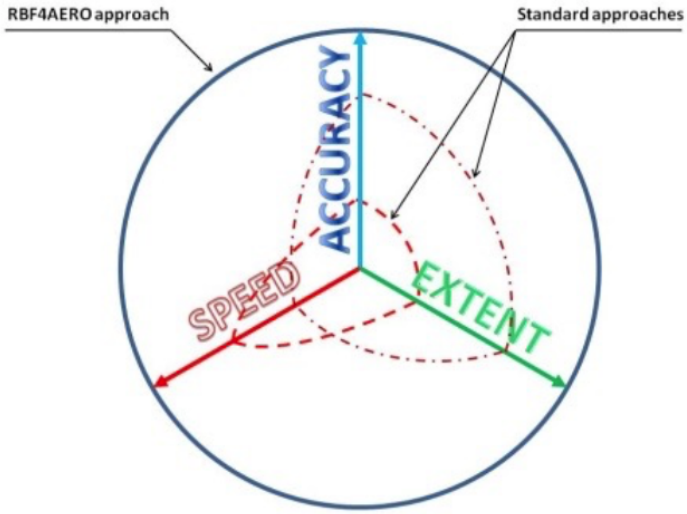


Figure 1: Trade-off between accuracy, time of computing, and number of analysed variants using traditional and the RBF4AERO proposed approach.

2 Description of Work

The scientific and technological enhancements offered by the RBF4AERO techniques require a convincing verification process before entering the industrial practice. For this reason, the Project work plan develops a robust industrial based process divided into three principal tasks:

1. Benchmark Technology Infrastructure Development
   The novel methodological procedure for the computational-driven optimisation proposed in the Project implies the development of an appropriate infrastructure to set-up the op-

timisation environment and enable the simulation of test cases of industrial relevance. A major role in achieving this goal will be played by parallel processing;

2. Benchmark Technology Verification
   Preliminary verification of results accuracy will be achieved on published state-of-the-art reference applications or available industrial cases;

3. Benchmark Technology Numerical Testing
   The optimisation procedure on demanding industrial applications will be carried out and numerically validated by the Project End Users.

Critical analysis of the numerical predictions of morphed configuration with respect to the baseline is supported and complemented by experimental outputs provided within the project. Finally, the effectiveness as well as the efficiency of the overall RBF4AERO optimization procedure is demonstrated.

3 Benchmark Technology Infrastructure capabilites

The architecture of the RBF4AERO numerical platform is thought as a control and communication manager framework where different modules are seamlessly integrated according to the available hardware and software. Looking at the platform as a whole, this integrated system is required to accomplish three principal functions: optimization, mesh morphing and simulation. The first two functions are carried out by software internal to the Benchmark Technology, whereas the latter one by external numerical solvers. An outline of the Benchmark Technology optimization process is depicted in Figure 2.

In specific, the optimization techniques implemented in the RBF4AERO platform are:

1. optimization algorithms based on Evolutionary Algorithms (EAs);

2. optimization based on a gradient algorithm and the continuous adjoint solver;

3. optimization based on Adjoint Self Sculpting.

The EA-based optimization is assisted by metamodels trained on data collected during the Design-of-Experiment (DoE) phase and supported by the Response Surface Models (RSM) that reduces the number of evaluation tool calls [1]. In the case the CFD solver is OpenFOAM and the objective function is the drag, lift or the pressure loss, the user can also exploit the capabilities of the continuous adjoint solver and of the innovative adjoint-morphing coupling. In particular, two algorithms foreseeing the coupling between the adjoint solver and the MT, called gradient-based [2] and Adjoint Self Sculpting algorithms, can be used to perform shape optimization. Moreover the Adjoint Preview feature, in the case multiple shape variations are available, can be adopted to identify the most influent ones. Relating to FSI two methods, respectively referred to as mode-superposition and two-way, are available [3]. According to the mode-superposition method, the CFD model is made flexible imposing structural displacements through mesh morphing by the superposition of its natural modes provided either by an analytic method or by means of a FEM model. The two-way technique foresees instead the coupled use of CFD and structural FEM solvers and the exchange of data between them. Using both approaches the MT allows to apply the deformation to selected wet surfaces only by constraining the other rigid bodies. With regard to icing, two methodological approaches have been designed [4]. According to the first one, termed frozen or constrained, icing simulations are carried out by

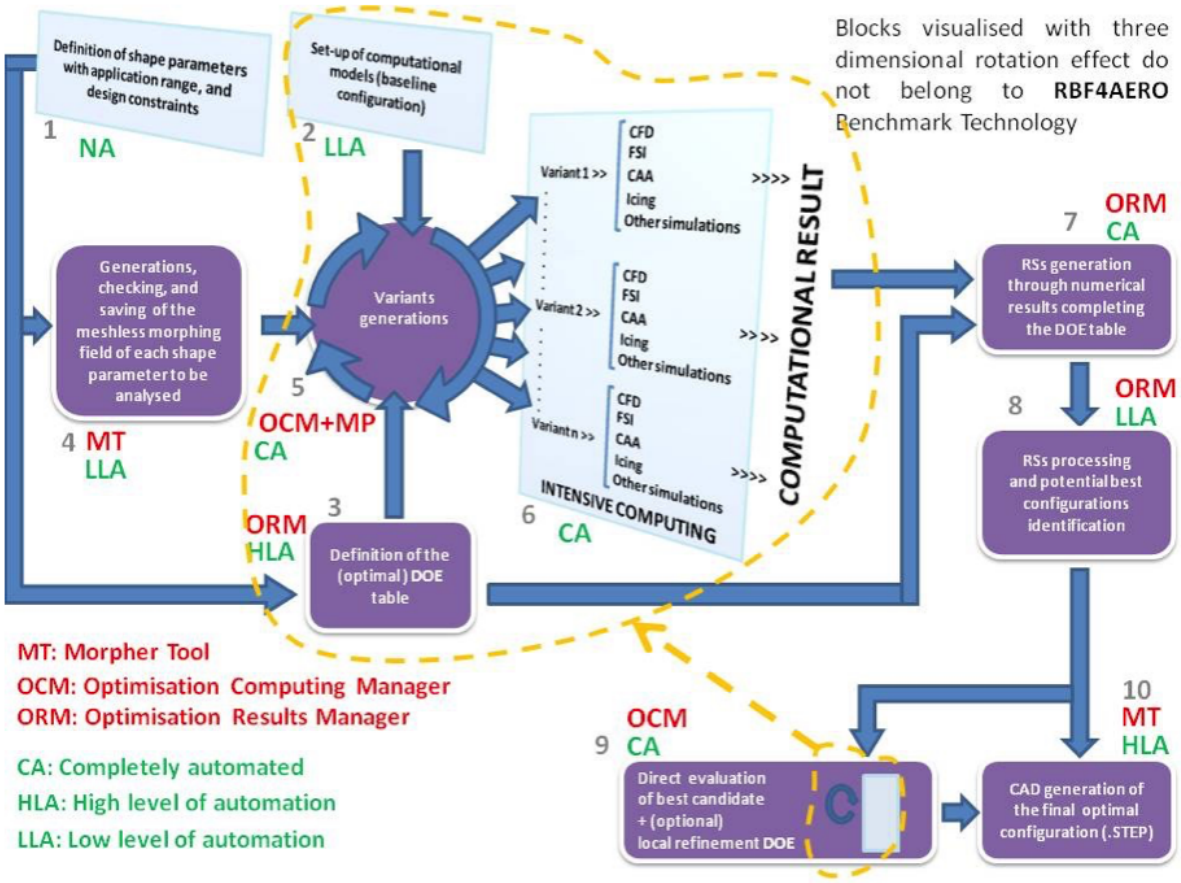


Figure 2: RBF4AERO Benchmark Technology optimization process.

imposing, at specific iterations of the CFD computing, the icing profiles previously calculated by means of an icing accretion tool at predefined instants of time. The second one, referred to as on-the-fly or evolutionary, foresees the use of an accretion code that, in conjunction with a CFD solver, modifies dynamically the numerical grid according to the calculated ice accretion.

4 Benchmark Management System

The **Benchmark Management System (BMS)** is used by the end-user for benchmark editing, benchmark submission, platform monitoring and results reporting. One of the main requirement of the software was to be as modular as possible, in order to be deployed in the largest possible number of system types and be able to manage the largest possible number of solvers. Three software modules have been developed to accomplish all the requirements:

- **Graphical User Interface (GUI)** give the end-user the possibility to interact with the platform by submitting user benchmarks to the BDS;
- **Workflow Manager (WM)** which executes the user submitted benchmarks by evenly distributing workload to platform nodes;
- **Benchmarking Database System (BDS)** which holds all information about user submitted benchmarks; it is the communication channel between GUI and WM.

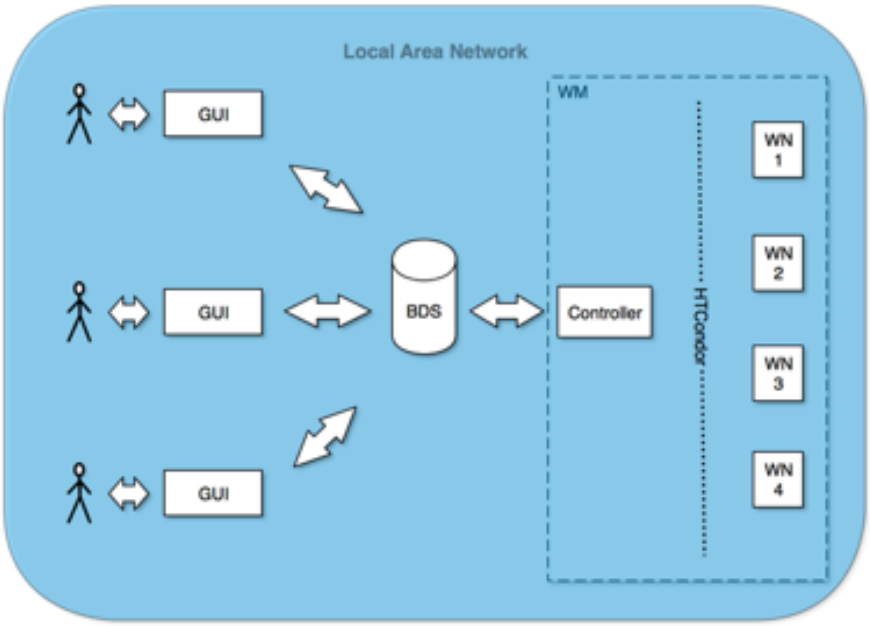


Figure 3: Benchmark Management System simplified scheme.

These three modules cooperate in order to assist the end-user during the whole optimization benchmark life-cycle.

4.1 Graphical User Interface

In order to guarantee compatibility on the majority of operating systems, GUI is based on the multi-platform Qt framework. The Graphical User interface is the only component which interacts with the end-user; it basically gives the possibility to:

- submit a generic benchmark
- show benchmark progresses and results
- monitor the available platform resources load

Benchmark submission is the procedure for the end-user to setup an optimization benchmark. The benchmark can be a shape optimization procedure based on the methods previously described, a multi-physics simulation or both. During this phase the user will input: the computational model prepared for a generic simulator/CFD, the RBF material needed during the morphing phase, all the optimization parameters needed for the specific benchmark type. Once the input is finished, the user can submit the benchmark which is then delivered to the Benchmarking Database System.

Benchmark visualization After the benchmark is submitted, it will be dispatched by the Workflow Manager. The user can follow all the progresses within the GUI, which is automatically updated while the information is inserted in the BDS. Benchmark visualization includes tabular views of data, various post processing graphs drawn using live data, platform logs for all messages and outputs/errors generated during the benchmark execution.

Resources monitoring For each platform computation node, the Workflow Manager is also responsible for updating node load information inside the BDS. This information is then received by the GUI which shows it to the end-user.

4.2 Workflow Manager

The workflow manager is the module which coordinates all the computing activities in the system. It queries the optimizer in order to get a table of candidate solutions, then represents the generated candidate solutions inside the BDS, executes every single candidate solution by scheduling jobs to be ran on the installed working nodes and gives back the results of the various jobs into the optimizer, writing the results also in the BDS as they can be read by the GUI and shown to the end-users. During jobs execution all the log information is stored by the WM on its log files, and all the important log messages produced by the running jobs are also stored in the BDS. When the user requests for these informations, the GUI will query the BDS and show these messages on a dedicated panel.

WM runs as a service on the hosting server and can be stopped and resumed at any moment, as all the benchmark and computed data are stored in the BDS and can be retrieved when the service is restarted. This means a lot of time saved when a running benchmark is stopped by the user or by the hosting machine malfunctioning.

Job scheduler

The **Job Scheduler** layer separates the WM component in two sub-layers: controller side and worker side. In the former, jobs are defined and delivered to a **Job Sender** that packs and entrusts them to the **Job Scheduler** layer; in the latter, executed on a potentially different system node, all the CFD/solver specific computation is done.

This way the controller node does not have to deal with node selection in job assignment, it has just to assure right dependencies between jobs. Before the assignment of a job to a particular node, the **Job Scheduler** checks if the node has access to all the solvers and resources needed in order to complete the job. Once the node has been selected, resources on that node will be allocated and the job will be prepared to be sent.

Finally, on the worker side, the workflow manager will receive the job from the **Job Scheduler** layer, will unpack the job and dispatch it to the right helper. When the job is completed, **Job Scheduler** will take care of copying the output results (a set of files on disk) to the controller node, which then parses them and updates the information in the **BDS**. Also, on a benchmark type basis, some important folders in the job working directory will be saved and referenced as a *Resource* inside the **BDS**. The end-user can request these *Resources* from a dedicated panel in the **GUI**. In our implementation we used the open-source workload management system **HTCondor**[1], which handles all the job scheduling operations during node selection and job sending.

[1] HTCondor software website: `https://research.cs.wisc.edu/htcondor/index.html`

Worker-side helpers

Each of the platform nodes communicates with the Controller node only through HTCondor communication protocols and can have any number and type of simulators/CFDs installed. The only constraint is that for each simulator, an Helper must be defined and installed. An Helper is a Python script with a predefined command line arguments set, defined in the Workflow Manager specifications. As the Python interpreter and HTCondor software are implemented in various operating systems, an RBF4AERO platform installation can have an heterogeneous set of worker nodes (WN) and can accomodate a huge number of simulators.

4.3 Benchmarking Database System

The Benchmarking Database System (BDS) is a relational database defined in a PostgreSQL DBMS installation. This module is the glue between the GUI and the WM, as they only communicate through queries to the DBMS. We have chosen this particular DBMS because it natively implements some triggering procedures which can notify tables update to both, GUI and Workflow Manager sides. The fact that all the information about benchmark management is stored inside BDS will also allow future exploitation of the platform using a web app. This will enable end-users to access RBF4AERO platform also outside the LAN where the WM and BDS are installed.

5 Conclusions

In this paper, an overview of the RBF4AERO benchmark technology platform and its working logic was presented. The complex modular structure that compose it was designed in order to be deployed in heterogeneous architectures and to be as much customisable as possible. Multi-objectives and multi-physics optimisation studies can be accomplished in a dramatically shorter times with respect to current practices.

6 Acknowledgements

This work was financially supported by the RBF4AERO project funded in part by the EUs 7th Framework Programme (FP7-AAT, 2007 - 2013) under Grant Agreement no. 605396.

REFERENCES

[1] D.H. Kapsoulis, V.G. Asouti, K.C. Giannakoglou, S. Porziani, E. Costa, C. Groth, U. Cella and M.E. Biancolini *Evolutionary Aerodynamic Shape Optimization through the RBF4AERO Platform*. ECCOMAS Congress 2016 VII European Congress on Computational Methods in Applied Sciences and Engineering, Crete Island, Greece, 510 June 2016.

[2] E.M. Papoutsis-Kiachagias, M. Andrejašič, S. Porziani, C. Groth, D. Eržen, M.E. Biancolini, E. Costa, and K.C. Giannakoglou *Combining an RBF-based Morpher with Continuous Adjoint for Low-Speed Aeronautical Optimization Applications*, ECCOMAS Congress 2016 VII European Congress on Computational Methods in Applied Sciences and Engineering, Crete Island, Greece, 510 June 2016.

[3] M. Andrejašič, D. Eržen, E. Costa, S. Porziani, M. E. Biancolini and C. Groth *A mesh morphing based FSI method used in aeronautical optimization applications*. ECCOMAS

Congress 2016 VII European Congress on Computational Methods in Applied Sciences and Engineering, Crete Island, Greece, 510 June 2016.

[4] M.E. Biancolini, E. Costa, C. Groth, G. Travostino and G. D'Agostini *Reliable mesh morphing approach to handle icing simulations in the aviation sector*. Aircraft Engineering and Aerospace Technology (submitted).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DATA ASSIMILATION USING MRI DATA

Geir Nævdal[1], **Ove Sævareid**[1], **and Rolf J. Lorentzen**[1]

[1]International Research Institute of Stavanger (IRIS)
Thormøhlensgate 55, N-5008 Bergen, Norway
e-mail: {gen,osae,rjl}@iris.no

Keywords: Contrast enhanced dynamic MRI data, multi-compartment porous media flow modelling, data assimilation, ensemble Kalman filter

Abstract. *Contrast enhanced dynamic MRI data is important for classifying anomalies in tissue and blood circulation, e.g. tumors, stokes, kidney failure. The resulting sequence of images providing spatial-temporal representation of contrast agent concentration must be given meaningful clinical interpretation. Current processing typically relies on localized strategies, where the images are analyzed voxel by voxel. The consensus seems to be that taking direct advantage of the global structure of the images potentially will allow for more accurate and robust interpretations.*

Based on a workflow that has been extensively studied within both petroleum and ground water research, the current paper combines explicit porous media flow modelling and state of the art data assimilation. Starting from a multi-compartment global flow model, we explore MRI interpretation via synthetic cases where model generated images are used as input to a parameter identification algorithm based on ensemble based data assimilation like the ensemble Kalman filter (EnKF).

From reasonable assumptions, the procedure demonstrates ability to identify relevant model parameters for two synthetic cases.

This study indicates that flow based assimilation of MRI data is viable and that it might provide valuable supplementing insight compared to state of the art. Regularizations are made on a firm statistical basis, eliminating the need for ad hoc parameters, and leading to estimates that are rigorous in terms of error thresholds. A priori knowledge can easily be incorporated in the procedure. The flow model allows explicit representation of basic material properties, and adds regularization to the procedure by enforcing globally consistent flow fields.

1 INTRODUCTION

Magnetic resonance imaging (MRI) enhanced by an injected contrast agent is a useful tool for characterizing the blood flow in organs like brain or kidney. The procedure provides a high resolution spatial-temporal representation of *contrast agent concentration* (CAC) over the object considered, and the aim of the subsequent interpretation is to extract clinically useful parameters. Assessment of brain blood supply relies on concepts like *cerebral blood volume* (CBV), *cerebral blood flow* (CBF), and *mean transit time* (MTT) of the blood flow.

The traditional parameter identification approach, see [1] for a detailed review, relates the observed average CAC across a *volume of interest* (VOI) to the CAC of the incoming (arterial) blood stream by the convolution

$$C_{\text{VOI}}(t) = (C_{\text{arterial}} * k)(t) \quad , \tag{1}$$

where the response function $k(t)$ is the flow-scaled residual and can be determined by explicit deconvolution or alternatively from various ad-hoc procedures. Given $k(t)$, the following estimates are available

$$\text{CBF} = \frac{\max\left(k(t)\right)}{\rho_{\text{VOI}}} \ , \quad \text{MTT} = \frac{\int_0^\infty k(\tau)d\tau}{\max\left(k(t)\right)} \ , \quad \text{CBV} = \text{MTT} \cdot \text{CBF} \tag{2}$$

where ρ_{VOI} denotes the mean density of the tissue constituting the VOI.

The significance of CBF is to identify the amount of blood that is actually feeding the local tissue. Applying the above approach voxel-wise based on modern high resolution imaging has certain limitations. It can be difficult to isolate the relevant arterial influx, and the observed CAC for the voxel will typically represent a superposition of blood in transit in the arterial and/or venular system as well as blood relevant to the local CBF. Voxel-wise temporal deconvolution also neglects the potential useful spatial structure of the observations.

These challenges are addressed by [2], where a systematic approach is formulated in terms of discrete spatial-temporal conservation relations for the CAC. The procedure can readily be extended to so-called multi-compartment models, where the total flowing volume of each voxel is partitioned into a set of distinct but interacting compartments. Each compartment type connects globally throughout the voxel-lattice thus constituting a set of interacting flow networks. Connection coefficients governing the flow between voxels and exchange between compartments can then in principle be identified from the evolving CAC distribution.

The current work is directly inspired by Sourbron [2], and in particular his section IV-D where he formulates a two-compartment model for "blood flow and perfusion". While Sourbron identifies the flow pattern by solving for a set of abstract connection coefficients, our approach is based on explicit physical modelling and identification of basic physical properties. We strongly believe that maintaining complete physical consistency throughout our computations may contribute important regularization to the problem.

It is natural to start to investigate if one can identify the volumes for the different compartments within a voxel by combining the use of a physical model for the flow of the contrast agent through some live tissue with the information obtained from a high resolution spatial-temporal representation of the CAC obtained from MRI. A natural approach is to use data assimilation methodology developed for characterization of the properties of a porous medium (e.g. flow in an oil reservoir or a hydrological model) developed over the last decades. A common approach is to use ensemble based methods, like the ensemble Kalman filter (EnKF) [3] for solving such problems (see e.g. [4], [5]). Alternatively, the Adaptive Gaussian Mixture filter (AGM) [6]

may be used. This filter is a combination of the traditional EnKF and particle filters, and was developed to better represent multi-modal posterior distributions.

The ensemble based data assimilation approaches works efficiently with large scale models. The models are updated by utilizing the information obtained from an ensemble of forecast (model simulations) with the available measurements. The EnKF and its variants are suitable for updating large scale non-linear models and is applied in a range of geophysical sciences, besides reservoir models such as meteorology (e.g. [7]) and oceanography (see e.g. [8]). Ensemble based data assimilation has also been used within medicine. In [9] a synthetic study using the local ensemble transform Kalman filter for improved forecasting for glioblastoma utilizing MRI data. In the recent paper [10] it is proposed that the cancer biology community makes an effort to utilize ideas from weather forecasting to predict and evaluate tumor growth and treatment response.

2 FORWARD MODEL

We consider a two-compartment model, or - in the terminology of traditional porous media modelling - a dual porosity, dual permeability model. One porous network (index a) is considered to represent the arterial blood flow and the other (index v) models the flow in the venular system. The interaction between the two pore networks represent the capillary perfusion feeding the local tissue, and the motivation for the model formulation is that this approach allows for an accurate identification of CBF.

We consider a spatial-temporal domain with coordinates $(\vec{x}, t)$, and the model formulation is equally valid in one, two, or three spatial dimensions. Filtration velocities, $\vec{u}_i$ for the respective networks are related to pressures p_i by Darcy's law. We assume fluid of constant viscosity μ, and permeabilities K_i that are functions of spatial position. Neglecting gravity, yields the following version of the Darcy law

$$\vec{u}_i = -\mu^{-1} K_i(\vec{x}) \nabla p_i \quad , \quad i = a, v. \tag{3}$$

We make the assumption of incompressibility for both fluid and porous media. The flow will be driven via boundary conditions, and the interior source terms q_i will only represent exchange between the compartments according to

$$\nabla \cdot \vec{u}_i = q_i \quad , \quad i = a, v. \tag{4}$$

As customary for dual models, we relate the transfer terms to the pressure difference between the compartments. The parameter σ_{av} scales for arterial/venular interface area of a unit volume, and the product $K_{av} \cdot \sigma_{av}$ constitutes a dimensionless conductivity factor at each location. Thus the volumetric flux per unit volume medium can be written

$$q = -q_a = q_v = \mu^{-1} K_{av}(\vec{x}) \sigma_{av}(\vec{x})(p_a - p_v). \tag{5}$$

Substituting equation (3) into (4) and then use equation (5) to eliminate the source terms yields a coupled system of two equations where p_a and p_v are the only unknowns. Given a domain, fluid and media properties, and suitable boundary conditions, we solve for the pressures using a standard finite volume scheme on a regular lattice. Subsequently we substitute the pressures into (3) and (5) to find velocities and transfer rates respectively.

The CAC is modelled as passive (i.e. not influencing the fluid flow) scalar field that enters along inflow boundary into the arterial network, is drained trough the capillary bed into the

venular network, and eventually leaves through the outflow boundary. The fractions of voxel volume occupied by the arterial and venular networks respectively, are given by the porosities ϕ_a and ϕ_v. As for equation (4), the source terms are exclusively related to transfer between compartments and the associated CACs are denoted c_i^*

$$\phi_i(\vec{x})\frac{\partial c_i}{\partial t} + \nabla \cdot (c_i \vec{u}_i) = q_i c_i^* \quad , \quad i = a, v. \tag{6}$$

Assuming normal operating conditions, the transfer will be one-way only, directed from arterial into venular, i.e. $q \geq 0$. We assume the CAC of the fluid leaving the arterial system at a given space-time location equals the overall arterial CAC at that location. This means $c_a^* = c_a$ and by combining equations (4) and (6) for $i = a$ we have

$$\phi_a(\vec{x})\frac{\partial c_a}{\partial t} + \vec{u}_a \cdot \nabla c_a = 0. \tag{7}$$

Thus the CAC travels unchanged along the characteristics $\phi_a \vec{dx} = \vec{u}_a dt$. By tracking the curve backwards until it hits an inflow boundary and take the time delay (time of flight) into consideration, c_a for arbitrary space and time can immediately be deduced from the appropriate boundary condition.

In order to associate a time delay with the transfer between the arterial and venular system, we introduce a third porosity, $\phi_{av}(\vec{x})$, characterizing the capillary beds. This additional compartment is assumed to be "discrete" in the sense that inter-voxel connections are omitted, a simplification justified by comparing the typical voxel size with that of a capillary bed. The time delays for given location can then be expressed as ϕ_{av}/q.

For the venular compartment, the situation is slightly more involved as the local CAC c_v normally will be different from the entering CAC c_{V}^* due to flow in the venular network and we have

$$\phi_v(\vec{x})\frac{\partial c_v}{\partial t} + \vec{u}_v \cdot \nabla c_v = q\,(c_v^* - c_v). \tag{8}$$

Again we identify a family of characteristics $\phi_v \vec{dx} = \vec{u}_v dt$, but this time CAC varies along the curves according to the right hand side of equation (8). By combining the time delay for the arterial and capillary flow, we can identify c_v^* at any point in space-time. Then c_v can be found at an arbitrary location by first tracking the characteristic curve backwards in time until one hits an inflow boundary or reaches the initial time. Starting from the relevant boundary or initial value, equation (8) is integrated forward along the characteristic curve to obtain the appropriate value for c_v.

To conclude this section, we relate our model formulation to the clinical parameters CBV, CBF and MTT. The unit used for CBF in a clinical setting will typically be either [mL/min/100g] or [mL/min/100mL], where the first is multiplied with a relevant density ρ_{VOI} of unit [g/mL] to obtain the second. Thus, starting from our inter-compartment flux q of units [1/s], the conversion factor is either $60 \times 100/\rho_{\mathrm{VOI}}$ or simply 60×100. Similarly, the unit used for CBV is normally [mL/100g] or [mL/100mL] which can be obtained by multiplying to our dimensionless capillary bed porosity ϕ_{av} with conversion factors of $100/\rho_{\mathrm{VOI}}$ or 100 respectively.

For the brain, [11] estimates the CBV for grey matter to be in the range 1.8 - 3.4 [mL/100mL] and the corresponding CBF in the range 61 - 103 [mL/min/100mL]. Thus the corresponding MTT is between 1.0 and 3.3 seconds. Similar estimates for white matter give a CBV between 0.9 and 1.7 [mL/100mL] and a CBF between 9 and 37 [mL/min/100mL], leading to MTT between 1.4 and 11 seconds. For our model parameters q and ϕ_{av}, typical values will then be 0.014 [1/s] and 0.026 [-] for grey matter, and 0.0038 [1/s] and 0.013 [-] for white matter.

3 ENSEMBLE BASED METHODS FOR DATA ASSIMILATION

Data assimilation concerns the subject of combining model and observations to improve the forecasting applicability of the model. The prototype application is weather forecasting were current observations from all available sources are taken into account before a new weather forecast is prepared. In that application, the observations are primarily used to update the initial values of the system equations that need to be solved for generating a new forecast.

As mentioned in the introduction data assimilation is used within many disciplines besides meteorology and atmospheric sciences, including oceanography, hydrology (e.g. groundwater flow) and petroleum (in particular reservoir modelling). Opposed to the initial value that one faces in meteorology, in our case, we are facing uncertain parameters in describing the blood flow through an organ. This means that we are closer to the problems faced within groundwater and reservoir modelling. For these applications an ensemble based approach is most common.

The data assimilation problem can be formulated as a minimization problem for a quadratic cost function

$$J(\vec{s}) = (\vec{s} - \vec{s}_b)^T C^{-1} (\vec{s} - \vec{s}_b) + \sum_{i=0}^{n} (y_i - G_i(\vec{s}))^T R_i^{-1} (y_i - G_i(\vec{s})). \tag{9}$$

Here $\vec{s}_b$ is the initial state estimate, C is a covariance matrix for a background error, R_i are measurement error covariance matrices, G_i are observation matrices and i is a time index and n the number of time steps. In weather forecasting variational approaches is the most common approach, whereas in other applications ensemble based methods are very popular. A variational approach will require $\nabla J(\vec{s})$. The ensemble based approaches have an advantage in that they do not require the access to the code developed for solving the forward problem.

The ensemble based approaches, like the EnKF, are typically derived based on the Kalman filter [12]. The Kalman filter can be derived as a recursive solution of the least squares problem (9). It is motivated from electrical and mechanical engineering with the aim of solving the state estimation for a linear dynamic system described by a forward model

$$\vec{s}_n = F\vec{s}_{n-1} + \vec{\epsilon}_n, \tag{10}$$

and an observational model

$$\vec{y}_n = G\vec{s}_n + \vec{\eta}_n.$$

Here, $\vec{s}$ is the state vector of the system, $\vec{y}$ denote the observations of the system, $\vec{\epsilon}_n$ denotes the model noise, which is assumed to be zero mean normally distributed with covariance matrix Σ and $\vec{\eta}_n$ is the measurement noise that is normally zero mean distributed with covariance matrix C_D. The model and measurement noise is independent, and the both the model and the measurement noise at different time instances are also independent.

Kalman [12] gave the optimal solution, given observations $\vec{y}_{o,n}$ at time step n. Starting from an initial estimate $\hat{\vec{s}}_0$ with covariance matrix C_0 representing the initial uncertainty, the following recursion provides the solution:

$$\begin{aligned}
\vec{s}_n &= F\hat{\vec{s}}_{n-1}, \\
C_n &= F\hat{C}_{n-1}F^T + \Sigma, \\
\hat{\vec{s}}_n &= \vec{s}_n + K_n(\vec{y}_{o,n} - G\vec{s}_n), \\
K_n &= C_n G^T (GC_n G^T + C_D)^{-1}, \\
\hat{C}_n &= (C_n^{-1} + G^T C_D^{-1} G)^{-1}.
\end{aligned}$$

The first two equations above represent the effect of the forward model, which propagates the mean and the covariance matrix. The third equation is the update equation taking into account the observations. The matrix K_n is denoted the Kalman gain matrix. The last equation give the posterior covariance matrix $\hat{C}_n$ after taking the information form the measurement at time n into account.

The Kalman filter was derived for a linear system, but was soon generalized to non-linear systems by the extended Kalman filter. The extended Kalman filter works by performing a linearization of a non-linear system around the mean, and works well for many mildly non-linear system. However, if the non-linearities are to large it breaks down. Another issue is that the propagation of the covariance matrices is not feasible for large-scale system.

The ensemble Kalman filter was suggested by Evensen [13], and that has been found to work for large scale system, and handle non-linearity better than the extended Kalman filter. The linear forward model (10) is now replaced with the non-linear extension

$$\vec{s}_n = F(\vec{s}_{n-1}) + \vec{\epsilon}_n \tag{11}$$

where we have kept the model noise $\vec{\epsilon}_n \sim N(0, \Sigma)$. The observation equation is

$$y_n = G\vec{s}_n + \vec{\eta}_n$$

where the measurement noise is $\vec{\eta}_n \sim N(0, C_D)$.

The ensemble Kalman filter solution is obtained by using an ensemble of N model realizations $\vec{S}_n = [\vec{s}_{n,1} \dots \vec{s}_{n,N}]$ to obtain the statistics required to solve the Kalman filter equations. The initial ensemble is generated by sampling from a normal distribution. All the ensemble members are propagated through the forward model $\vec{s}_{n,i} = F(\vec{s}_{n-1,i}) + \epsilon_{n,i}$. Here n is the time index and i is the counting of the ensemble members. Forecasted values of the observations can be computed as $y_{n,i} = G\vec{s}_{n,i} + \vec{\eta}_{n,i}$. (The noise need to be added to each ensemble member to propagate the statistics correct.) The Kalman gain is now approximated by calculating the mean of the ensemble, $\bar{\vec{s}}_n = \frac{1}{N}\sum_{i=1}^{N} \vec{s}_{n,i}$. Formally, we can also calculate the covariance matrix $\bar{C}_n = \frac{1}{N-1}\sum_{i=1}^{N}(\vec{s}_{n,i} - \bar{\vec{s}}_n)(\vec{s}_{n,i} - \bar{\vec{s}}_n)^T$ and estimate the Kalman gain matrix $K_n = \bar{C}_n G^T (G\bar{C}_n G^T + C_D)^{-1}$. From this each of the ensemble members are updated as

$$\hat{\vec{s}}_{n,i} = \vec{s}_{n,i} + K_n(\vec{y}_{o,n} - \vec{y}_{n,i}). \tag{12}$$

What is presented here is a formal derivation of the Kalman filter update equation using the ensemble Kalman filter. Care needs to be taken in the implementation, and different approaches exist on how to handle this, see e.g. [3], [4], [9].

The AGM filter utilizes the same set of equations for updating of the state variables, but in addition weights are computed (and adaptively adjusted to avoid degeneracy) for each ensemble member. The weights are then used to resample the ensemble to better represent non-Gaussian behavior. For more details, see [6], [14].

In our application, the primary aim is to estimate parameters, $\vec{p}$. This can be done by extending the state vector $\vec{s}$ to the form

$$\vec{s}_e = \begin{bmatrix} \vec{s} \\ \vec{p} \end{bmatrix}.$$

In the forward step (11) the parameters are not changed, that is,

$$\vec{s}_{e,n} = F(\vec{s}_{e,n-1}) = \begin{bmatrix} F(\vec{s}_n) \\ \vec{p}_n \end{bmatrix}.$$

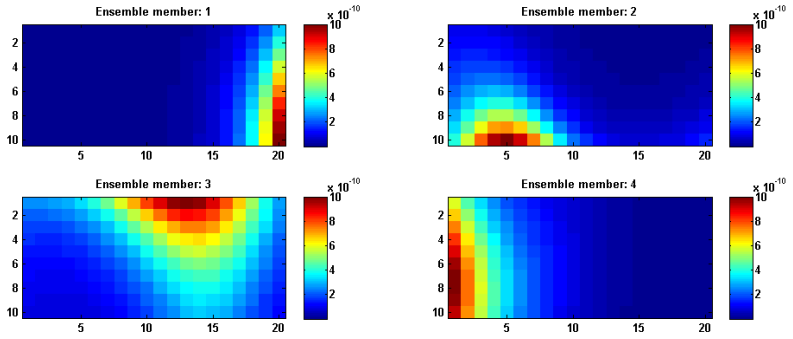


Figure 1: Four first ensemble members for permeability. The units are [m^{-2}].

However, in the update equation (12) we are now working on the extended state vector $\vec{s}_e$. As is common in reservoir engineering applications for parameter estimation we set the model noise $\vec{\epsilon}_n = 0$ in (11).

4 DETERMINING PARAMETERS IN A TWO-COMPARTMENT MODEL USING MRI DATA

We demonstrate the methodology on two examples of increasing complexity. In the first example the tissue perfusion (CBF) is estimated and in the second example the cerebral blood volume (CBV) is estimated. In both cases we consider a a square tissue field, with dimensions 10 cm $\times$ 10 cm, with resolution $n_x = 20$ and $n_y = 10$. The arterial inlet function is represented by a pulse ($C_{arterial} = 1$) lasting 4 seconds, and entering the field at the left boundary of the field. The pressure at the same boundary is 13300 Pa. At the venular compartment, the pressure boundary condition is 133 Pa at the left edge. All other edges are assumed to be sealed. The blood viscosity is $4 \cdot 10^{-3}$ Pa·s. The MRI data are simulated as the superposition of the compartment concentrations times the corresponding porosity:

$$m_{mri}(\vec{x}) = \phi_a \cdot c_a + \phi_{av} \cdot c_{av} + \phi_v \cdot c_v. \tag{13}$$

To these measurements we add independent (temporally and spatially) measurement noise of 10^{-4}.

In both the following examples the simulator is run for 60 seconds with time step of 0.6 seconds.

4.1 Estimating perfusion

In this example we assume that the permeabilities $K_i(\vec{x})$ defined in equation (3) are unknown, but we assume that the permeabilities are equal for the arterial and venular system, i.e. K_a = K_v. We further assume that the product $K_{av} \cdot \sigma_{av}$ defined in equation (5) is given by $10 \cdot K_i$. This gives one unknown parameter for each voxel in the simulated system. The ensemble members are generated as exponentials of a Gaussian randoms fields. The random fields have a spatial distribution having a Gaussian variogram [15, page 91] with practical range $\frac{10}{\sqrt{3}}$ in the x-direction and $\frac{20}{\sqrt{3}}$ in the y-direction. The fields are shifted so that the maximum value is 10^{-9}. Four realizations of the ensemble are shown on Figure 1. In this example 500 ensemble members are used. We have assumed that the porosity values for the arterial, capillary and venular systems are known. The applied fields are shown on Figure 2. The Adaptive Gaussian Mixture (AGM) filter is then applied for estimation of the true permeability field. The true field

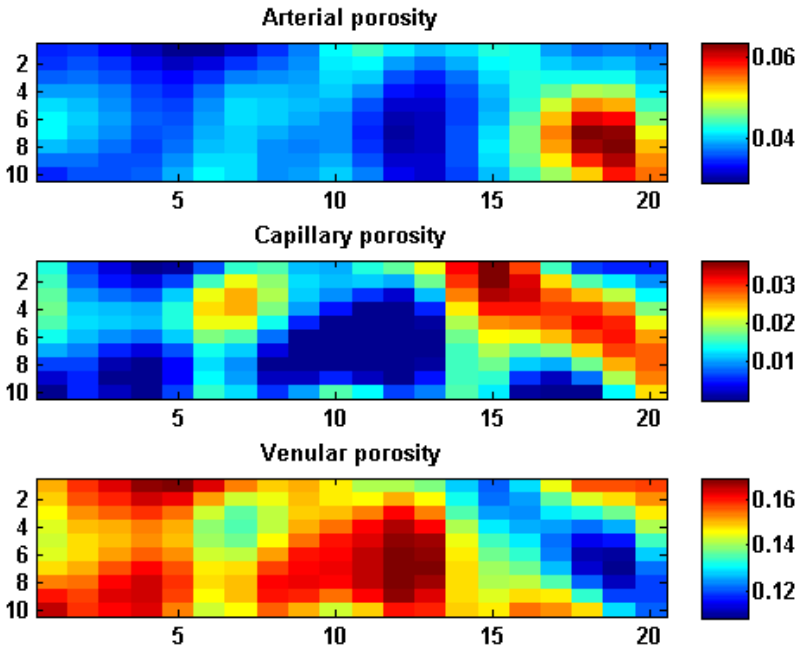


Figure 2: Porosity values for arterial, capillary and venular system.

is generated the same way as the initial ensemble. The field is updated at 24 points in time, evenly distributed within the simulation interval from 0 to 60 seconds. The true field is shown on Figure 3 (top) and the mean value of the final estimated permeability is shown on Figure 3 (bottom). As can be seen, the field is almost perfectly reconstructed. The corresponding true and estimated perfusion fields are shown on Figure 4. Finally, we show the simulated MRI data at selected points in time on Figure 5. Here we can clearly see how a rather inaccurate MRI image at the early time steps are improved as more data are assimilated.

4.2 Estimation of porosities for individual compartments

In this example we consider estimation of the porosity fields $\phi_a(\vec{x})$, $\phi_v(\vec{x})$ and $\phi_{av}(\vec{x})$ of the two-compartment model using MRI observations of the CAC as a function of time. The region we consider is the same as in the previous example, but now we have constant permeabilities $K_{av} = 2 \cdot 10^{-9} \mathrm{m}^{-2}$, $K_a = K_v = 2 \cdot 10^{-10} \mathrm{m}^{-2}$. For the porosity we generate three spatially varying fields. These fields are constructed by first generating two fields $u(\vec{x})$ and $v(\vec{x})$ a spatial distribution having a Gaussian variogram [15, page 91] with practical range $\frac{6}{\sqrt{3}}$ in the x-direction and $\frac{12}{\sqrt{3}}$ in the y-direction. Then $\phi_a(\vec{x}) = 0.003 \cdot u(\vec{x}) + 0.0375$ giving the field shown at Figure 6 (top). Then we set $\phi_v(\vec{x}) = 0.005 \cdot (-0.8 \cdot u(\vec{x}) + 0.6 \cdot v(\vec{x})) + 0.15$ and obtain the field shown in Figure 7 (top). Further we make the assumption that the $\phi_a u(\vec{x}) + \phi_v u(\vec{x}) + \phi_{av} u(\vec{x}) = 0.2$ and this defines ϕ_{av} as shown in Figure 8 (top). (Note that $\phi_a u(\vec{x})$ and $\phi_v u(\vec{x})$ are generated such that they have an expected correlation of -0.8, but for the realization giving our "true" fields the correlation is actually higher.)

While working with the EnKF we generate an initial ensemble of size 200 of fields ϕ_a, ϕ_v and ϕ_{av} using the same construction as for the true field, but with the additional ϕ_v is truncated such that $\phi_v + \phi_a < 0.199$. The condition $\phi_v + \phi_a + \phi_{av} = 0.2$ corresponds to having full knowledge of CBV from the CAC data. While generating the initial ensemble we added a constraint that

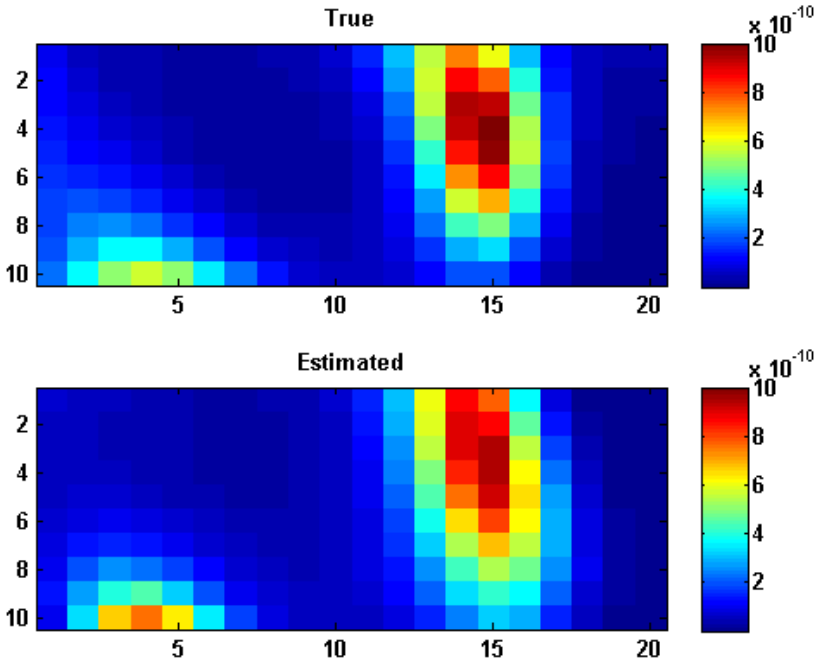


Figure 3: True permeability field (top) and mean value of estimated fields (bottom). The units are [m^{-2}].

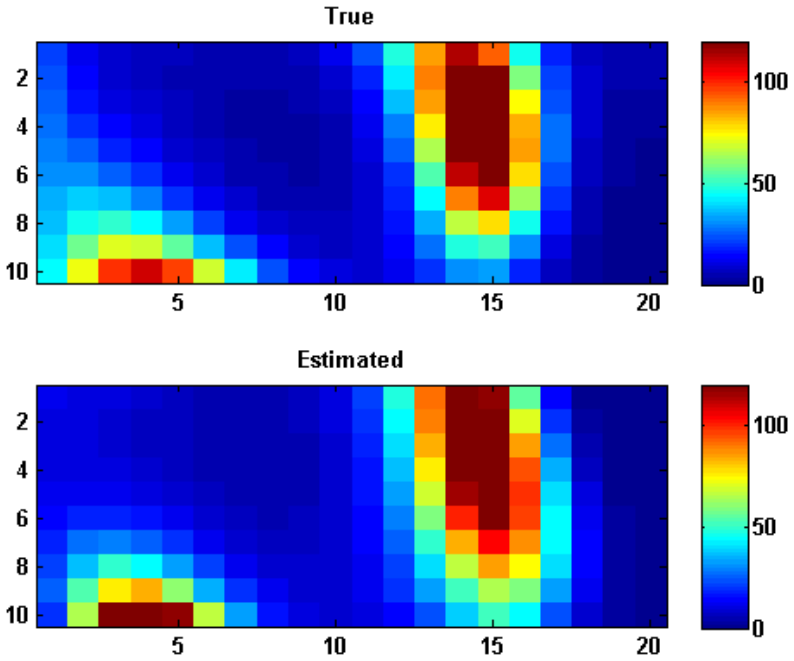


Figure 4: Estimated perfusion - CBF in units [mL/min/100mL].

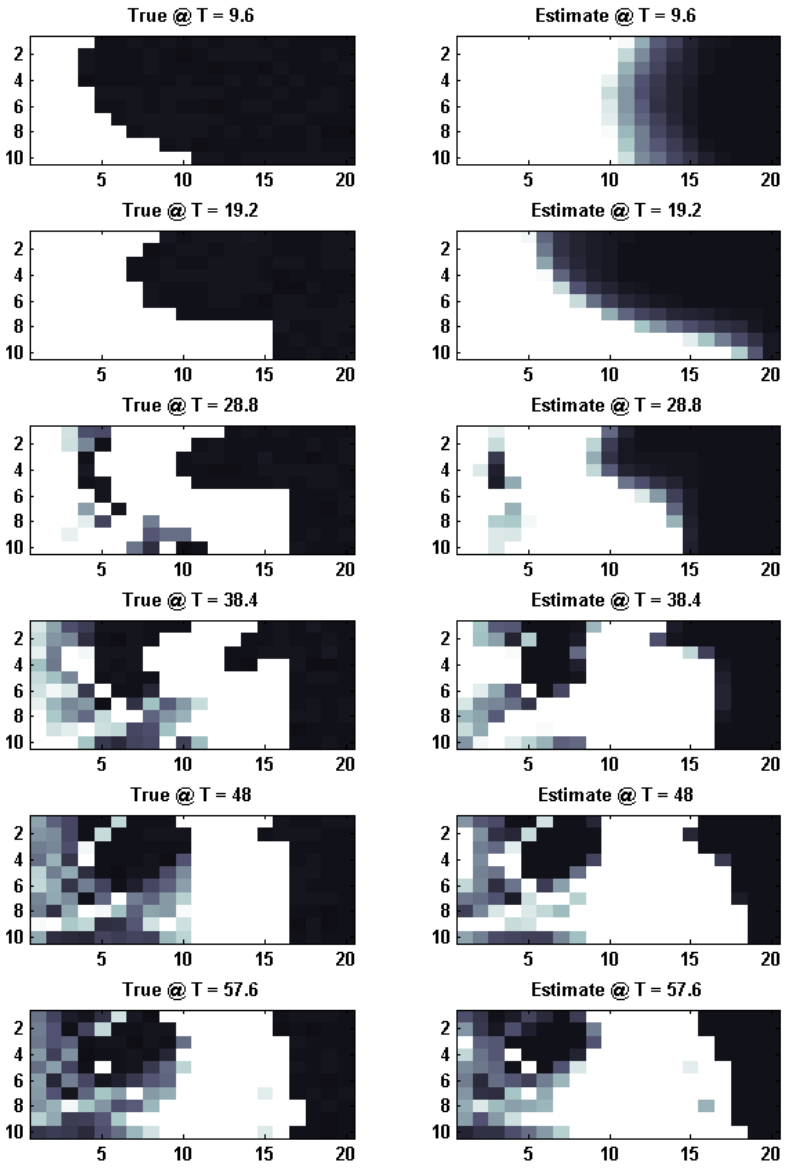


Figure 5: Simulated MRI data. Black indicate low intensity and white indicate high intensity. The range is between 0 and 10^{-2}.

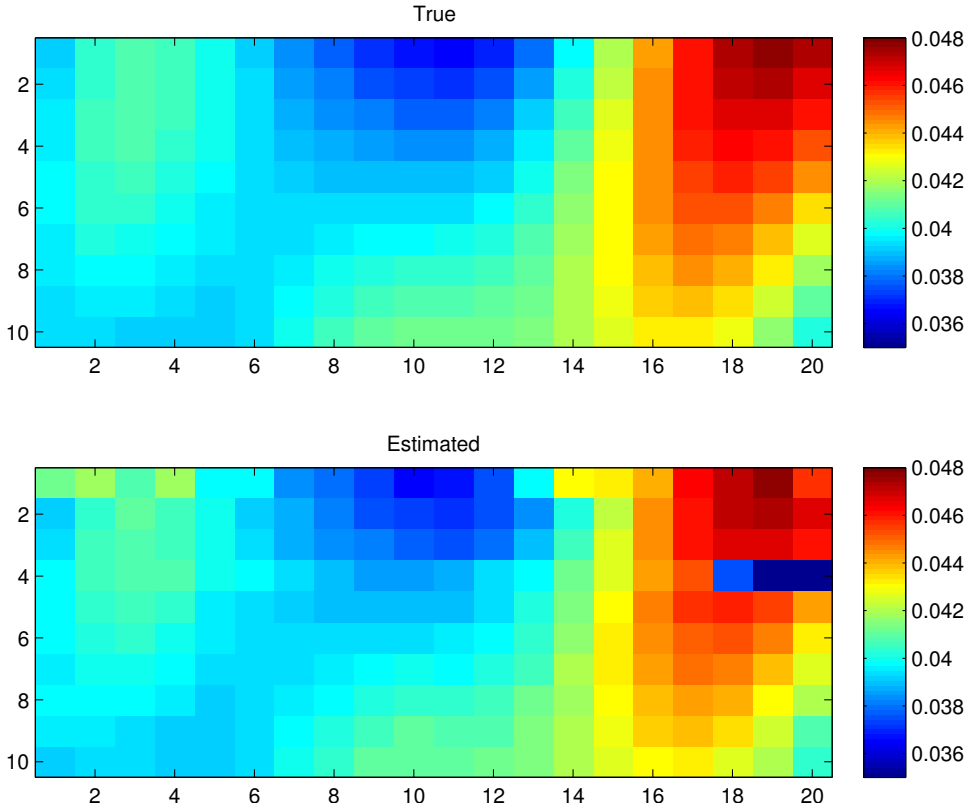


Figure 6: True and estimated ϕ_a.

all the porosity values always should be above 0.001. In the cases when the EnKF update step adjusted them below this value, they were truncated.

We observe the CAC for every 0.6 seconds for one minute with same type of measurement noise as in the previous example. This give a very large data set to assimilate. For the EnKF this poses additional challenges to avoid ensemble collapse. Therefore we restrict the use of each CAC observation to only update the porosities in its own location. Similar type of localization was used in [16] for updating a petroleum reservoir model using a seismic data set. The final estimates are shown in Figures 6 (bottom), 7 (bottom) and 8 (bottom). We note that the match all over is exceedingly good, except for a few locations. To avoid some extreme values to flatten the colors of the image, we have truncated the fields. For $\phi_v(\vec{x})$ there are two values above 0.16, the highest one 0.22. For $\phi_{av}(\vec{x})$ there are two locations having a value above 0.02, the highest one is 0.068. For four locations we came below 0.01, the lowest value was 0.001 (the truncation value). For $\phi_a(\vec{x})$ we got two values below 0.036, the lowest one was 0.001. The extreme values should be easily identified in the figures.

5 SUMMARY AND CONCLUSIONS

A two-compartment model is built with the aim of studying blood flow in live tissue. In a couple of synthetic examples we show that estimating important parameters as porosities of the compartments and the permeability seems feasible. In further studies, we need to generalize the estimation methodology to estimate these parameters jointly. In further studies there is also a clear need to proceed towards more realism in the biological modeling.

There are several issues that has been simplified in the modelling. The boundary condition is now assumed known, but in reality there will be uncertainty associated with it. Dispersion

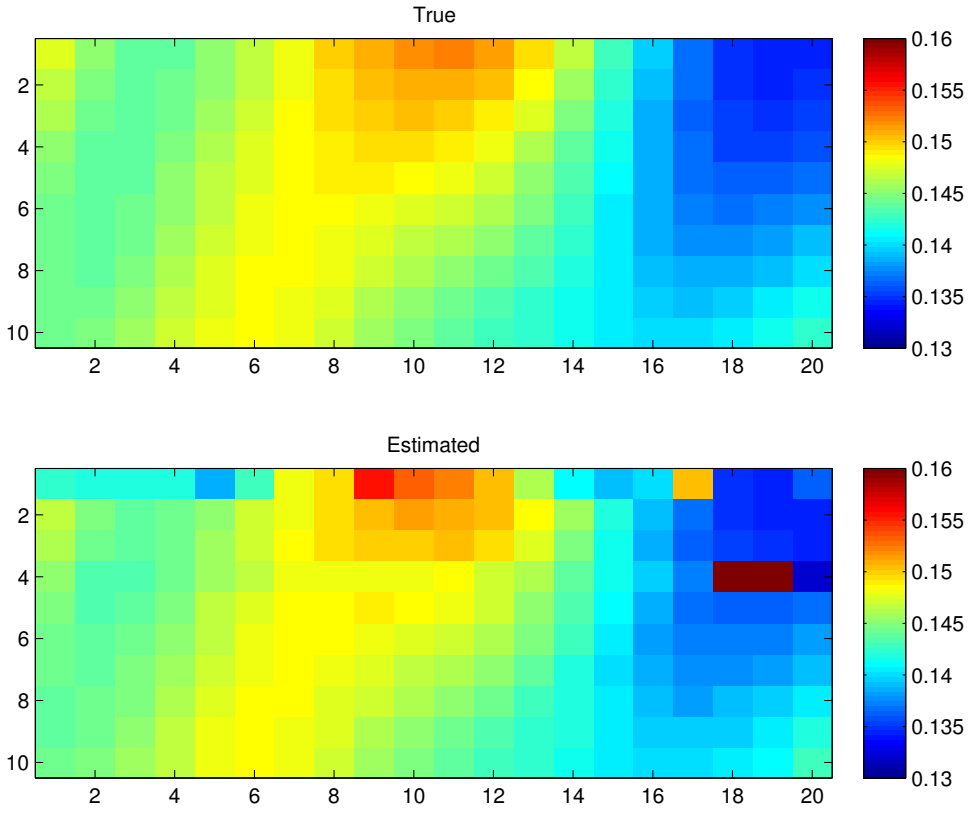


Figure 7: True and estimated ϕ_v.

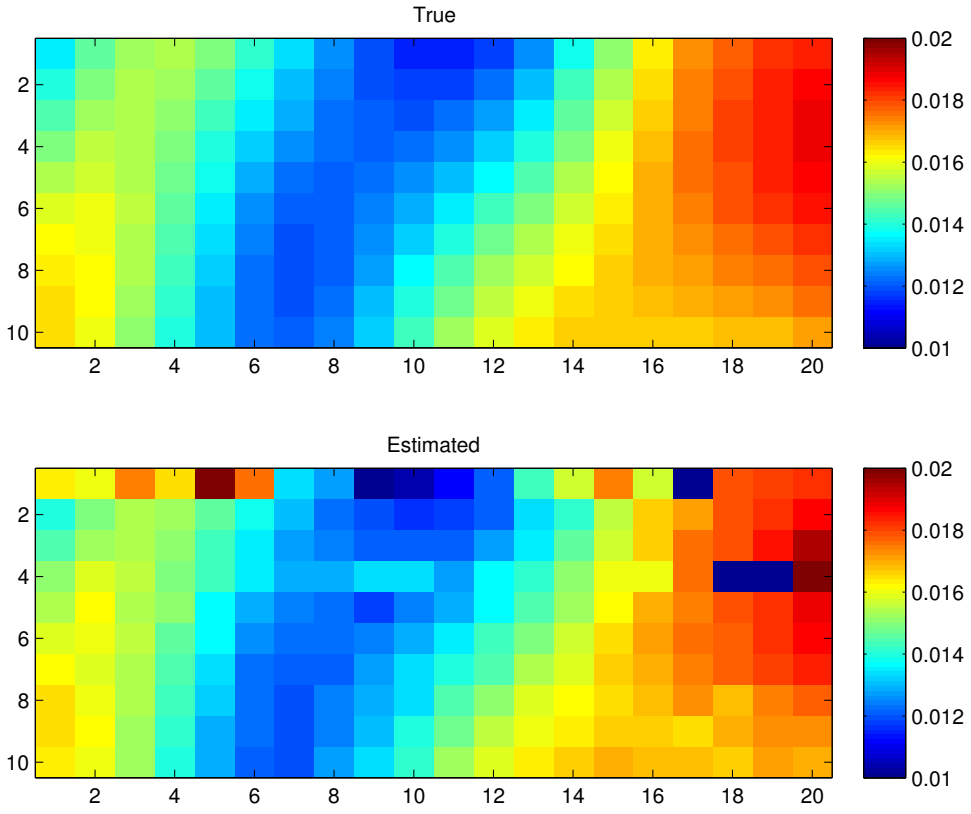


Figure 8: True and estimated ϕ_{av}.

effects are ignored, as well as the fact that the tissue will be in constant movement due to the heart pulse. However, while interpreting MRI images, any internal movement of the body has to be adjusted for to compare images at different time instances. Another interesting feature that might be added is the modelling is accounting for defects in the blood-brain barrier.

Both the EnKF and AGM that is used for data assimilation here, requires an initial ensemble based on a prior model. In the examples presented here, both the true model and the initial ensemble has been built based on the authors prior experience within geoscience. For modelling flow in live tissue, the prior assumptions need to be developed based on understanding of the biology of the actual system.

REFERENCES

[1] A. Fieselmann, M. Kowarschick, A. Ganguly, J. Hornegger, and R. Fahrig, Deconvolution-based CT and MR brain perfusion measurement: theoretical model revisited and practical implementation details. *International Journal of Biomedical Imaging*, **2011**, 20 pages, 2011.

[2] S. Sourbron, A tracer-kinetic field theory for medical imaging. *IEEE Transactions on Medical Imaging*, **33**, 4, 935–946, Apr. 2014.

[3] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2007.

[4] S. I. Aanonsen, G. Nævdal, D. S. Oliver, A. C. Reynolds, and B. Vallès, The ensemble Kalman filter in reservoir engineering – a review. *SPE Journal*, **14**, 3, 393–412, Sep. 2009.

[5] D. S. Oliver and Y. Chen, Recent progress on reservoir history matching: a review. *Computational Geosciences*, **15**, 185–221, 2010.

[6] A. S. Stordal, H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallès, Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. *Computational Geosciences*, **15**, 2, 293–305, 2011.

[7] M. Ehrendorfer, A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift*, **16**, 6, 795–818, Dec. 2007.

[8] F. Counillon, I. Bethke, N. Keenlyside, M. Bentsen, L. Bertino, and F. Zheng, Seasonal-to-decadal predictions with the ensemble Kalman filter and the Norwegian Earth System Model: a twin experiment. *Tellus A*, **66**, 21074, 2014.

[9] E. J. Kostelich, Y. Kuang, J. M. McDaniel, N. Z. Moore, N. L. Martirosyan, and M. C. Preul, Accurate state estimation from uncertain data and models: an application of data assimilation to mathematical models of human brain tumors. *Biology Direct*, **6**, 64, 2011.

[10] T. E. Yankeelov, V. Quaranta, K. J. Evans, and E. C. Rericha, Toward a science of tumor forecasting for clinical oncology. *Cancer Research*, **75**, 6, 918–923, Mar. 2015.

[11] S. Sourbron, M. Ingrisch, A. Siefert, M. Reiser, and K. Herrmann, Quantification of cerebral blood flow, cerebral blood volume, and blood-brain-barrier leakage with DCE-MRI. *Magnetic Resonance in Medicine*, **62**, 1, 205–217, 2009.

[12] R. E. Kalman, A new approach to linear filtering and prediction problems. *Transactions of the AMSE, Journal of Basic Engineering (Series D)*, **82**, 34–45, 1960.

[13] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, C5, 10, 143–10, 162, 1994.

[14] A. S. Stordal, R. Valestrand, H. A. Karlsen, G. Nævdal, and H. J. Skaug, Comparing the adaptive Gaussian mixture filter with the ensemble Kalman filter on synthetic reservoir models. *Computational Geosciences*, **16**, 2, 467–482, Mar. 2012.

[15] N. Remy, A. Boucher, and J. Wu, *Applied Geostatistics with SGeMS*. Cambridge University Press, 2009.

[16] Y. Dong, Y. Gu, and D. S. Oliver, Sequential assimilation of 4D seismic data for reservoir description using the ensemble Kalman filter. *Journal of Petroleum Science and Engineering*, **53**, 83–99, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

FULL-WAVEFORM INVERSION FOR THE RECONSTRUCTION OF ELECTROMAGNETIC PROPERTY PROFILES USING PLANE MICROWAVES

Namho Joh[1] **and Jun Won Kang**[1]

[1] Department of Civil Engineering, Hongik University
Wausan-ro 94, Mapo-gu, Seoul 04066, Republic of Korea
e-mail: johst99@naver.com, jwkang@hongik.ac.kr

Keywords: Perfectly-Matched-Layers (PMLs), Plane electromagnetic waves, Lagrangian, Full-waveform Inversion, Maxwell's equations.

Abstract. *This study is concerned with microwave full-waveform inversion method for the reconstruction of electromagnetic properties in heterogeneous one-dimensional media. Perfectly-Matched-Layers (PMLs) are introduced as wave-absorbing boundary to reduce the originally infinite medium to a finite computational domain. The microwave full-waveform inversion method is based on an optimization approach constrained by PML-endowed Maxwell's equations for plane electromagnetic waves. The approach attempts to minimize the objective functional consisting of the L^2 norm of a misfit between calculated and observed electric fields. Specifically, Lagrangian was constructed comprising the objective functional augmented by the PML-endowed Maxwell's equations via Lagrange multipliers. Enforcing stationarity of the Lagrangian yields time-dependent state, adjoint, and time-independent control problems, which can be iteratively solved to update the electromagnetic properties of the medium such as permittivity and permeability. We used Fletcher-Reeves conjugate gradient method for updating the electromagnetic property profiles, and explored Tikhonov and Total Variation regularization methods to relieve the ill-posedness inherent in the full-waveform inversion. Numerical results are reported showing the successful reconstruction of both smooth and sharply-varying electromagnetic property profiles with appropriate regularization methods.*

1 INTRODUCTION

In the last several decades, microwave nondestructive evaluation (NDE) of civil structures has been paid considerable attention by both academics and industry owing to the advantages such as low cost, good penetration in dielectric materials, good resolution, and contactless feature of sensors. The method has been found to be potentially excellent in detecting cracks, voids, flaws, defects, moisture contents, layers, etc. of composite structures by itself or with the combination with other NDE methods. Recent applications of the method includes, but not limited to, damage exploration of FRP sheets [1], quality control and condition assessment of reinforced concrete structures [2,3].

Microwave NDE methods generally use electromagnetic waves with frequencies between 300 MHz and 100 GHz to inspect and characterize structures and materials. For dielectric materials, microwaves may have wavelengths of orders of centimeters to millimeters, and it has been shown that cracks and layer thickness of heterogeneous dielectric structures could be measured using microwaves of such wavelengths [4,5]. However, microwave NDE methods have been more implemented in academic research environment, rather than for practical applications so far. One of the reasons for this is probably that the microwave NDE technique and the interpretation of test data are fairly complicated. Despite the difficulties and limited availability of commercial microwave NDE systems, microwave NDE methods have achieved great progress in the development of sophisticated inversion procedure and microwave imaging system for structure monitoring and biomedical applications. In particular, recent development of innovative inverse scattering methods with various optimization algorithms deserves close attention [6,7].

This work seeks the development of a microwave inverse scattering method for subsurface imaging of structures. The proposed inversion method is based on using full waveforms rather than using the waveform of specific frequencies. This kind of method is called full-waveform inversion method, and it has been shown to be effective in unveiling unknown material properties in seismic imaging and geophysical probing applications, etc. [8,9]

The target for reconstruction in this work is the spatial distribution of permittivity and permeability profiles of heterogeneous media. To this end, an optimization approach constrained by Maxwell's equations for plane electromagnetic waves was introduced. The approach attempts to minimize the objective functional consisting of the L^2 norm of a misfit between calculated and measured electric fields. Specifically, Lagrangian was constructed comprising the objective functional augmented by the Maxwell's equations. Enforcing stationarity of the Lagrangian yields time-dependent state, adjoint, and time-independent control problems, which can be solved iteratively to update the permittivity and permeability profiles of the medium. To satisfy far-field radiation condition of electromagnetic waves, Perfectly-Matched-Layers (PMLs) were introduced as wave-absorbing boundaries next to the finite computational domain. Formulation and numerical examples of plane microwave inversion are presented in the subsequent Chapters.

2 FORWARD PROBLEM

To conduct full-waveform inversion analysis, forward solutions of wave equations are indispensable since the solutions should be utilized in the overall inversion process. In this work, plane microwaves are treated as probing waves for the characterization of material profiles, such that the domain of wave propagation can be modelled as a one-dimensional one. Fig. 1 illustrates the one-dimensional domain truncated by PMLs on both sides. For the forward modelling of plane electromagnetic waves in the PML-truncated domain, modified Maxwell's equations discussed in [10-12] are introduced. Therefore, the forward problem to find electric and magnetic fields in the PML-truncated domain can be formulated as the following: Find $E_x = E_x(z,t)$ and $H_y = H_y(z,t)$, such that:

$$\mu \frac{\partial H_y}{\partial t} + c_{\text{ref}} g \mu H_y + \frac{\partial E_x}{\partial z} = 0 \tag{1a}$$

$$\epsilon \frac{\partial^2 E_x}{\partial t^2} + c_{\text{ref}} g \epsilon \frac{\partial E_x}{\partial t} + \frac{\partial^2 H_y}{\partial z \partial t} = -\frac{\partial J}{\partial t} - c_{\text{ref}} g J \tag{1b}$$

and

$$E_x(-L_{tl}, t) = E_x(L_{tr}, t) = 0 \tag{2a}$$

$$H_y(-L_{tl}, t) = H_y(L_{tr}, t) = 0 \tag{2b}$$

$$E_x(z, 0) = \frac{\partial E_x}{\partial t}(z, 0) = 0 \tag{2c}$$

$$H_y(z, 0) = 0 \tag{2d}$$

where $E_{\mathrm{x}}(z,t)$ and $H_{\mathrm{y}}(z,t)$ are the electric and magnetic fields polarized in the x and y directions, respectively. J is current density, and $\epsilon\,(=\epsilon_0\epsilon_r)$ and $\mu\,(=\mu_0\mu_r)$ denote the permittivity and permeability of the medium, respectively. $c_{\mathrm{ref}}\,(=1/\sqrt{\epsilon_0\mu_0})$ denotes reference wave velocity, which was taken as the velocity of electromagnetic waves in free space. $g(z)$ represents an attenuation function that enforces amplitude decay of waves within the PML. The value of the attenuation function $g(z)$ is zero within the regular domain $(-L_l \le z \le L_{\mathrm{r}})$, whereas $g(z) > 0$ within the PML $(-L_{tl} \le z \le -L_l$ and $L_r \le z \le L_{tr})$.

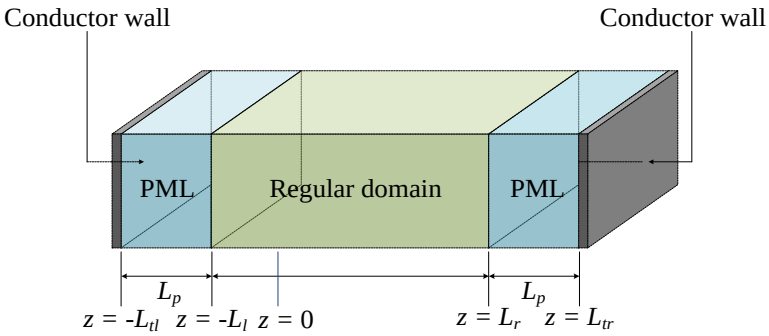


Figure 1: schematics of the PML-truncated domain

3 INVERSE PROBLEM

3.1 PDE-constrained optimization

The inverse medium problem of interest here is to reconstruct the permittivity profile of heterogeneous medium surrounded by PMLs, given measured electric fields to prescribed excitation at a point in the domain. The inverse problem can be formulated as the following optimization problem constrained by governing partial differential equations:

$$J \coloneqq \frac{1}{2}\sum_{i=1}^{N_r}\int_{-L_{tl}}^{L_{tr}}\int_0^T [E_x(z_i,t) - E_m(z_i,t)]^2\delta(z - z_i)dtdz + R(\epsilon) \tag{3}$$

subject to Eqs. (1)-(2).

The objective functional J in Eq. (3) consists of a misfit functional represented as the L^2 norm of difference between measured electric field $E_m(z_i,t)$ and calculated electric field $E_x(z_i,t)$. N_r denotes the number of receivers, and z_i is the location of receivers. To alleviate the ill-posedness of the inverse problem, regularization term $R(\epsilon)$ was introduced as a function of permittivity to be recovered. In this work, Tikhonov (TN) and Total Variation (TV) regularization schemes were used to investigate the regularization effect on the reconstructed permittivity profile of the PML-truncated domain. If TN regularization scheme

is used, $R(\epsilon)$ can take the form:

$$R^{TN}(\epsilon) \coloneqq \frac{r_\epsilon}{2}\int_{L_{tl}}^{L_{tr}}\left(\frac{d\epsilon}{dz}\right)^2 dz \tag{4}$$

whereas if TV regularization scheme is used, $R(\epsilon)$ can be formulated as:

$$R^{TV}(\epsilon) \coloneqq r_\epsilon \int_{L_{tl}}^{L_{tr}}\left[\left(\frac{d\epsilon}{dz}\right)^2 + \beta\right] dz \tag{5}$$

where r_ϵ is a regularization factor controlling the amount of penalty on the gradient of permittivity. In Eq. (5), a small parameter β was included to make $R^{TV}(\epsilon)$ differentiable when $d\epsilon/dz = 0$. It has been known that TN scheme is suited for the reconstruction of smooth profiles, whereas TV scheme is more relevant for the reconstruction of sharply-varying target profiles.

3.2 Lagrangian functional

To solve the PDE-constrained optimization problem, augmented Lagrangian method was introduced, by which the constrained optimization problem can be replaced by unconstrained optimization problem. The objective functional in Eq. (3) was augmented by side-imposing the PML-endowed Maxwell's equations via Lagrange multipliers to form the following Lagrangian functional.

$$\begin{aligned}
&\mathcal{L}\left(E_x,\, H_y,\, \lambda_{E_x},\, \lambda_{H_y}, \epsilon\right) \\
&\quad \coloneqq \frac{1}{2}\int_{-L_{tl}}^{L_{tr}}\int_0^T [E_x(z,t) - E_m(z,t)]^2 \delta(z) dt dz + R_\varepsilon(\varepsilon) \\
&\qquad + \int_{-L_{tl}}^{L_{tr}}\int_0^T \lambda_{H_y}\left(\mu\frac{\partial H_y}{\partial t} + c_{\text{ref}} g \mu H_y + \frac{\partial E_x}{\partial z}\right) dt dz \\
&\qquad + \int_{-L_{tl}}^{L_{tr}}\int_0^T \lambda_{E_x}\left(\epsilon\frac{\partial^2 E_x}{\partial t^2} + c_{\text{ref}} g \epsilon \frac{\partial E_x}{\partial t} + \frac{\partial^2 H_y}{\partial z \partial t} = -\frac{\partial J}{\partial t} - c_{\text{ref}} g J\right) dt dz
\end{aligned} \tag{6}$$

The permittivity profile $\epsilon(z)$ is sought to be reconstructed in an attempt to minimize the Lagrangian functional. The boundary and initial conditions in Eqs. (2a)-(2d) are explicitly imposed in the inversion process.

3.3 First-order optimality conditions

Optimal solutions of the state (E_x, H_y), adjoint $(\lambda_{E_x}, \lambda_{H_y})$, and control (ϵ) variables of the Lagrangian functional can be sought by enforcing the first variations of the Lagrangian to vanish. First of all, enforcing the vanishing of the first variations of $\mathcal{L}$ with respect to the Lagrange multipliers λ_{E_x} and λ_{H_y} results in the system of equations identical to Eqs. (1a) and (1b) for the forward problem.

Second, enforcing the vanishing of the first variations of $\mathcal{L}$ with respect to the state variables E_x and H_y yields the following mixed adjoint problem: Find $\lambda_{E_x} = \lambda_{E_x}(z,t)$ and $\lambda_{H_y} = \lambda_{H_y}(z,t)$, such that:

$$\epsilon\frac{\partial^2 \lambda_{E_x}}{\partial t^2} - c_{\text{ref}} g \epsilon \frac{\partial \lambda_{E_x}}{\partial t} - \frac{\partial \lambda_{H_y}}{\partial z} = -[E_x(z_i,t) - E_m(z_i,t)]\,\delta(z - z_i) \tag{7a}$$

$$\mu \frac{\partial^2 \lambda_{H_y}}{\partial t} - c_{\text{ref}} g \mu \lambda_{H_y} - \frac{\partial \lambda_{E_x}}{\partial z \partial t} = 0 \tag{7b}$$

and

$$\lambda_{E_x}(z,T) = \frac{\partial \lambda_{E_x}}{\partial t}(z,T) = 0 \tag{8a}$$

$$\lambda_{H_y}(z,T) = 0 \tag{8b}$$

The adjoint equations (7a) and (7b) are also PML-endowed, and coupled in terms of λ_{E_x} and λ_{H_y}. The source term of the adjoint problem depends on the misfit between computed and measured electric fields as indicated in Eq. (7a). It is noted that the adjoint problem is a final value problem, as opposed to the initial-value state problem.

Third, enforcing the vanishing of the first variations of $\mathcal{L}$ with respect to the control variable ϵ entails the following time-invariant control problem: Find $\epsilon = \epsilon(z)$, such that:

$$-R_\epsilon \frac{\partial^2 \epsilon}{\partial z^2} + \int_0^T \left(\lambda_{E_x} \frac{\partial^2 E_x}{\partial t^2} + c_{\text{ref}} g \lambda_{E_x} \frac{\partial E_x}{\partial t} \right) dt = 0 \tag{9a}$$

$$\frac{d\epsilon}{dz}(-L_{tl}) = \frac{d\epsilon}{dz}(L_{tr}) = 0 \tag{9b}$$

To reach Eq. (9a), TN regularization scheme was used in Eq. (4). If, instead, TV regularization scheme were used, the control equation would become

$$-R_\epsilon \beta \frac{\partial^2 \epsilon}{\partial z^2} \left[\left(\frac{d\epsilon}{dz} \right)^2 + \beta \right]^{-\frac{3}{2}} + \int_0^T \left(\lambda_{E_x} \frac{\partial^2 E_x}{\partial t^2} + c_{\text{ref}} g \lambda_{E_x} \frac{\partial E_x}{\partial t} \right) dt = 0 \tag{10}$$

The control problem is a boundary-value problem in terms of ϵ, which can be solved once the state and adjoint solutions are obtained.

4 MATERIAL PROPERTY UPADTE

The state, adjoint, and control problems can be solved simultaneously to obtain the optimal solutions of the inverse problem. However, the computational cost is expensive and the approach may not be practical for highly ill-posed inverse problems. In this work, a reduced space approach was introduced, where the permittivity profile is updated repeatedly as inversion iteration progresses. At each inversion iteration, the state variables E_x and H_y are calculated by solving Eqs. (1)-(2) assuming the distribution of permittivity profile within the domain. Then the adjoint variables λ_{E_x} and λ_{H_y} are calculated by solving Eqs. (7) using the state solutions obtained earlier. Finally, the control variable ϵ is updated using the calculated state and adjoint solutions. The permittivity profile at *k*-th inversion iteration can be updated as

$$\boldsymbol{\epsilon}_{k+1} = \boldsymbol{\epsilon}_k + \alpha \boldsymbol{d}_k \tag{11}$$

in which α is step length, and $\boldsymbol{d_k}$ is search direction vector. $\boldsymbol{\epsilon_k}$ is the vector of nodal permittivity values at *k*-th inversion iteration. The search direction vector $\boldsymbol{d_k}$ was determined by using Fletcher-Reeves conjugate gradient method as the following:

$$\boldsymbol{d}_k = \begin{cases} -\boldsymbol{g}_k & (k=0) \\ -\boldsymbol{g}_k + \dfrac{\boldsymbol{g}_k \cdot \boldsymbol{g}_k}{\boldsymbol{g}_{k-1} \cdot \boldsymbol{g}_{k-1}} \mathbf{d}_{k-1} & (k \geq 1) \end{cases} \tag{12}$$

where $\boldsymbol{g_k}$ denotes the vector of reduced gradient of the Lagrangian. The continuous form of the reduced gradient can be taken as the left-hand-side terms of Eq. (9a) with $g = 0$ in the regular domain as

$$\nabla_\epsilon \mathcal{L} = -R_\epsilon \frac{\partial^2 \epsilon}{\partial z^2} + \int_0^T \left(\lambda_{E_x} \frac{\partial^2 E_x}{\partial t^2} \right) dt \tag{13}$$

If TV regularization scheme were used, the reduced gradient would be

$$\nabla_\epsilon \mathcal{L} = -R_\epsilon \beta \frac{\partial^2 \epsilon}{\partial z^2} \left[\left(\frac{d\epsilon}{dz} \right)^2 + \epsilon \right]^{-\frac{3}{2}} + \int_0^T \left(\lambda_{E_x} \frac{\partial^2 E_x}{\partial t^2} \right) dt \tag{14}$$

In updating the vector $\boldsymbol{\epsilon_k}$ by Eq. (11), an inexact line search method was used, where the following Armijo condition was imposed.

$$J(\boldsymbol{\epsilon}_k + \alpha \boldsymbol{d}_k) \leq J(\boldsymbol{\epsilon}_k) + \mu \alpha \boldsymbol{g}_k \cdot \mathbf{d}_k \tag{15}$$

In Eq. (15), μ is usually chosen to be small. $\mu = 10^{-10}$ was used herein.

5 NUMERICAL RESULTS

Let us consider a one-dimensional PML-truncated domain with the regular domain occupying $-0.1\ \text{m} \leq z \leq 20\ \text{m}$ and the PML placed at $-5.1\ \text{m} \leq z \leq -0.1\ \text{m}$ as well as at $20\ \text{m} \leq z \leq 25\ \text{m}$, as shown in Fig. 2. As target permittivity profiles for reconstruction, two profiles were considered. The first one is a smoothly-varying profile represented by the following function:

$$\epsilon(z) = \epsilon_{Air} \times \exp\left[-\frac{(t-10)^2}{10} \right] + \epsilon_{Air} \tag{16}$$

where ϵ_{Air} denotes the permittivity of air and was taken as $\varepsilon_{Air} = 8.847 \times 10^{-12}\text{F/m}$. Fig. 3(a) shows the smoothly-varying profile corresponding to Eq. (16). The second profile is a sharply-varying profile as shown in Fig. 3(b), where it is assumed that concrete layer with $\epsilon_{conc} = 4.5\epsilon_{Air}$ exists in the range of $8.95\text{m} \leq \text{z} \leq 10.95\text{m}$.

In this setting, plane electromagnetic waves were assumed to be generated at origin ($z = 0$) by the acceleration of charges ($\partial J/\partial t$). As the source for electromagnetic waves, a wide-band Gaussian pulse type signal of the following expression was considered:

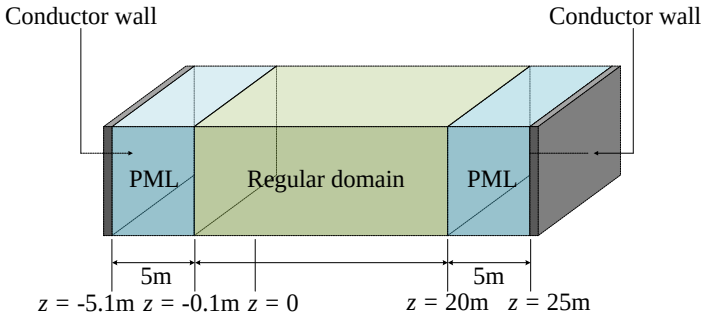
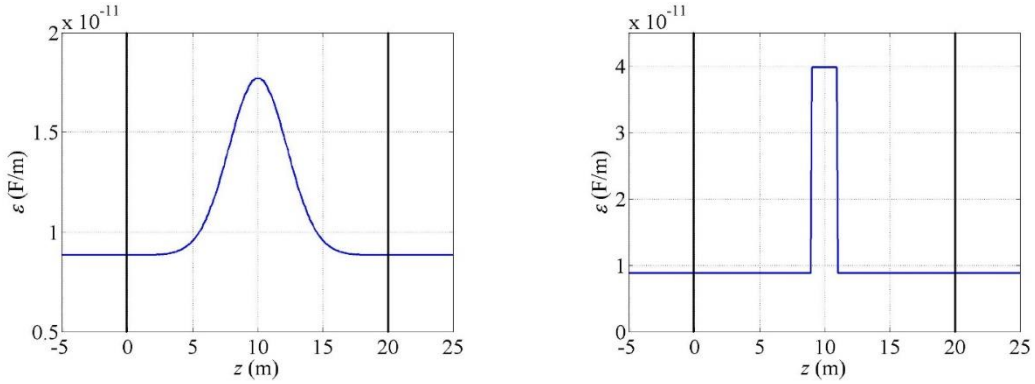


Figure 2: One-dimensional PML-truncated domain for material profile reconstruction

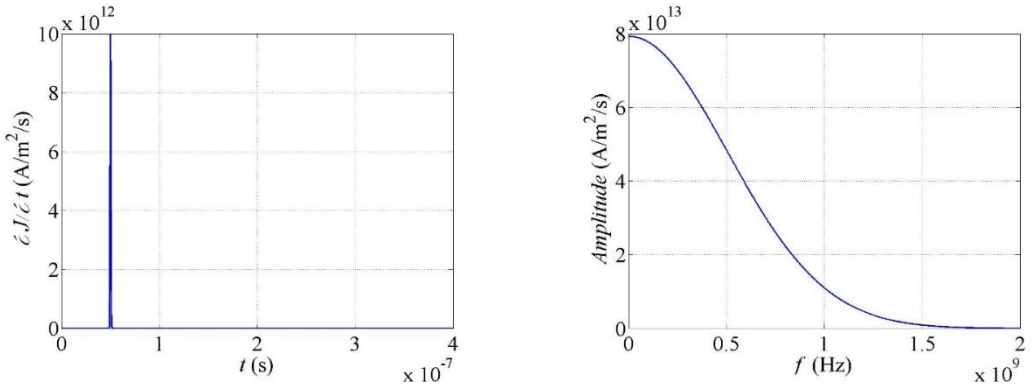


(a) Smooth profile (b) Sharply-varying layered profile

Figure 3: Target permittivity profiles

$$\frac{\partial J}{\partial t} = 10^{13}\exp\left[-\frac{(t-5\times10^{-8})^2}{2\times10^{-19}}\right] \tag{17}$$

Fig. 4 shows the time history and frequency spectrum of the Gaussian pulse, whose maximum frequency is about 2 GHz. In the simulation of wave propagation by finite element method, time step was set as $\Delta t = 8\times10^{-11}$ s and element length was set as $L_e = 0.1$ m. Electric field $E_x(z,t)$ and magnetic field $H_y(z,t)$ were spatially approximated with linear and quadratic shape functions, respectively.

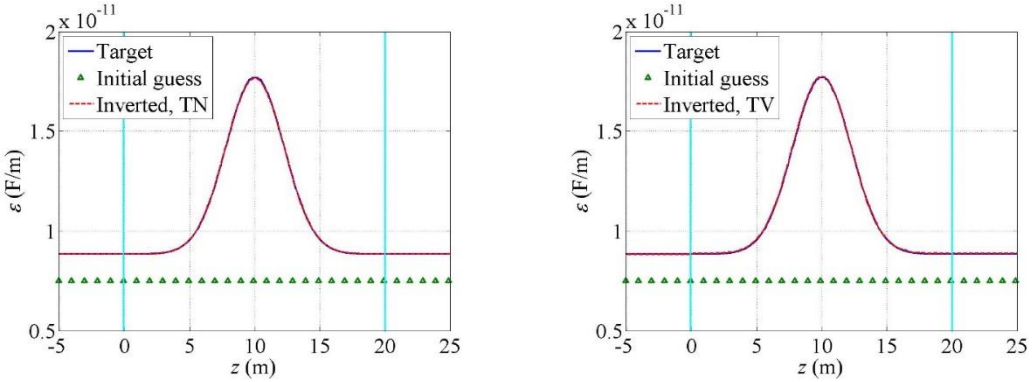


(a) Time history (b) Frequency spectrum

Figure 4: Time history and frequency spectrum of microwave source signal

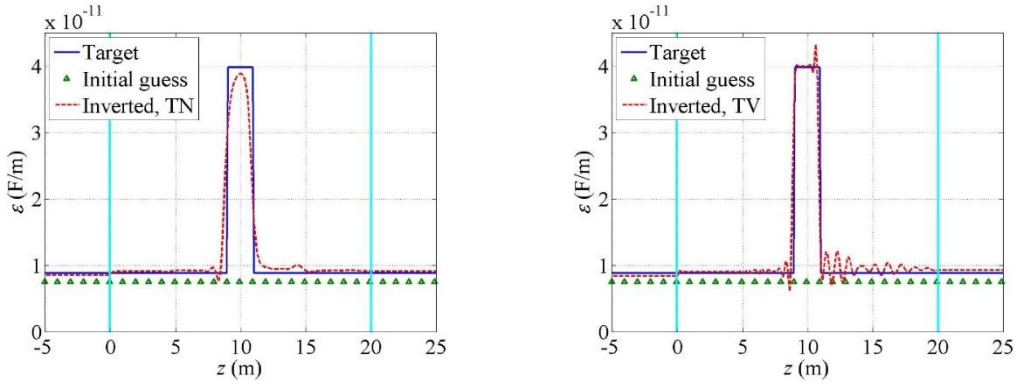
Fig. 5 shows the smoothly-varying permittivity profiles reconstructed using TN and TV regularization schemes. The inversion started with homogeneous initial guess of $\epsilon(z) = 0.75\times10^{-11}\mathrm{F/m}$, and the target profile was recovered at about 500 iterations with both TN and TV regularizations. Regularization factor was set as $r_\epsilon = 10^{23}$ for both regularization cases, and the small auxiliary parameter $\beta = 10^{-54}$ was used for TV regularization. As can be seen in Fig. 5, both TN and TV regularization schemes assisted the inversion process excellently in capturing the smoothly-varying target profile. Fig. 6 shows the sharply-varying permittivity profiles reconstructed using TN and TV regularization schemes. Initial guess, regularization factor, and the parameter β used for TV regularization were the same as in the case of smooth profile inversion. The layered target profile was recovered at about 2,000 iterations for both cases of TN and TV regularizations. The inversion with TV regularization

reconstructed the sharply-varying material interfaces better than the inversion with TN regularization, but slight oscillations appeared in the recovered profile when TV scheme was used.



(a) TN regularization (b) TV regularization

Figure 5: Inverted smoothly-varying permittivity profiles



(a) TN regularization (b) TV regularization

Figure 6: Inverted sharply-varying permittivity profiles

6 CONCLUSIONS

- A microwave full-waveform inversion method was developed for the reconstruction of permittivity profile in PML-truncated domains.
- The inversion of microwaves was conducted in a PDE-constrained optimization framework, using PML-endowed Maxwell's equations for plane electromagnetic waves as constraint conditions.
- The permittivity profile of medium was updated iteratively satisfying the first-order optimality conditions of augmented Lagrangian.
- Both smooth and layered permittivity profiles were reconstructed successfully by the full-waveform inversion method with the aid of Tikhonov and Total Variation regularization schemes. Sharply-varying layered profile could be effectively reconstructed using the Total Variation regularization.

REFERENCES

[1] P. Feng, X.Z. Lu, L.P. Ye, Experimental research and finite element analysis of square concrete columns confined by FRP sheets under uniaxial compression. *Proceedings of 17th Australasian conference on the mechanics of structures and materials*, Gold Coast, Australia, 2002.

[2] I. L. Al-Qadi, D.K. Ghodgaonkar, V.K. Varadan, Effect of moisture on asphaltic concrete at microwave frequencies. *IEEE Transactions on Geoscience and Remote Sensing*, **29(5)**, 710-717, 1991.

[3] D.K. Ghodgaonkar, Majid, B.W.A. Wan Mahmood, Majid, B.A. Rosnoizam, Accurate Measurement of Electromagnetic Properties of Concrete for nondestructive Evaluation at Microwave Frequencies. *Creating with Concrete, Universiiy of Dundee International Congress*, Dundee, Scotland, UK, September 6-10, 672-679, 1999.

[4] H.C. Rhim, O. Büyüköztürk, Electromagnetic properties of concrete at microwave frequency range. *Materials Journal,* **95(3)**, 262-271, 1998.

[5] H.C. Rhim, O. Büyüköztürk, Wideband microwave imaging of concrete for nondestructive testing. *Journal of Structural Engineering*, **126(12)**, 1451-1457, 2000.

[6] M. Donelli, A. Massa, Computational approach based on a particle swarm optimizer for microwave imaging of two-dimensional dielectric scatterers. *Microwave Theory and Techniques, IEEE Transactions on*, **53(5)**, 1761-1776, 2005

[7] M. Benedetti, M. Donelli, A. Martini, M. Pastorino, A. Rosani, A. Massa, An innovative microwave-imaging technique for nondestructive evaluation: Applications to civil structures monitoring and biological bodies inspection. . *Instrumentation and Measurement, IEEE Transactions on*, **55(6)**, 1878-1884, 2006

[8] R.G. Pratt, R.M. Shipp, Seismic waveform inversion in the frequency domain, Part 2: Fault delineation in sediments using crosshole data. *Geophysics*, **64(3)**, 902-914, 1999.

[9] S. Operto, C. Ravaut, L. Improta, J. Virieux, A. Herrero, P. Dell'Aversana, Quantitative imaging of complex structures from dense wide-aperture seismic data by multiscale traveltime and waveform inversions: a case study. *Geophysical prospecting*, **52(6)**, 625-651, 2004.

[10] J.W. Kang, L.F. Kallivokas, The Inverse Medium Problem in 1D PML-truncated Heterogeneous Semi-infinite Domains. *Inverse Problems in Science and Engineering*, **18(6)**, 759-786, 2010.

[11] J.W. Kang, L.F. Kallivokas, Mixed Unsplit-field Perfectly-matched-layers for Transient Simulations of Scalar Waves in Heterogeneous Domains. *Computational Geosciences*, **14(4)**, 623-648, 2010.

[12] J.W. Kang, L.F. Kallivokas, The Inverse Medium Problem in Heterogeneous PML-truncated Domains Using Scalar Probing Waves. *Computer Methods in Applied Mechanics and Engineering*, **200(1-4)**, 265-283, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

REDUCED BASIS METHOD FOR POISSON-BOLTZMANN EQUATION

Cleophas Kweyu[1], Martin Hess[2], Lihong Feng[3], Matthias Stein[4], and Peter Benner[5]

[1, 2, 3, 5] Computational Methods in Systems and Control Theory (CSC)
[4] Molecular Simulations and Design (MSD)
[1, 2, 3, 4, 5] Max Planck Institute of Dynamics of Complex Technical Systems
Sandtorstr. 1, 39106, Magdeburg (Germany)
e-mail: {kweyu, hessm, feng, matthias.stein, benner}@mpi-magdeburg.mpg.de

Keywords: Reduced basis method, Poisson-Boltzmann equation, ionic strength, finite differences scheme, preconditioned conjugate gradient, algebraic multigrid.

Abstract. *The Poisson-Boltzmann equation (PBE) is a nonlinear elliptic PDE that arises in biomolecular modeling and is a fundamental tool for structural biology. It is used to calculate electrostatic potentials of biomacromolecules in liquid solutions. To efficiently compute the electrostatic potential numerically, a very large domain is required to sufficiently accommodate both the biomacromolecule and the electrolyte. This yields high degrees of freedom in the resultant algebraic system of equations ranging from several hundred thousands to a few millions. This poses great computational challenges to conventional numerical techniques, especially when many simulations for varying parameters, for instance, the ionic strength, are to be run. The reduced basis method (RBM) greatly reduces this computational complexity by constructing a reduced order model of typically low dimension. We discretize the linearized PBE (LPBE) with a centered finite differences scheme and solve the resultant linear system by the preconditioned conjugate gradient (PCG) method with algebraic multigrid as the preconditioner. We then apply the RBM to the high-fidelity full order model (FOM) and present the numerical results. We notice that the RBM reduces the model order from* $\mathcal{N} = 1,614,177$ *to* $N = 6$ *at an accuracy of* 10^{-9} *and reduces computational time by a factor of approximately over* 1300.

1 INTRODUCTION

Electrostatics plays a crucial role in almost all processes that involve biomolecules in ionic solutions, for instance, in areas such as protein structural stability, biomolecular recognition, enzyme catalysis, among others [1]. Biomacromolecules such as nucleic acids and proteins have quite low internal dielectric constants ranging from 2 to 5 because of the fixed or frozen state of their constituent polar groups and side chains that cannot significantly reorient themselves in response to an applied external dielectric field [2, 3, 4]. On the other hand, water as the solvent has a high dielectric constant of around 80 because its dipoles are able to reorient freely. In principle, the Poisson-Boltzmann equation (PBE) is capable of modeling a system comprising of two dielectric materials with totally different dielectric properties. In this approach, the biomacromolecule is treated as a low dielectric cavity in which partial charges are discretely embedded at atomic positions, whereas the solvent is treated as a continuum through the Boltzmann distribution [1, 2, 3].

The PBE and its linear form have been widely applied in the field of structural biology. First, they are used to determine the electrostatic potential at the surface of a biomacromolecule. This is anticipated to provide information about the concentration of charged solute molecules in its vicinity which may, through inspection, indicate possible docking sites for biomolecules. Secondly, they determine the electrostatic potential outside the biomacromolecule that provides information on the free energy of interaction of micromolecules at different positions in the proximity of the biomacromolecule. Thirdly, they determine the free energy of a biomacromolecule or of its different states which provides information about its stability. Finally, they are used to determine the electrostatic field that is used to derive mean atomic forces which are added to the interatomic calculations in standard molecular dynamics [2, 3, 5].

In this study, we consider a charged protein molecule immersed in some electrolyte, for example, a salt solution at physiological concentration, and determine the electrostatic potential triggered by the interaction between the two particles. The kind of electrolyte considered is monovalent, in which the cations and anions are present in a ratio of $1:1$, which implies that the ionic strength equals the concentration of the ions. The ionic strength acts as a physical parameter of the LPBE, and we obtain the electrostatic potential under the variation of this parameter. The PBE is given as:

$$-\vec{\nabla}.(\epsilon(x)\vec{\nabla}u(x)) + \bar{k}^2(x)\sinh(u(x)) = (\frac{4\pi e_c^2}{k_B T})\sum_{i=1}^{N_m} z_i\delta(x-x_i), \quad u(\infty)=0 \tag{1}$$

where $u(x) = \frac{e_c\psi(x)}{k_B T}$ is the dimensionless potential and $\psi(x)$ the electrostatic potential in centimeter-gram-second (cgs) units at $x \in \mathbb{R}^3$. The dielectric coefficient $\epsilon(x)$ and the modified Debye-Hückel screening parameter $\bar{k}$, a function of the ionic strength of the solvent, are discontinuous functions at the interface between the charged biomacromolecule and the solvent, and at an ion exclusion region (Stern layer) surrounding the molecule, respectively. k_B is the Boltzmann constant and T the temperature. The term $k^2 = \frac{8\pi e_c^2 I}{1000\epsilon k_B T}$ describes ion accessibility and concentration, where I is the ionic strength given by $I = \frac{1}{2}\sum_{i=1}^{N} c_i z_i^2$. The right-hand side term represents the molecular charge density of the biomacromolecule consisting of N_m point partial charges $q_i = z_i e_c$ at atomic positions x_i and the distributions $\delta(.)$ reflect the point charge behaviour. Details on how to map the dielectric coefficient and the Debye-Hückel parameter to a computational grid can be found in [6].

The PBE (1) has some features which pose great challenges in both analytical and numerical

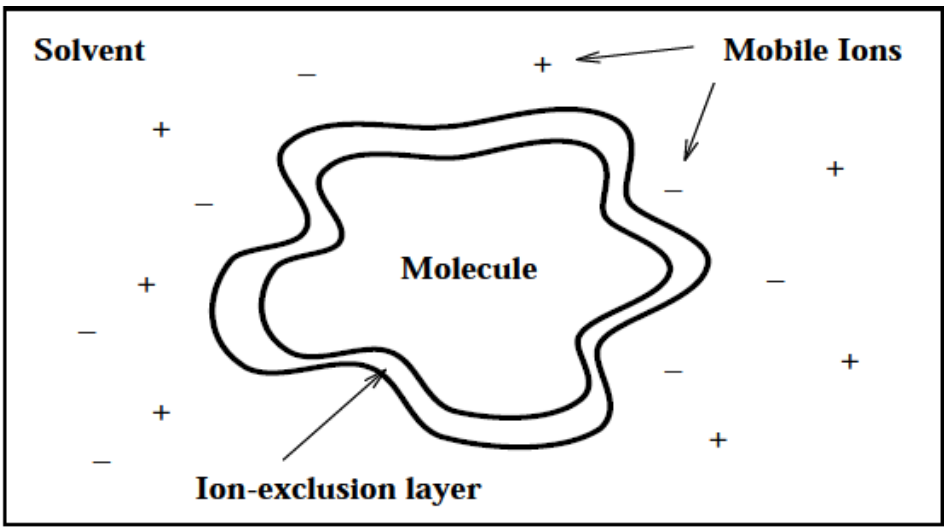


Figure 1: 2-D view of the 3-D Debye-Hückel model, a reproduction of Figure 1.1 in [7].

approaches. These include the infinite (unbounded) domain, delta functions, rapid nonlinearity, and discontinuous coefficients [7, 8]. Figure 1 shows the two dimensional view (cross-section) of the three dimensional Debye-Hückel model of electrostatic interactions between a biomolecule and a salt solution [7]. Proteins are not highly charged as compared to nucleic acids and, therefore, it suffices to consider the linearized PBE (LPBE) and still obtain accurate results. The PBE can be linearized under the assumption that the electrostatic potential is very small (relative to k_BT) or that the ions' electrostatic energy is much less than their thermal energy [1]. Therefore, the nonlinear function $\sinh(u(x))$ can be expanded by a Taylor series and only the first term is retained. We obtain the linearized PBE (LPBE) given by;

$$-\vec{\nabla}.(\epsilon(x)\vec{\nabla}u(x)) + \bar{k}^2(x)u(x) = (\frac{4\pi e_c^2}{k_BT})\sum_{i=1}^{N_m} z_i\delta(x-x_i). \tag{2}$$

However, we must note that the LPBE yields very inaccurate results for highly charged biomacromolecules such as nucleic acids (DNA and RNA) or polylysine, though it can be applied to low charged biomacromolecules such as proteins [3]. More information about the PBE, including its derivation from first principles can be found in [7].

Many different kinds of numerical methods have been used to solve the PBE and LPBE [9]. The finite difference method (FDM) is the most widely used method and has been implemented in many programs, for example, Amber, Delphi, and adaptive Poisson-Boltzmann solver (APBS) [10]. Other methods such as the finite element method (FEM), the boundary element method, the interface method, and an adaptive method have also been developed and a thorough review can be found in [9]. APBS, for instance, has versions for both FDM and FEM. As stated earlier, an infinite domain, $u(\infty) = 0$, coupled with the fact that molecules of interest such as enzymes and proteins have a spatial scale of 10 to over 100 Ångströms, provides a great challenge to numerical techniques. Therefore, it is customary to choose a truncated domain of at least threefold the size of the biomacromolecule to accurately approximate the boundary conditions [7]. Nevertheless, this leads to a very large algebraic system which becomes computationally expensive if many simulations for different ionic strengths have to be run. The reduced basis method may largely reduce the computational complexity due to its capability of constructing a reduced order model of typically low dimension, which can replace

the high-fidelity system in the multi-query tasks, and can be solved at a low computational cost [11]. In this paper, we apply the reduced basis method to the system of PBE, and have achieved significant acceleration of orders of magnitude.

The outline of this paper is as follows: In Section 2, we provide the basics of RBM which include the problem formulation, the solution manifold, the greedy algorithm, and the *a posteriori* error estimation. In Section 3 we provide numerical results of the FOM (via the FDM) and those of the ROM. Conclusions and future work are given in the end.

2 ESSENTIALS OF THE REDUCED BASIS METHOD

RBM is a projection-based parametrized model order reduction (PMOR) technique that aims at computing numerical solutions to parametrized PDEs in real-time and many-query scenarios by exploiting an offline/online procedure which ensures accurate approximation of the high-fidelity solution in a rapid and inexpensive manner. For a thorough review, see [12]. We consider a physical domain $\Omega \subset \mathbb{R}^3$ with boundary $\partial\Omega$, and a parameter domain $D \subset \mathbb{R}$. Since the exact solution is seldom available, we discretize the PDE (LPBE) with a centered finite differences (FD) scheme. The discrete problem of the LPBE now becomes, for any $\mu \in D$, find $u^{\mathcal{N}}(\mu)$ that satisfies the linear system

$$A(\mu)u^{\mathcal{N}}(\mu) = f, \quad \mu \in D, \tag{3}$$

where $A(\mu) \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ and $f \in \mathbb{R}^{\mathcal{N}}$. We notice that the $\mathcal{N}\times\mathcal{N}$ system is indeed computationally expensive to solve for an accurate approximation of $u(\mu)$ and therefore, we apply model order reduction (in this case, RBM) to save computational costs by providing an accurate approximation of $u^{\mathcal{N}}(\mu)$ at a greatly reduced dimension $N \ll \mathcal{N}$.

2.1 The solution manifold and the greedy algorithm

A very important assumption in RBM is the assumption that a typically very low dimensional solution manifold already covers all the high-fidelity solutions of (3) under variation of parameters [11]:

$$\mathcal{M}^{\mathcal{N}} = \{u^{\mathcal{N}}(\mu) : \mu \in D\}. \tag{4}$$

The RB approximation space is then built upon this solution manifold and is given by the subspace spanned by the snapshots, i.e. the subspace spanned by the high-fidelity $u^{\mathcal{N}}(\mu)$ solutions corresponding to a number of samples of the parameters

$$\text{range}(V) = \text{span}\{u^{\mathcal{N}}(\mu_1), ..., u^{\mathcal{N}}(\mu_N)\}, \quad \forall \mu_1, ..., \mu_N \in D. \tag{5}$$

The greedy algorithm is used to generate the reduced basis space (5) through an iterative procedure where a new basis is computed at each iteration. The procedure, as given by [13] is shown in Algorithm 1. The RB approximation is then formulated as, for any given $\mu \in D$, find $u_N(\mu) \in X_N$ which satisfies

$$A_N(\mu)u_N(\mu) = f_N, \tag{6}$$

where $A_N = V^T AV$ and $f_N = V^T f$. V is the orthonormal matrix computed from the greedy algorithm. From the fact that $N \ll \mathcal{N}$, solving the small dimensional ROM is much cheaper than solving the high-fidelity model (3) [11]. Efficient implementation of Algorithm 1 depends on an efficient error estimation $\Delta_N(\mu)$ of the ROM. In the next subsection, we introduce an *a posteriori* error estimation derived from the residual of the approximate RB solution.

Algorithm 1 Greedy algorithm

Input: A training set Ξ including all the samples of μ, i.e., $\Xi := \{\mu_1, \ldots, \mu_m\}$.
Output: RB basis represented by the projection matrix V.

Choose $\mu^* \in \Xi \subset D$ arbitrarily
Solve (3) for $u^{\mathcal{N}}(\mu^*)$
$S_1 = \{\mu^*\}$
$V_1 = [u^{\mathcal{N}}(\mu^*)]$
$N = 1$
while $\max_{\mu \in \Xi} \Delta_N(\mu) \geq \epsilon$ **do**
    $\mu^* = \arg\max_{\mu \in \Xi} \Delta_N(\mu)$
    Solve (3) for $u^{\mathcal{N}}(\mu^*)$
    $S_{N+1} = S_N \cup \mu^*$
    $V_{N+1} = [X_N \quad u^{\mathcal{N}}(\mu^*)]$
    Orthonormalize the columns of V_{N+1}
    $N = N + 1$
end while

2.2 *A posteriori* error estimation

A posteriori error estimators are computable indicators that estimate the actual solution error by utilizing the residual of the approximate RB solution. An error estimator is required to fulfill three major properties: it is supposed to be sharp (close to the unknown actual error), asymptotically correct (tend to zero with increasing RB space dimension N, at a similar rate as the actual error), and computationally cheap. Therefore, they guarantee both reliability and efficiency of the reduction process [14].

We first compute the residual;

$$\begin{aligned} r_N(u_N;\mu) &= f - A(\mu)u_N(\mu) \\ &= A(\mu)u^{\mathcal{N}}(\mu) - A(\mu)u_N(\mu) \\ &= A(\mu)e(\mu) \end{aligned} \tag{7}$$

where the error $e(\mu) := u^{\mathcal{N}}(\mu) - u_N(\mu)$ is given by

$$e(\mu) = (A^{\mathcal{N}}(\mu))^{-1} r_N(u_N;\mu) \tag{8}$$

We obtain an upper bound for the 2-norm of the error by taking the 2-norm on both sides of equation (8), that is;

$$\|e(\mu)\|_2 \leq \|(A^{\mathcal{N}})^{-1}(\mu)\|_2 \|r_N(u_N;\mu)\|_2 = \frac{1}{\sigma_{min}(A^{\mathcal{N}}(\mu))} \|r_N(u_N;\mu)\|_2 =: \Delta_N(\mu), \tag{9}$$

where $\sigma_{min}(A^{\mathcal{N}}(\mu))$ is the smallest singular value of $A^{\mathcal{N}}(\mu)$ [14]. The quantity $\Delta_N(\mu)$ is a rigorous error bound, and can be used to select snapshots within the greedy algorithm in the offline stage and consequently to measure the accuracy of the RB approximation in the online stage [13]. For efficient computation of the norm of the residual and error bounds, see [15, 14]. It is computationally expensive, especially in the offline-online context, to compute $\sigma_{min}(A^{\mathcal{N}}(\mu))$ as it entails solutions of large-scale eigenvalue problems [13]. Therefore, in our computations, we use the norm of the residual as our error estimator, an upper bound which satisfies the inequality

(9) and provides an estimation of the true error that works well for our problem. It also provides rapid convergence as depicted in the numerical results in Figures 6 and 7. It is given by

$$\|e(\mu)\|_2 \approx \|r_N(u_N;\mu)\|_2 = \Delta_N(\mu). \tag{10}$$

3 NUMERICAL RESULTS

3.1 Finite differences results

We consider the LPBE in (2), a parameter domain $\mu \in D = [0.05, 0.15]$, and a physical domain $\Omega = [7.4415, 64.5655]\text{Å} \times [-1.5120, 44.2180]\text{Å} \times [-20.2755, 35.8775]\text{Å}$. The parameter domain is chosen for a feasible physiological process and μ resides in the second term in the kappa function. Information about the molecular charge density (right-hand side of the LPBE) was obtained from a PQR file provided by the MSD group at our institute. The file contains 1228 protein atoms (constituents of amino acid residues) at various atomic positions (P), with respective charges (Q), and radii (R). The PQR file is generated from the PDB (Protein Data Bank) file by the PDB2PQR software which allows users to add charges and assign atomic radii as additional parameters. The PDB file provides a standard representation for macromolecular structure data (in terms of atomic coordinates), derived from X-ray crystallography and nuclear magnetic resonance (NMR) studies for electrostatic potential analysis [16].

We discretize the LPBE with a centered finite differences scheme and the resulting parametrized linear system (3) is of more than 1.6×10^6 degrees of freedom. We solve this full order model (FOM) by the preconditioned conjugate gradient (PCG) method with algebraic multigrid as the preconditioner at different samples of ionic strength. We use a tolerance of 10^{-10} for the PCG and the solution converges within 50 iterations. The choice of this tolerance directly affects that of the greedy algorithm proportionally and therefore, it is prudent to ensure that the truth solution is highly accurate. Zero Dirichlet boundary conditions are used to avoid the dependence of the boundary data on the parameter as well. Figures 2-5 below plot cross-sections of the $z-$ axis and show the exponential decay of the high-fidelity solutions $(u^{\mathcal{N}}(\mu), \mathcal{N} = 1,614,177)$ with variation of the parameter μ. This is attributed to the large force constant (332 kcal/mol) of electrostatic interactions. In the absence of ions (that is, at $\mu = I = 0$), these interactions are long ranged, but in the presence of ions (that is, $\mu > 0$), they are damped and gradually decay to zero [1]. The computational time taken to obtain the high-fidelity solution $u^{\mathcal{N}}(\mu)$ is on average approximately 90 seconds, and varies depending on the value of the ionic strength used.

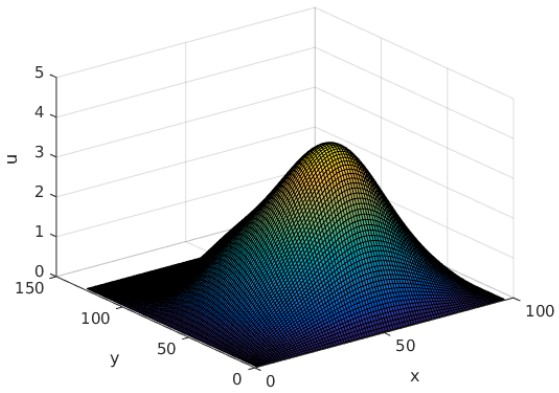


Figure 2: $u^{\mathcal{N}}(\mu)$ at μ = 0.05

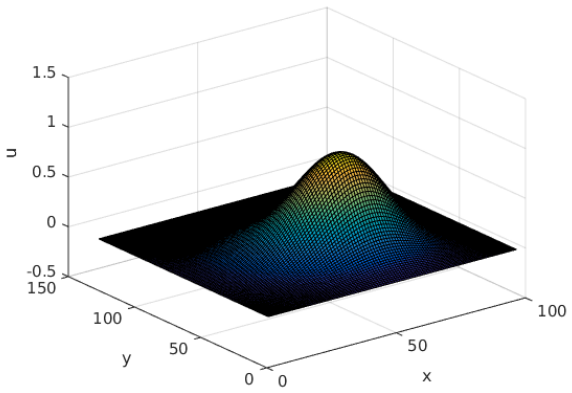


Figure 3: $u^{\mathcal{N}}(\mu)$ at μ = 0.5

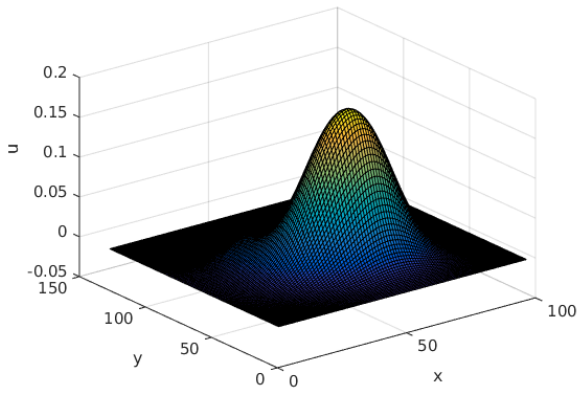


Figure 4: $u^{\mathcal{N}}(\mu)$ at $\mu = 1.5$

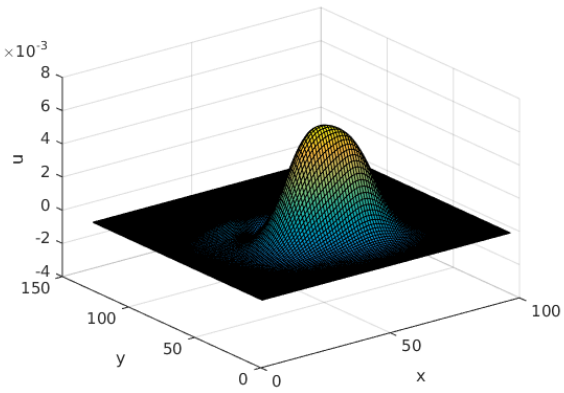


Figure 5: $u^{\mathcal{N}}(\mu)$ at $\mu = 5$

3.2 Reduced basis results

The *a posteriori* error estimates the true error between the high-fidelity solution and the reduced basis solution. It also provides the exponential decay of the error in the greedy algorithm until a reasonable tolerance is achieved. Figures 6 and 7 show the decays of the error estimator and the true error during the greedy algorithm at the current RB dimension $i = 1, \ldots, N$. They satisfy the asymptotic correctness property stated in subsection 2.2, and it is evident that the error estimator is an upper bound to the true error. We also observe a high convergence rate of the error estimator with up to two orders of magnitudes. The error estimators in Figures 6 and 7 are the maximal error and relative maximal error, respectively, and are defined as, maximal error: $\Delta_N^{\max} = \max_{\mu \in \Xi} \|r_N(u_N;\mu)\|_2$, relative maximal error: $\frac{\Delta_N^{\max}}{\|u_N(\mu^*)\|_2}$, $\mu^* = \arg\max_{\mu \in \Xi} \|r_N(u_N;\mu)\|_2$.

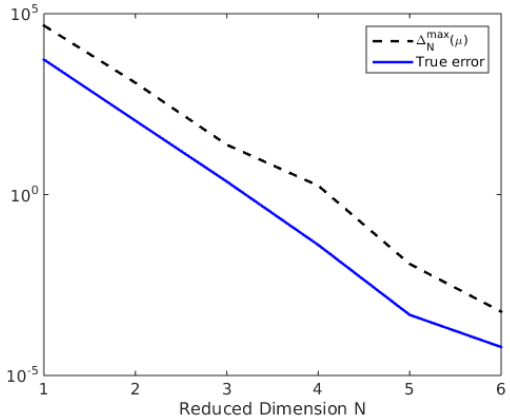


Figure 6: Maximal versus true error

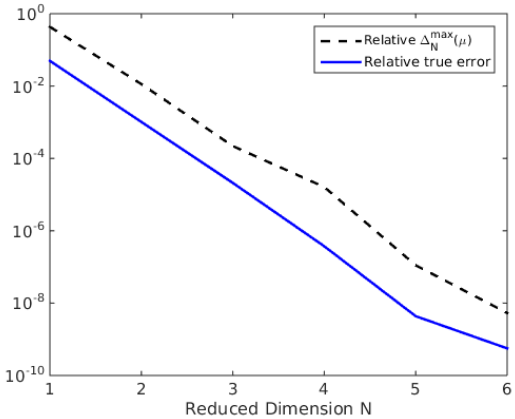


Figure 7: Relative maximal versus true error

Figure 8 shows the error estimator and the true error of the finally constructed reduced model over the whole parameter domain $\mu_i \in \Xi$ for $i = 1, ..., 100$. It is evident that the error estimate for the final RB approximation at dimension $N = 6$ is indeed an upper bound of the true error and a trend that both quantities behave similarly is clearly visible from the graphs. The

computational time taken to obtain the approximate solution $u_N(\mu)$ is on average approximately 0.065 seconds. Figure 9 validates the true error of the ROM of dimension $N = 6$, i.e., the final ROM constructed by the greedy algorithm, at some 100 randomly selected parameter values $\mu \in D$ which are different from those in Ξ, used in Algorithm 1. It is evident that the true error is still smaller than the error estimate.

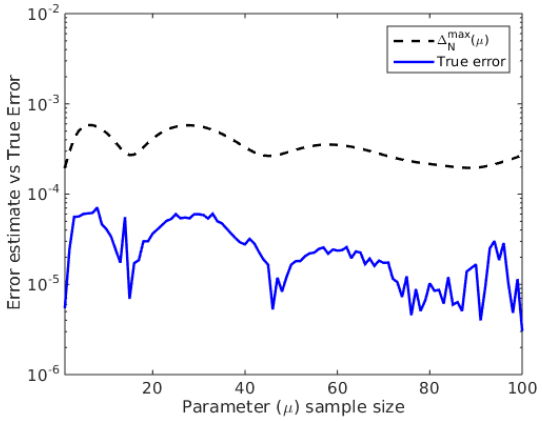


Figure 8: Error estimate versus true error

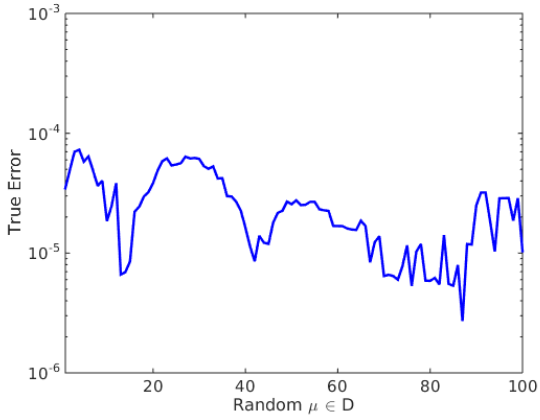


Figure 9: True error

4 CONCLUSIONS

In this paper, we have presented the RBM and its application to the LPBE with ionic strength as the meaningful parameter. The RBM reduces the high dimensional full order model by a factor of $269,029.5$ and the computational time by a factor of approximately over 1300. The error estimator provided fast convergence to the reduced basis approximation. Our future research is on employing non-zero Dirichlet boundary conditions from Debye-Hückel model for a spherical non-polarizable ion with uniform charge. These boundary conditions also vary with the parameter, which is non-affinely inherent in them and therefore we need to employ the discrete empirical interpolation method (DEIM) for a speed-up in the online stage.

5 ACKNOWLEDGEMENTS

We would like to thank the following organizations for financial and material support on this project: International Max Planck Research School (IMPRS) and Max Planck Society (MPI).

REFERENCES

[1] F. Fogolari, A. Brigo, H. Molinari, The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of Molecular Recognition*, **15**, 377–392, 2002.

[2] F. Fogolari, P. Zuccato, G. Esposito, P. Viglino, Biomolecular electrostatics with the linearized Poisson-Boltzmann equation. *Biophysical Journal*, **76**, 1–16, 1999.

[3] B. Honig, A. Nicholls, Classical electrostatics in biology and chemistry. *Science, New Series*, **268**, 1144–1149, 1995.

[4] http://www.edusoft-lc.com/zap/manuals/100/chaptwo.html. Last visited 2016-02-07.

[5] K.A. Sharp, B. Honig, Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Chem.*, **19**, 301–332, 1990.

[6] N.A. Baker, M.J. Holst, F. Wang, The adaptive multilevel finite element solution of the Poisson-Boltzmann equation on massively parallel computers. *IBM J. Res. Devel.*, **45**, 427–438, 2001.

[7] M.J. Holst, *Multilevel methods for the Poisson-Boltzmann equation, PhD Thesis*. Numerical Computing group, University of Illinois at Urbana-Champaign, 1993.

[8] M. Holst, N. Baker, F. Wang, Adaptive multilevel finite element solution of the Poisson-Boltzmann equation I: algorithms and examples. *Journal of Computational Chemistry*, **21**, 1319–1342, 2000.

[9] B.Z. Lu, Y.C. Zhou, M.J. Holst, J.A. McCammon, Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications. *Commun. Comput. Phys.*, **3**, 973–1009, 2008.

[10] J. Wang, R. Luo, Assessment of linear finite difference Poisson-Boltzmann solvers. *J. Comp. Chem.*, **31**, 1689–1698, 2010.

[11] J.L. Eftang, *Reduced basis methods for parametrized partial differential equations, PhD Thesis*. Norwegian University of Science and Technology, Trondheim, 2011.

[12] P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model order reduction methods for parametrized dynamical systems. *SIAM Review*, **57**, 483–531, 2015.

[13] M.W. Hess, P. Benner, Fast evaluation of time-harmonic Maxwell's equations using the reduced basis method. *IEEE Transactions on Microwave Theory and Techniques*, **61**, 2265–2274, 2013.

[14] A. Quarteroni, A. Manzoni, F. Negri, *Reduced basis methods for parametrized partial differential equations: an introduction*. Springer International Publishing, 2015.

[15] J.S. Hesthaven, G. Rozza, B. Stamm, *Certified reduced basis methods for parametrized partial differential equations*. Springer International Publishing, 2015.

[16] http://www.poissonboltzmann.org/docs/file-format-info/. Last visited 2016-02-25.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

AN ITERATIVE MODEL ORDER REDUCTION SCHEME FOR A SPECIAL CLASS OF BILINEAR DESCRIPTOR SYSTEMS APPEARING IN CONSTRAINT CIRCUIT SIMULATION

Pawan Goyal[1], **Peter Benner**[1]

[1] Max Planck Institute for Dynamics of Complex Technical Systems,
Sandtorstr. 1, 39106, Magdeburg, Germany.
e-mail: {goyalp,benner}@mpi-magdeburg.mpg.de

Keywords: Model order reduction, Descriptor systems, Bilinear systems, Electrical circuits.

Abstract. *We focus on interpolatory-based model order reduction for a special class of bilinear descriptor systems in the $\mathcal{H}_2$-optimal framework, appearing in constraint circuit simulations. The straightforward extension of the $\mathcal{H}_2$-optimality conditions for ODE systems to descriptor systems generically may produce an unbounded error in the $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm, or both. This arises due to the inappropriate use of the polynomial part of the system. To ensure bounded error, one needs to deal with the polynomial part of the systems properly. To do so, we first transform these descriptor systems into equivalent ODE systems by means of oblique projectors, as it is widely done in the literature for linear index-2 ODEs. This enables us to employ bilinear iterative rational Krylov algorithm (*B-IRKA*) which provides us locally $\mathcal{H}_2$-optimal reduced-order systems on convergence, if it converges. Unfortunately, the direct implementation of* B-IRKA *on equivalent ODEs requires the expensive explicit computation of the oblique projectors. Therefore, as one of our contributions, we show how to apply* B-IRKA *to the equivalent bilinear ODE system without an explicit computation of the projectors. We demonstrate the efficiency of the proposed technique by means of several constraint circuit examples and compare the quality of the reduced-order systems with the ones obtained by using the projection matrices determined by applying* IRKA *on the corresponding linear part.*

1 Introduction

We discuss the interpolatory-based model order reduction technique for bilinear descriptor systems which are of the form

$$\begin{aligned} E_{11}\dot{x}_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + \sum_{k=1}^{m} N_k x_1(t)u_k(t) + B_1 u(t), \\ 0 &= A_{21}x_1(t) + B_2 u(t), \\ y(t) &= C_1 x_1(t) + C_2 x_2(t) + Du(t), \end{aligned} \tag{1}$$

where $x_1(t) \in \mathbb{R}^{n_1}$, $x_2(t) \in \mathbb{R}^{n_2}$ are the generalized states, $y(t) \in \mathbb{R}^p$ and $u(t) \in \mathbb{R}^m$ are the output and input of the system, respectively, and all the matrices are of appropriate dimensions. It is assumed that E_{11} and $A_{21}E_{11}^{-1}A_{12}$ are invertible. This implies that the dynamical system (1) is a Hessenberg index-2 differential algebraic system [1] in case of $N_k = 0$. Generally, these special bilinear systems (1) arise from semi-discretization of Navier-Stokes equations or constraint RLC circuits. As a motivating example, we consider a constraint transmission circuit as shown in Figure 1.

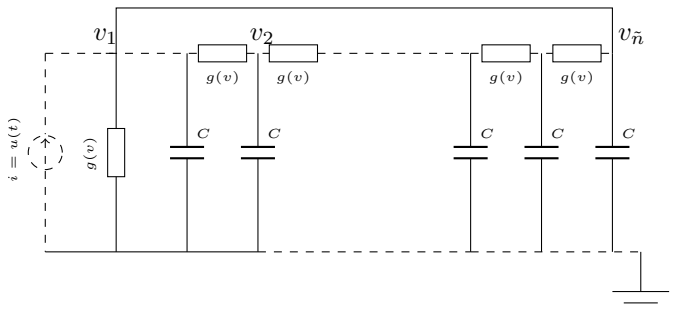


Figure 1: A constraint nonlinear transmission line.

The above transmission circuit contains nonlinear diodes, $g(v) = e^{40v_D} + v_D - 1$, where v_D is the voltage difference across the nodes. Using Kirchhoff's current law, we can model the dynamics of the circuit as a quadratic-bilinear descriptor system, having an index-2 matrix pencil $\lambda E - A$ (details are presented in Section 5). Nonetheless, such quadratic systems, having an index-2 matrix pencil, can be approximated as bilinear systems via Carleman bilinearization. The approximated bilinear systems have a similar structure as (1). We refer to Section 5 for a couple of more constraint circuit examples, having the same structure as well. Therefore, there is a need to develop efficient model reduction techniques for such bilinear structured systems.

To have high accuracy in modeling of physical phenomena or for a better understanding of the underlying process, the governing differential equations are discretized very fine over the spatial domain. As a consequence, we obtain large-scale complex dynamical systems whose numerical simulations, optimization and control study become a huge numerical burden and inefficient. Thus, model order reduction (MOR) brings forth a solution to the immense numerical demand for such large-scale systems. MOR aims to replace these large-scale systems by small surrogate systems of much smaller dimensions, while capturing the important dynamical characteristics of the original system.

MOR for linear systems has proven to be successful and efficient, see, cf. [2, 3, 4]. The extension to the bilinear systems has drawn significant attention in recent times. Many of MOR techniques have been extended from linear systems to bilinear systems. For example,

Gramian-based MOR such as balanced truncation has been extended to bilinear systems in [5] and interpolatory-based methods are extended to bilinear systems in [6, 7, 8, 9]. Obviously, Gramian-based MOR requires the solutions of two generalized Lyapunov equations which are computationally cumbersome, although efficient methods are proposed to solve these generalized Lyapunov equations in [10, 11].

On the other hand, interpolatory-based MOR seeks to determine the projection matrices such that each subsystem of the reduced bilinear system interpolates the corresponding subsystem of the original system at predefined interpolation points. For the first time in [12], the problem related to the $\mathcal{H}_2$-optimal MOR for bilinear systems was considered, wherein the Gramian-based Wilson conditions [13] for the $\mathcal{H}_2$ optimality for linear systems are extended to bilinear systems. A similar problem, but in a rather different way, was again considered in [7], where the first-order necessary conditions for $\mathcal{H}_2$-optimality are derived by taking the derivative of the $\mathcal{H}_2$-norm of the error system with respect to the elements of the reduced-order system's realization. Based on these conditions, an iterative method, the so-called bilinear iterative rational Krylov algorithm *(B-IRKA)* is proposed, extending the iterative rational Krylov algorithm *(IRKA)* for linear systems to bilinear systems.

Notwithstanding, there are many up-front questions when it comes to MOR of bilinear descriptor systems. The Gramian-based MOR for bilinear descriptor systems is still an open question, but interpolatory conditions can be readily extended to descriptor systems by just replacing I by E. However, as discussed for linear descriptor systems in [14] that straightaway extending interpolation conditions for ODE systems to descriptor systems can lead to poor reduced-order systems due to mismatch of the polynomial parts of the systems. This may give rise to an unbounded error in the $\mathcal{H}_2$-norm. This holds for bilinear systems as well. Therefore, a special attention to the polynomial part is required to ensure a bounded error. Considering a special class of bilinear descriptor systems whose kth order subsystem has a constant polynomial part, the modified interpolation conditions are proposed in [15]. Therein, a special attention to the polynomial parts of the bilinear systems is paid. Later, the problem of obtaining locally $\mathcal{H}_2$-optimal reduced-order systems of such descriptor systems is considered in [16].

In this paper, we focus on MOR for bilinear descriptor systems (1), having an index-2 matrix pencil, in the $\mathcal{H}_2$-optimal framework. In this regard, we first transform the system (1) into an equivalent bilinear ODE system by means of projections as it is done in [17] for Stokes-type quadratic-bilinear descriptor systems. This allows us to employ the version of *B-IRKA* which is extended in [18] from the $E = I$ case to the $E \neq I$ case. However, the direct implementation of *B-IRKA* requires the explicit computation of the projectors which is highly undesirable. Therefore, as one of the main contributions in this paper, we show how to apply *B-IRKA* without explicit computation of the projectors.

As mentioned before, the application of MOR for bilinear systems can also be seen in the reduction of quadratic-bilinear systems [8, 19]. This is done by first determining an approximate bilinearized system by using Carleman bilinearization process and then employing MOR techniques for bilinear systems. This way, one loses the quadratic-bilinear structure of the system and blows up the dimension of the state vector but it is possible to achieve a reduced bilinear system of much smaller dimension. The Carleman bilinearization process for ODE systems has been widely studied; see [19] and it is recently extended to descriptor systems in [20], having an index-1 matrix pencil $\lambda E - A$. In this paper, we extend the Carleman bilinearization process for quadratic-bilinear descriptor systems, with an index-2 matrix pencil, to obtain approximate bilinearized systems, under the assumption that $x_2(t)$ in (1) is a scalar variable. We also show that the bilinearized systems preserve the index of the matrix pencil.

In the following section, we briefly outline *B-IRKA* for bilinear ODE systems. In Section 3, we present the transformation of the bilinear descriptor systems (1) into equivalent ODE systems and apply *B-IRKA* to obtain reduced-order systems. Therein, we also discuss the computational issues and present implementation details. In Section 4, the Carleman bilinearization process for quadratic-bilinear descriptor systems, having index-2 of the matrix pencil $\lambda E - A$, is presented. In Section 5, we demonstrate the efficiency of the proposed MOR technique by means of several constraint electrical circuit examples.

2 $\mathcal{H}_2$-Optimal Model Reduction of Bilinear Systems

In this section, we briefly review MOR for bilinear ODE systems in the $\mathcal{H}_2$-optimal framework. We consider bilinear systems in the following form:

$$\Sigma := \begin{cases} E\dot{x}(t) = Ax(t) + \sum_{k=1}^{m} N_k x(t)u_k(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the input and output of the system, respectively, and all other matrices are of appropriate sizes. Here, E is considered to be nonsingular. The $\mathcal{H}_2$-norm expression for bilinear systems (2) is defined in [12] and can be given as follows:

$$\|\Sigma\|_{\mathcal{H}_2} = \operatorname{tr}\left(\sum_{k=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}\sum_{l_1,\ldots,l_k=1}^{m} h_k^{(l_1,\ldots,l_k)}(h_k^{(l_1,\ldots,l_k)})^T dt_1 dt_2\cdots dt_k\right),$$

in which $h_k^{(l_1,\ldots,l_k)} = Ce^{\tilde{A}t_k}\tilde{N}_{l_1}e^{\tilde{A}t_{k-1}}\tilde{N}_{l_2}\cdots\tilde{N}_{l_k}e^{\tilde{A}t_1}\tilde{b}_{l_k}$ with $\tilde{A} = E^{-1}A$, $\tilde{N}_i = E^{-1}N_i$, $\tilde{B} = E^{-1}B$ and $\tilde{b}_i$ is the ith column of the matrix $\tilde{B}$. Assuming the bilinear system is bounded input bounded output stable, the $\mathcal{H}_2$-norm can also be computed in terms of the Gramians associated with the bilinear system, i.e.,

$$\|\Sigma\|_{\mathcal{H}_2}^2 = \operatorname{tr}(CPC^T) = \operatorname{tr}(B^TQB),$$

where the reachability Gramian P and the observability Gramian Q are the solutions to the following generalized Lyapunov equations:

$$APE^T + EPA^T + \sum_{k=1}^{m} N_kPN_k^T + BB^T = 0$$

and

$$A^TQE + E^TQA + \sum_{k=1}^{m} N_k^TQN_k + C^TC = 0,$$

respectively. The main goal of $\mathcal{H}_2$-optimal MOR is to determine an r-order reduced system, satisfying

$$\Sigma_r = \arg\min_{\|\Sigma_r\|_{\mathcal{H}_2}<\infty} \|\Sigma - \Sigma_r\|_{\mathcal{H}_2}.$$

As noted, this problem for $E = I$ is investigated in [7] and an iterative scheme is proposed to determine reduced-order systems, satisfying $\mathcal{H}_2$-optimality conditions. This is extended to $E \neq I$ in [18]. In the following, we sketch *B-IRKA* to obtain reduced-order systems, satisfying the necessary conditions for $\mathcal{H}_2$-optimality.

Algorithm 1 MOR for Bilinear ODE Systems.

1: **Input:** E, A, N_k, B, C.
2: Make an initial guess of $\hat{E}, \hat{A}, \hat{N}_k, \hat{B}, \hat{C}$.
3: **while** no convergence **do**
4: Determine nonsingular S, R such that $S\hat{A}R = \Lambda$ and $S\hat{E}R = \hat{I}$, where $\hat{I}$ is the identity matrix.
5: Compute $\tilde{B} = \hat{B}^T S^T, \quad \tilde{C} = \hat{C}R$ and $\tilde{N}_k = R^T \hat{N}_k^T S^T$.
6: Determine the projection matrices V and W:

$$\operatorname{vect}(V) = -(\Lambda \otimes E + \hat{I} \otimes A + \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k)^{-1}(\tilde{B}^T \otimes B)\operatorname{vect}(I_m),$$

$$\operatorname{vect}(W) = -(\Lambda \otimes E + \hat{I} \otimes A + \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k)^{-T}(\tilde{C}^T \otimes C^T)\operatorname{vect}(I_p),$$

where $I_q \in \mathbb{R}^{q\times q}$ is the identity matrix, m and p are the numbers of inputs and outputs, respectively, and the operator $\operatorname{vect}(\cdot)$ yields a column vector by stacking columns of the matrix on top of each other.
7: Compute the reduced-order system matrices:

$$\hat{E} = W^T E V, \quad \hat{A} = W^T A V, \quad \hat{N}_k = W^T N_k V, \quad \hat{B} = W^T B, \quad \hat{C} = CV.$$

8: **end while**
9: **Output:** $\hat{E}^{opt} = \hat{E}, \quad \hat{A}^{opt} = \hat{A}, \quad \hat{N}_k^{opt} = \hat{N}_k, \quad \hat{B}^{opt} = \hat{B}$ and $\hat{C}^{opt} = \hat{C}$.

3 MOR for Bilinear Descriptor Systems

In this section, we focus on MOR for bilinear descriptor systems (1), having an index-2 matrix pencil, in the $\mathcal{H}_2$-optimal framework. We begin with the case $B_2 = 0$, i.e.,

$$E_{11}\dot{x}_1(t) = A_{11}x_1(t) + A_{12}x_2(t) + \sum_{k=1}^{m} N_k x_1(t)u_k(t) + B_1 u(t), \tag{2a}$$

$$0 = A_{21}x_1(t), \quad x_1(0) = 0, \tag{2b}$$

$$y(t) = C_1 x_1(t) + C_2 x_2(t) + Du(t), \tag{2c}$$

where the dimensions of the matrices are the same as in (1). We assume that E_{11} and $A_{21}E_{11}^{-1}A_{12}$ are invertible, and E_{11} is symmetric. As a first step, we transform the bilinear descriptor system into an equivalent bilinear ODE system. Following the same steps as for the Stokes-type quadratic-bilinear descriptor system in [17], we obtain the following system:

$$E_{11}\dot{x}_1(t) = \Pi A_{11}x_1(t) + \sum_{k=1}^{m} \Pi N_k x_1(t)u_k(t) + \Pi B_1 u(t), \quad x_1(0) = 0, \tag{3a}$$

$$y(t) = \mathcal{C}x_1(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)} x_1(t)u_k(t) + \mathcal{D}u(t), \tag{3b}$$

where

$$\mathcal{C} = C_1 - C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}A_{11}, \quad \mathcal{C}_N^{(k)} = -C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}N_k,$$
$$\mathcal{D} = D - C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}B_1$$

and

$$\Pi = I - A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}. \tag{4}$$

Throughout this paper, for simplicity we assume $A_{21} = A_{12}^T$. However, $A_{21} \neq A_{12}^T$ can be treated in the current bilinear framework as well, if one extends the arguments used in [14]. By $A_{12}^T x_1(t) = 0$, we have that $\Pi^T x_1(t) = x_1(t)$, cf., e.g., [21]. Replacing $x_1(t)$ by $\Pi^T x_1(t)$ in (3) and premultiplying by Π, we obtain the following system:

$$\Pi E_{11}\Pi^T \dot{x}_1(t) = \Pi A_{11}\Pi^T x_1(t) + \sum_{k=1}^{m} \Pi N_k \Pi^T x_1(t) u_k(t) + \Pi B_1 u(t), \tag{5a}$$

$$y(t) = \mathcal{C}\Pi^T x_1(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)} \Pi^T x_1(t) u_k(t) + \mathcal{D}u(t), \qquad x_1(0) = 0. \tag{5b}$$

As stated in [21], the dynamical system (5) evolves in the $n_1 - n_2$ dimensional subspace $\ker(\Pi)$. Therefore, with the decomposition of Π,

$$\Pi = \phi_1 \phi_2^T \tag{6a}$$

with $\phi_1, \phi_2 \in \mathbb{R}^{n_1 \times n_1 - n_2}$ assuring

$$\phi_1^T \phi_2 = I, \tag{6b}$$

the system is completely described through $\tilde{x}_1(t) = \phi_1^T x_1(t)$ which satisfies

$$\phi_2^T E_{11} \phi_2 \dot{\tilde{x}}_1(t) = \phi_2^T A_{11} \phi_2 \tilde{x}_1(t) + \sum_{k=1}^{m} \phi_2^T N_k \phi_2 \tilde{x}_1(t) u_k(t) + \phi_2^T B_1 u(t), \tag{7a}$$

$$y(t) = \mathcal{C}\phi_2 \tilde{x}_1(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)} \phi_2 \tilde{v} u(t) + \mathcal{D}u(t), \quad \tilde{x}_1(0) = 0. \tag{7b}$$

Thus, MOR of the system (7) is equivalent to MOR of the system (2). However, the advantage of the system (7) is that $\phi_2^T E_{11} \phi_2$ is nonsingular, allowing us to employ Algorithm 1, if bilinear terms are neglected in the output equation. This leads a locally $\mathcal{H}_2$-optimal reduced-order system, provided it converges. Unfortunately, to determine the system matrices of (7), we require the explicit computation of the basis matrix ϕ_2 which is not readily available. Moreover, it might also appear that the realization of the system (7) becomes dense after multiplication with ϕ_2 which makes the computation of the reduced-order systems expensive. To overcome this, in what follows, we show how to avoid the explicit computation of ϕ_2 in the application of *B-IRKA*.

Remark 3.1. *In this paper, we neglect the nonlinear terms and the control part in the output equation as far as the computation of the projection matrices is concerned. We focus on the linear relation between the state vector and the output. Nonetheless, the bilinear terms in the output equation are projected afterwards.*

Computational issues

We consider the following associated bilinear ODE system to compute the projection matrices $\mathcal{V}$ and $\mathcal{W}$:

$$\phi_2^T E_{11} \phi_2 \dot{\tilde{x}}_1(t) = \phi_2^T A_{11} \phi_2 \tilde{x}_1(t) + \sum_{k=1}^{m} \phi_2^T N_k \phi_2 \tilde{x}_1(t) u_k(t) + \phi_2^T B_1 u(t), \tag{8a}$$

$$\tilde{y}(t) = \mathcal{C}\phi_2 \tilde{x}_1(t), \quad \tilde{x}_1(0) = 0. \tag{8b}$$

In the view of resolving the computational issues, we aim to determine $\mathcal{V}$ and $\mathcal{W}$ such that the system matrices $E_{11}, A_{11}, N_k, B_1, \mathcal{C}$ and $\mathcal{C}_N$ can be directly reduced using the projection matrices as shown in Algorithm 2.

Algorithm 2 MOR for Bilinear DAEs, having an Index-2 Matrix Pencil (Involving Projector).

1: **Input:** $E_{11}, A_{11}, N_k, B_1, \mathcal{C}, \mathcal{C}_N^{(k)}$.
2: Make an initial guess of $\hat{E}, \hat{A}, \hat{N}_k, \hat{B}, \hat{\mathcal{C}}$.
3: **while** no convergence **do**
4: Compute nonsingular matrices Y and Z such that $Y\hat{A}Z = \Lambda$ and $Y\hat{E}Z = \hat{I}$.
5: Define $\tilde{B} = \hat{B}^T Y^T, \quad \tilde{C} = \hat{C}Z \quad$ and $\quad \tilde{N}_k = Z^T \hat{N}_k^T Y^T$.
6: Determine

$$L = -(\hat{I} \otimes \phi_2)\left(\Lambda \otimes (\phi_2^T E_{11}\phi_2) + \hat{I} \otimes (\phi_2^T A_{11}\phi_2) + \sum_{k=1}^{m} \tilde{N}_k^T \otimes (\phi_2^T N_k \phi_2)\right)^{-1} (\hat{I} \otimes \phi_2^T).$$

7: Determine the projection matrices $\mathcal{V}$ and $\mathcal{W}$:

$$\begin{aligned} \operatorname{vect}(\mathcal{V}) &= L(\tilde{B}^T \otimes B)\operatorname{vect}(I_m), \\ \operatorname{vect}(\mathcal{W}) &= L^T(\tilde{C}^T \otimes \mathcal{C}^T)\operatorname{vect}(I_p). \end{aligned}$$

8: Compute the reduced-order system matrices:

$$\begin{aligned} &\hat{E} = \mathcal{W}^T E_{11}\mathcal{V}, \qquad \hat{A} = \mathcal{W}^T A_{11}\mathcal{V}, \qquad \hat{N}_k = \mathcal{W}^T N_k \mathcal{V}, \\ &\hat{B} = \mathcal{W}^T B_1, \qquad \hat{C} = \mathcal{C}\mathcal{V}, \qquad \mathcal{C}_N^{(k)} = \hat{\mathcal{C}}_N^{(k)}\mathcal{V}. \end{aligned}$$

9: **end while**
10: **Output:**
$\hat{E}^{opt} = \hat{E}, \;\; \hat{A}^{opt} = \hat{A}, \;\; \hat{N}_k^{opt} = \hat{N}_k^{opt}, \;\; \hat{B}^{opt} = \hat{B}, \;\; \hat{\mathcal{C}}^{opt} = \hat{\mathcal{C}}, \;\; \hat{\mathcal{C}}_N^{(k)opt} = \hat{\mathcal{C}}_N^{(k)}$.

We notice that the projection matrices can be directly applied to the original system matrices, but in order to compute the projection matrices $\mathcal{V}$ and $\mathcal{W}$, we still require the matrix ϕ_2 explicitly. Therefore, our next goal is to construct the matrices $\mathcal{V}$ and $\mathcal{W}$ without resorting to ϕ_2.

Lemma 3.1. *Let ϕ_2 be the matrix as defined in* (6) *and $\mathcal{F}$ be a matrix such that $(\hat{I}\otimes\phi_2^T)\mathcal{F}(\hat{I}\otimes\phi_2)$ is invertible. Define $\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{F}}$ as follows:*

$$\begin{aligned} \mathcal{X}_{\mathcal{F}}^{\mathcal{I}} &:= (\hat{I} \otimes \phi_2)\Big((\hat{I} \otimes \phi_2^T)\mathcal{F}(\hat{I} \otimes \phi_2)\Big)^{-1}(\hat{I} \otimes \phi_2^T), \\ \mathcal{X}_{\mathcal{F}} &:= (\hat{I} \otimes \Pi)\mathcal{F}(\hat{I} \otimes \Pi^T), \end{aligned} \tag{9}$$

where Π is defined in (4)*. Then the matrices $\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{F}}$ satisfy the following relation:*

$$\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}\mathcal{X}_{\mathcal{F}} = (\mathcal{X}_{\mathcal{F}}\mathcal{X}_{\mathcal{F}}^{\mathcal{I}})^T = \hat{I} \otimes \Pi^T.$$

Proof. We begin with

$$\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}\mathcal{X}_{\mathcal{F}} = (\hat{I} \otimes \phi_2)\Big((\hat{I} \otimes \phi_2^T)\mathcal{F}(\hat{I} \otimes \phi_2)\Big)^{-1}(\hat{I} \otimes \phi_2^T)(\hat{I} \otimes \Pi)\mathcal{F}(\hat{I} \otimes \Pi^T).$$

We decompose $\Pi = \phi_1\phi_2^T$ and use properties of the Kronecker product to get

$$\begin{aligned} \mathcal{X}_{\mathcal{F}}^{\mathcal{I}}\mathcal{X}_{\mathcal{F}} &= (\hat{I} \otimes \phi_2)\Big((\hat{I} \otimes \phi_2^T)\mathcal{F}(\hat{I} \otimes \phi_2)\Big)^{-1}(\hat{I} \otimes \phi_2^T)(\hat{I} \otimes \phi_1)(\hat{I} \otimes \phi_2^T)\mathcal{F}(\hat{I} \otimes \Pi^T) \\ &= (\hat{I} \otimes \phi_2)\Big((\hat{I} \otimes \phi_2^T)\mathcal{F}(\hat{I} \otimes \phi_2)\Big)^{-1}(\hat{I} \otimes \phi_2^T\phi_1)(\hat{I} \otimes \phi_2^T)\mathcal{F}(\hat{I} \otimes \Pi^T). \end{aligned}$$

Since $\phi_2^T\phi_1 = I$ from (6), we obtain

$$\begin{aligned}\mathcal{X}_{\mathcal{F}}^I\mathcal{X}_{\mathcal{F}} &= (\hat{I}\otimes\phi_2)\Big((\hat{I}\otimes\phi_2^T)\mathcal{F}(\hat{I}\otimes\phi_2)\Big)^{-1}(\hat{I}\otimes\phi_2^T)\mathcal{F}(\hat{I}\otimes\phi_2)(\hat{I}\otimes\phi_1^T)\\ &= (\hat{I}\otimes\phi_2)(\hat{I}\otimes\phi_1^T) = \hat{I}\otimes\Pi^T.\end{aligned}$$

A similar argument can be given for the other equality. □

Using Lemma 3.1 and properties of the Kronecker product, we observe that the projection matrices $\mathcal{V}$ and $\mathcal{W}$, computed in step 7 of Algorithm 2, satisfy:

$$(\hat{I}\otimes\Pi)\mathcal{F}(\hat{I}\otimes\Pi^T)\,\mathrm{vect}(\mathcal{V}) = (\hat{I}\otimes\Pi)(\tilde{B}^T\otimes B), \tag{10a}$$

$$(\hat{I}\otimes\Pi)\mathcal{F}^T(\hat{I}\otimes\Pi^T)\,\mathrm{vect}(\mathcal{W}) = (\hat{I}\otimes\Pi)(\tilde{C}^T\otimes \mathcal{C}), \tag{10b}$$

where $\mathcal{F} = -\Big(\Lambda\otimes E_{11} + \hat{I}\otimes A_{11} + \sum_{k=1}^m \tilde{N}_k^T\otimes N_k\Big)$. Note that $\Pi_\otimes = \hat{I}\otimes\Pi$ is also an oblique projector. It can be verified that $(\Pi_\otimes)^2 = \Pi_\otimes$, $\ker(\Pi_\otimes) = \mathrm{range}\left(\hat{I}\otimes A_{12}\right)$, and $\mathrm{range}(\Pi_\otimes) = \ker\left(\hat{I}\otimes A_{12}^T E_{11}^{-1}\right)$. Using these properties, it can be shown that

$$(\hat{I}\otimes A_{12}^T)Z = 0 \qquad \text{if and only if} \qquad (\hat{I}\otimes\Pi^T)Z = Z. \tag{11}$$

In the following lemma, we show the way of circumventing the explicit computation of Π to solve (10) for $\mathrm{vect}(\mathcal{V})$ or $\mathrm{vect}(\mathcal{W})$ and reveal the connection between the solution of (10) and saddle point problems.

Lemma 3.2. *Consider $Z = (\hat{I}\otimes\Pi^T)Z$ and $(\hat{I}\otimes\Pi)\mathcal{F}(\hat{I}\otimes\Pi^T)Z = (\hat{I}\otimes\Pi)G$. Then, the matrix Z solves*

$$\begin{bmatrix}\mathcal{F} & \hat{I}\otimes A_{12}\\ \hat{I}\otimes A_{12}^T & 0\end{bmatrix}\begin{bmatrix}Z\\ \Xi\end{bmatrix} = \begin{bmatrix}G\\ 0\end{bmatrix}. \tag{12}$$

Proof. Since $Z = (\hat{I}\otimes\Pi^T)Z$ using the properties of $\hat{I}\otimes\Pi^T$ as stated in (11), we have $(\hat{I}\otimes A_{12}^T)Z = 0$. This implies that the second block of the equation (12) is satisfied.

Moreover, $(\hat{I}\otimes\Pi)\mathcal{F}Z - (\hat{I}\otimes\Pi)G = 0$ implies that the columns of $\mathcal{F}Z - G$ lie in $\ker\left(\hat{I}\otimes\Pi\right)$ = $\mathrm{range}\left(\hat{I}\otimes A_{12}\right)$. Therefore, there exists Ξ, satisfying $\mathcal{F}Z - G = -(\hat{I}\otimes A_{12})\Xi$ which is nothing but the first block of the equation (12). This concludes the proof. □

Therefore, using Lemma 3.2, we can determine $\mathrm{vect}(\mathcal{V})$ and $\mathrm{vect}(\mathcal{W})$ without explicitly computing Π by solving the corresponding saddle point problems. All these theoretical analysis gives rise to Algorithm 3 for MOR of the system (7).

Remark 3.2. *As discussed in [21], the general $B_2 \neq 0$ index-2 problems can be brought back to a problem with $B_2 = 0$ type by decomposing $x_1(t)$ as follows:*

$$x_1(t) = x_0(t) + x_u(t), \tag{13}$$

Algorithm 3 MOR for Bilinear DAEs, having an Index-2 Matrix Pencil.

1: **Input:** $E_{11}, A_{11}, N_k, B_1, \mathcal{C}, \mathcal{C}_N^{(k)}$
2: Make an initial choice of $\hat{E}, \hat{A}, \hat{N}_k, \hat{B}, \hat{C}$.
3: **while** no convergence **do**
4: Compute nonsingular matrices Y and Z such that $Y\hat{A}Z = \Lambda$ and $Y\hat{E}Z = \hat{I}$.
5: Define $\tilde{B} = \hat{B}^T Y^T, \quad \tilde{C} = \hat{C}Z \quad$ and $\quad \tilde{N}_k = Z^T \hat{N}_k^T Y^T$.
6: Determine the projection matrices $\mathcal{V}$ and $\mathcal{W}$:

$$\begin{bmatrix} \mathcal{F} & \hat{I} \otimes A_{12} \\ \hat{I} \otimes A_{12}^T & 0 \end{bmatrix} \begin{bmatrix} \operatorname{vect}(\mathcal{V}) \\ \Gamma \end{bmatrix} = \begin{bmatrix} (\tilde{B}^T \otimes B)\operatorname{vect}(I_m) \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \mathcal{F}^T & \hat{I} \otimes A_{12} \\ \hat{I} \otimes A_{12}^T & 0 \end{bmatrix} \begin{bmatrix} \operatorname{vect}(\mathcal{W}) \\ \Delta \end{bmatrix} = \begin{bmatrix} (\tilde{C}^T \otimes \mathcal{C})\operatorname{vect}(I_p) \\ 0 \end{bmatrix},$$

where $\mathcal{F} = -(\Lambda \otimes E_{11} + \hat{I} \otimes A_{11} + \sum_{k=1}^m \tilde{N}_k^T \otimes N_k)$.
7: Compute the reduced-order system matrices:

$$\hat{E} = \mathcal{W}^T E_{11} \mathcal{V}, \qquad \hat{A} = \mathcal{W}^T A_{11} \mathcal{V}, \qquad \hat{N}_k = \mathcal{W}^T N_k \mathcal{V},$$
$$\hat{B} = \mathcal{W}^T B_1, \qquad \hat{C} = \mathcal{C}\mathcal{V}, \qquad \mathcal{C}_N^{(k)} = \hat{\mathcal{C}}_N^{(k)} \mathcal{V}.$$

8: **end while**
9: **Output:**
$\hat{E}^{opt} = \hat{E}, \;\; \hat{A}^{opt} = \hat{A}, \;\; \hat{N}_k^{opt} = \hat{N}_k, \;\; \hat{B}^{opt} = \hat{B}, \;\; \hat{\mathcal{C}}^{opt} = \hat{\mathcal{C}}, \;\; \hat{\mathcal{C}}_N^{(k)opt} = \hat{\mathcal{C}}_N^{(k)}$.

where $x_u(t) = -\underbrace{E_{11}^{-1}A_{12}(A_{12}^T E_{11}^{-1} A_{12})B_2}_{\Upsilon} u(t)$ *and* $x_0(t)$ *satisfies* $A_{12}^T x_0(t) = 0$. *After doing the algebraic calculations as done for the case* $B_2 = 0$ *case, we get*

$$\Pi E_{11}\Pi^T \dot{x}_0(t) = \Pi A_{11}\Pi^T x_0(t) + \sum_{k=1}^m \Pi N_k \Pi^T x_0(t) u_k(t) + \Pi \mathcal{B}\tilde{u}(t), \tag{14a}$$

$$\Pi^T x_0(0) = \Pi^T (x_0 - x_u(0)), \tag{14b}$$

$$y(t) = \mathcal{C}\Pi^T x_0(t) + \sum_{k=1}^m \mathcal{C}_N^{(k)} \Pi^T x_0(t) u_k(t) + \mathcal{D}\tilde{u}(t) - C_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} B_2 \dot{u}(t), \tag{14c}$$

where

$$\mathcal{B} = [\mathcal{B}_1, \mathcal{B}_u^{(1)}, \ldots, \mathcal{B}_u^{(m)}] \;\; \textit{with} \;\; \mathcal{B}_u^{(k)} = -N_i\Upsilon, \qquad \tilde{u}(t) = \left([1, u(t)^T] \otimes u(t)^T\right)^T,$$

$$\mathcal{C} = C_1 - C_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{11}, \qquad \mathcal{C}_N^{(k)} = -C_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1} N_k,$$

$$\mathcal{D} = \Big[D - C_1\Upsilon - C_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1} \mathcal{B}_1, \; C_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1} [B_u^{(1)}, \ldots, B_u^{(m)}]\Big].$$

Although the system (14) *has terms associated with* u, $u{\cdot}u_k$ *which are functions of* u *eventually, but we treat them as different inputs of the system as far as MOR is considered. Now, it can be easily seen that determining reduced-order systems of the system* (14) *is analogous to the system* (5). *Therefore, Algorithm 3 can be readily applied to the system* (14) *to obtain locally* $\mathcal{H}_2$*-optimal reduced-order systems, having neglected bilinear terms in the output equation.*

4 Carleman Bilinearization for Quadratic-Bilinear Descriptor Systems with Index-2 Matrix Pencil

In this section, we study the Carleman bilinearization process for quadratic-bilinear descriptor systems. The Carleman bilinearization process can be applied to approximate quadratic-bilinear systems; see, e.g., [8, 19]. This process was recently extended to quadratic-bilinear descriptor systems, having an index-1 matrix pencil $\lambda E - A$ in [20]. For simplicity of notation, we consider a single-input quadratic-bilinear descriptor system in the following form:

$$\begin{aligned} E_{11}\dot{x}_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + Hx_1(t) \otimes x_1(t) + Nx_1(t)u(t) + B_1u(t), \\ 0 &= A_{21}x_1(t) + B_2u(t), \\ y(t) &= C_1x_1(t) + C_2x_2(t), \end{aligned} \tag{15}$$

where $x_1(t) \in \mathbb{R}^{n_1}$ and $x_2(t) \in \mathbb{R}$ are the generalized states, and $y(t) \in \mathbb{R}^p$ and $u(t) \in \mathbb{R}$ are the output and the input to the system, respectively. All other matrices are fixed by the size of the state vectors, the input and the output. We assume that matrices E_{11} and $A_{21}E_{11}^{-1}A_{12}$ are invertible. This implies that the matrix pencil $\lambda E - A$ is index-2 pencil. Moreover, it is assumed that the system has only one constraint due to which $x_2(t)$ is considered to be a scaler variable, rather than a vector. With these assumptions, we aim to determine an approximate bilinear system via Carleman bilinearization, preserving the properties like the index of the matrix pencil $\lambda E - A$. Without loss of generality, we can assume that $B_2 = 0$, otherwise it can be brought back to the $B_2 = 0$ case by an appropriate change of variables as stated in Remark 3.2. As a first step, we develop a differential equation for $x_1^{\otimes} = x_1(t) \otimes x_1(t)$, neglecting cubic and higher order terms:

$$\begin{aligned} (E_{11} \otimes E_{11})\frac{d}{dt}x_1^{\otimes}(t) &= E_{11}\dot{x}_1(t) \otimes E_{11}x_1(t) + E_{11}x_1(t) \otimes E_{11}\dot{x}_1(t) \\ &= \mathcal{L}(A_{11}, E_{11})x_1^{\otimes}(t) + \mathcal{L}(A_{12}, E_{11})x_1(t)x_2(t) \\ &\quad + \mathcal{L}(B_1, E_{11})x_1(t)u(t) + \mathcal{L}(N, E_{11})x_1^{\otimes}u(t) \end{aligned} \tag{16}$$

in which $\mathcal{L}(\mathcal{A}, \mathcal{B}) = \mathcal{A} \otimes \mathcal{B} + \mathcal{B} \otimes \mathcal{A}$. We also get additional constraints as

$$0 = (E_{11} \otimes A_{21})x_1^{\otimes}(t). \tag{17}$$

Combining (15),(16) and (17), we get the following bilinear system as an approximation to (15):

$$\begin{aligned} \tilde{E}_{11}\dot{\tilde{x}}_1(t) &= \tilde{A}_{11}\tilde{x}_1(t) + \tilde{A}_{12}\tilde{x}_2(t) + \tilde{N}\tilde{x}_1(t)u(t) + \tilde{B}_1(t), \\ 0 &= \tilde{A}_{21}\tilde{x}_1(t), \\ \tilde{y}(t) &= \tilde{C}_1\tilde{x}_1(t) + \tilde{C}_2\tilde{x}_2(t), \end{aligned}$$

where

$$\tilde{E}_{11} = \begin{bmatrix} E_{11} & 0 \\ 0 & \mathcal{L}(E_{11}, E_{11}) \end{bmatrix}, \quad \tilde{A}_{11} = \begin{bmatrix} A_{11} & H \\ 0 & \mathcal{L}(A_{11}, E_{11}) \end{bmatrix}, \quad \tilde{A}_{12} = \begin{bmatrix} A_{12} & 0 \\ 0 & \mathcal{L}(A_{12}, E_{11}) \end{bmatrix},$$

$$\tilde{A}_{21} = \begin{bmatrix} A_{21} & 0 \\ 0 & (A_{21} \otimes E_{11}) \end{bmatrix}, \quad \tilde{N} = \begin{bmatrix} N & 0 \\ \mathcal{L}(B_1, E_{11}) & \mathcal{L}(N, E_{11}) \end{bmatrix},$$

$$\tilde{B}_1 = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad \hat{C}_1 = \begin{bmatrix} C_1^T \\ 0 \end{bmatrix}^T, \quad \tilde{C}_2 = \begin{bmatrix} C_2^T \\ 0 \end{bmatrix}^T, \quad \tilde{x}_1 = \begin{bmatrix} x_1 \\ x_1^{\otimes} \end{bmatrix}, \quad \tilde{x}_2 = \begin{bmatrix} x_2 \\ x_1x_2 \end{bmatrix}.$$

Next, we show that the matrix pencil $\lambda\tilde{E} - \tilde{A}$ is also an index-2 pencil. For this, we need to prove that $\tilde{A}_{21}\tilde{E}_{11}^{-1}\tilde{A}_{12}$ is a full rank matrix. Therefore, we consider the following:

$$\begin{aligned}\tilde{A}_{21}\tilde{E}_{11}^{-1}\tilde{A}_{12} &= \begin{bmatrix} A_{21} & 0 \\ 0 & (A_{21}\otimes E_{11}) \end{bmatrix} \begin{bmatrix} E_{11}^{-1} & 0 \\ 0 & \mathcal{L}(E_{11}^{-1}, E_{11}^{-1}) \end{bmatrix} \begin{bmatrix} A_{12} & 0 \\ 0 & \mathcal{L}(A_{12}, E_{11}) \end{bmatrix} \\ &= \begin{bmatrix} A_{21}E_{11}^{-1}A_{12} & 0 \\ 0 & E_{11}\left(\left(A_{21}E_{11}^{-1}A_{12}\right) I + E_{11}^{-1}A_{12}A_{21}\right) \end{bmatrix}.\end{aligned}$$

Since $E_{11}^{-1}A_{12}$ and A_{21} both are vectors, this implies that the eigenvalues of $E_{11}^{-1}A_{12}A_{21}$ are zero and $\left(A_{21}E_{11}^{-1}A_{12}\right)$ with multiplicities $n-1$ and 1, respectively. Also, we know that if σ_i are the eigenvalues of matrix P, then the matrix $Q := I + P$ has the eigenvalues $1+\sigma_i$. This shows that the matrix $\left(\left(A_{21}E_{11}^{-1}A_{12}\right) I + E_{11}^{-1}A_{12}A_{21}\right)$ has eigenvalues $A_{21}E_{11}^{-1}A_{12}$ and $2\cdot A_{21}E_{11}^{-1}A_{12}$ with multiplicities $n-1$ and 1, respectively. Hence, $\tilde{A}_{21}\tilde{E}_{11}^{-1}\tilde{A}_{12}$ is invertible and so, the bilinearized system has an index-2 matrix pencil.

5 Numerical Experiments

In this section, we investigate the efficiency of the proposed iterative algorithm for bilinear descriptor systems, having an index-2 matrix pencil and compare the quality of the determined reduced-order systems with the ones obtained by using the projection matrices determined by linear *IRKA* [14, Algo. 6.2]. The stopping criterion for Algorithm 3 is based on the relative change in the eigenvalues of the reduced-order system. If this change is below the square root of the machine precision, then the iteration is stopped. We randomly select the initial guess of the reduced-order matrices in Algorithm 3, and also choose a scaling factor γ as suggested in [7] for a smooth convergence of *B-IRKA*. All the simulations are done on a CPU 2x Intel Xeon E5620, 12 MB Cache, 48 GB DDR3 RAM, MATLAB® Version 7.11.0.584 (R2010b) 64-bit(glnxa64).

5.1 Nonlinear RC Circuit

Nonlinear transmission line circuits are considered to be standard test cases the community of interpolation-based MOR; see, e.g., [22] for the unconstrained RC circuit and [15] for the constraint RC circuit. Here, we consider a variant of the constraint transmission line circuit as shown in Figure 1 (in Introduction), where it is assumed that the voltages at the first and last nodes are the same. The electrical component i.e., diode I-V, has nonlinear characteristics $g(v_D) = e^{40v_D} + v_D - 1$, where v_D is the voltage across the node. Using Kirchhoff's current law at each node, we get the following set of equations:

$$\begin{aligned}\dot{v}_1 &= -2v_1 + v_2 + 2 - e^{40v_1} - e^{40(v_1 - v_2)} + u(t),\\ \dot{v}_i &= -2v_i + v_{i-1} + v_{i+1} + e^{40(v_{i-1}-v_i)} - e^{40(v_i - v_{i+1})}, \quad 1 < i < \tilde{n},\\ \dot{v}_{\tilde{n}} &= -v_{\tilde{n}} + v_{\tilde{n}-1} - 1 + e^{40(v_{\tilde{n}-1}-v_{\tilde{n}})}\end{aligned} \tag{18}$$

with a constraint

$$0 = v_1 - v_{\tilde{n}}.$$

The system of equations (18) can be written as a quadratic-bilinear descriptor system by appropriately introducing the new state variables, as shown in [22] for a nonlinear RC circuit example. The dynamics of the system, in state-space representation, is given as follows:

$$\begin{aligned}\dot{x}(t) &= Ax(t) + G\lambda + Hx(t)\otimes x(t) + Nx(t)u(t) + Bu(t),\\ 0 &= G^T x(t),\end{aligned}$$

where $x(t) \in \mathbb{R}^{\tilde{n}}$ and λ are state vector, containing voltages at each node, and an appropriate Lagrangian multiplier, respectively. We observe the voltage at the first node. Since we have only one constraint in the system dynamics, this allows us to employ Carleman bilinearization as discussed in Section 4 to obtain an approximate bilinearized descriptor system. We set $\tilde{n} = 15$, leading to a bilinearized system of order $n = 2 \cdot \tilde{n} + 4 \cdot \tilde{n}^2 + 1 = 961$. We apply the $\mathcal{H}_2$-optimal model reduction method given in Algorithm 3 by setting the order of the reduced-order system to $r = 10$. We choose the scaling factor $\gamma = 0.01$ in order to achieve convergence of *B-IRKA*. We also determine a reduced-order system by using linear *IRKA* of the same order. In Figure 2, we compare the quality of the reduced-order systems with the original system by computing transient responses for an input $u(t) = (\sin(10\pi t) + 1)/2$.

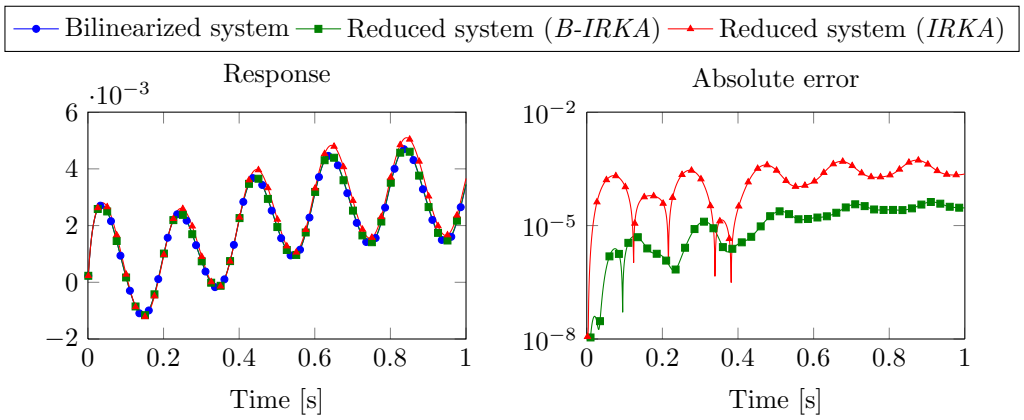


Figure 2: Comparison of the transient response of the systems for an input $u(t) = (\sin(10\pi t) + 1)/2$.

We observe that the reduced-order system obtained by modified *B-IRKA* captures the dynamics of the system better as compared to the reduced-order system obtained by linear *IRKA*.

5.2 Resistance-varying RC circuit

As our second example, we consider the RC circuit as shown in Figure 3 in which the ith node is connected to the $(i-1)$st and the $(i+1)$st nodes via resistances, and connected to the ground via capacitors. Moreover, the first node is connected to the ground via a variable resistance, and the voltage at the first node is influenced by the current (the input u_1). We also add an extra control u_2, controlling the voltage difference between the first and last nodes. Now, we apply Kirchhoff's current law at each node to obtain the following set of ODEs:

$$\begin{aligned} C\dot{v}_1(t) &= \tfrac{1}{R}(-v_1 + v_2) + \tfrac{1}{R_v}(0 - v_1) + u(t), \\ C\dot{v}_i(t) &= \tfrac{1}{R}(-2v_i + v_{i-1} + v_{i+1}), \qquad (2 \leq i \leq n-1), \\ C\dot{v}_n(t) &= \tfrac{1}{R}(-v_n + v_{n-1}) \end{aligned}$$

along with a constraint

$$0 = v_1 - v_n - u_2(t).$$

We set all the capacitors (C) and the constant resistance (R) equal to 1, and consider that the variable resistance R_v varies with respect to the parameter δ as follows:

$$R_v = \frac{R}{1+\delta}.$$

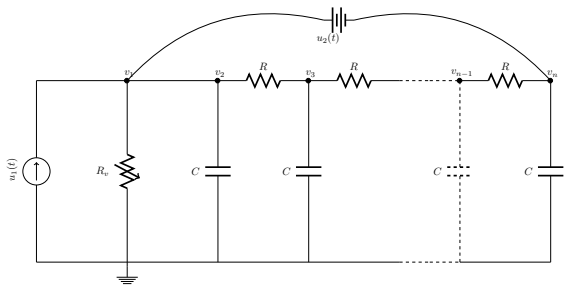


Figure 3: A constraint RC-circuit diagram.

Combining all these equations together, we obtain the dynamics of the RC circuit which are described by the following descriptor system:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + G^T\lambda(t) + \delta Nx(t) + B_1u_1(t), \\ 0 &= Gx(t) + B_2u_2(t), \\ y(t) &= Cx(t), \end{aligned} \tag{19}$$

where $x(t) \in \mathbb{R}^n$ is the state vector containing the voltage at each node, $\lambda \in \mathbb{R}$ is the Lagrange multiplier, $G = [1, 0, \ldots, 0, -1]$ is a constraint matrix, $B_1 = [1, 0, \ldots, 0]$ and $B_2 = 1$. The voltage at the second node is the output of interest, and it yields $C = [0, 1, 0, \ldots, 0]$. For this example, we first transform the system (19) into an equivalent system with $\mathcal{B}_2 = 0$ as suggested in Remark 3.2, leading to the following system:

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) + \delta N\tilde{x}(t) + G^T\lambda(t) + \mathcal{B}\tilde{u}(t), \\ 0 &= G\tilde{x}(t), \\ y(t) &= C\tilde{x}(t) + D\tilde{u}(t), \end{aligned} \tag{20}$$

where $\mathcal{B} = [B_1,\ A\mathcal{G},\ N\mathcal{G}]$ and $D = [0,\ C\mathcal{G},\ 0]$ in which $\mathcal{G} = -G^T(GG^T)^{-1}B_2$, and $\tilde{u}(t) = [u_1(t),\ u_2(t),\ \delta u_2(t)]$. Now, the system (20) can be seen as a linear parameter-varying system in the parameter δ. It is shown in [23] that the special class of parametric systems is closely related to MOR of bilinear systems. Therefore, we reformulate the linear system (20) appropriately as a bilinear system with four inputs and one output as follows:

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) + \sum_{i=1}^{4} N_i\tilde{x}(t)u_i(t) + G^T\lambda(t) + \mathcal{B}_b\tilde{u}_b(t), \\ 0 &= G\tilde{x}(t), \\ y(t) &= C\tilde{x}(t) + D_b\tilde{u}_b(t), \end{aligned} \tag{21}$$

where $\begin{bmatrix} N_1, N_2, N_3, N_4 \end{bmatrix} = \begin{bmatrix} 0, 0, 0, N \end{bmatrix}$, $\mathcal{B}_b = [\mathcal{B}, 0]$ and $D_b = [D, 0]$ with inputs $\tilde{u}_b(t) = [\tilde{u}^T(t), \delta]^T$. We consider $n = 1000$, leading to the order of the system $\tilde{n} = 1001$. Next, we determine reduced bilinear systems of order $r = 15$ by employing the $\mathcal{H}_2$-optimal model reduction method given in Algorithm 3 and by using linear ***IRKA***. These reduced bilinear systems again can be rewritten into reduced linear parametric systems. This allows us to determine the quality of the reduced-order systems by comparing the relative H_∞-norm $\frac{\|H-\hat{H}\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$ of the error system varying parameter values δ which is shown in Figure 4.

We observe that the reduced parametric system obtained from ***B-IRKA*** captures the dynamics of the original system for a wide parameter range much better as compared to the one obtained

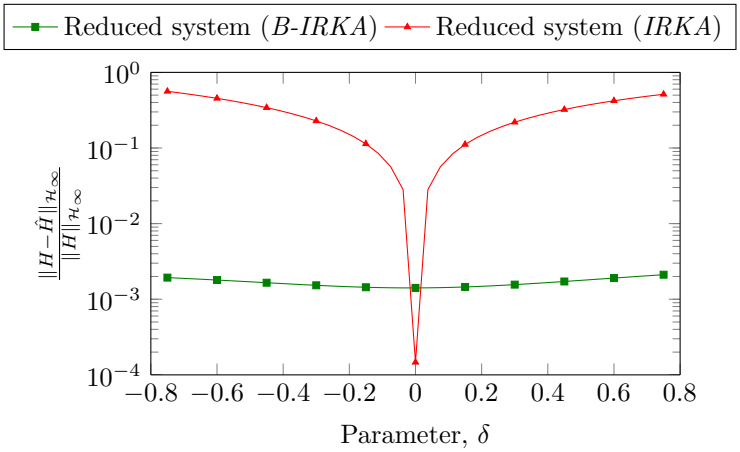


Figure 4: Relative H_∞ error versus the parameter δ for the reduced linear parametric systems obtained from *B-IRKA* and *IRKA*.

by using linear *IRKA*. However, one can see the drop in the relative H_∞ error in Figure 4. This is due to an obvious reason that the projection matrices obtained by employing *IRKA* capture the dynamics of the system quite accurately for $\delta = 0$, but fail to capture the dynamics of the system as the parameter δ moves away from $\delta = 0$. On the contrary the reduced parametric system obtained from *B-IRKA* performs quite well over a wide parameter range.

5.3 Parameter dependant RLC circuit

Lastly, we consider an RLC circuit as shown in Figure 5. The governing equations of the RLC circuit can be written as follows:

$$\begin{aligned} C\tfrac{d}{dt}v_j &= i_j - i_{j-1}, \quad j = 1, \ldots, g-1, \\ C\tfrac{d}{dt}v_g &= i_g, \\ L\tfrac{d}{dt}i_j + Ri_j &= v_{j-1} - v_j \quad j = 2, \ldots, g, \end{aligned}$$

where v_j and i_j are the voltage at the jth node and the current passing through the $(j-1)$st inductor, respectively. Also, $V(t)$ is a control voltage source of the system and i_1 is the current passing through this voltage source. Since the voltage source is connected to the first node via ground, this leads to a constraint $0 = v_1 - V(t)$. We set all capacitors and inductors to 1, and consider variable resistances, depending linearly on the parameter p as follows:

$$R = 1 + p.$$

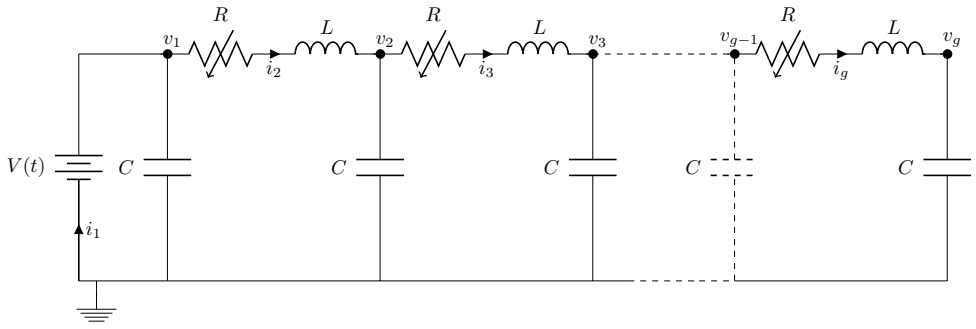


Figure 5: A variable resistance RLC circuit.

With these relations, we can write the system in state-space form as:

$$\begin{aligned} \frac{d}{dt}x_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + pNx_1(t), \\ 0 &= A_{12}^T x_1(t) + B_2 u(t), \end{aligned} \tag{22}$$

where $x_1(t)$ contains the voltages at each node and the currents passing through each inductor, and $x_2(t)$ contains the current though the voltage source. The voltage at the last node is observed. We choose $g = 500$ which results in the order of the system (22) $n = 1001$. As a first step, we convert system (22) to an equivalent system by using an appropriate change of the state variable so that the constraint equation becomes independent of the input, leading to the following system:

$$\begin{aligned} \frac{d}{dt}\tilde{x}_1(t) &= A_{11}\tilde{x}_1(t) + \tilde{A}_{12}x_2(t) + pN\tilde{x}_1(t) + \tilde{B}u(t), \\ 0 &= A_{12}^T\tilde{x}_1(t), \\ y(t) &= C_1\tilde{x}_1(t). \end{aligned}$$

Next, we treat the above system as a bilinear system by considering the parameter p as an input to the system. We determine reduced bilinear systems of order $r = 10$, by employing the proposed *B-IRKA* and *IRKA*, and then convert back to have linear parametric reduced-order systems. In order to compare the quality of the reduced-order systems, we plot the relative H_∞-norm of the error system in Figure 6.

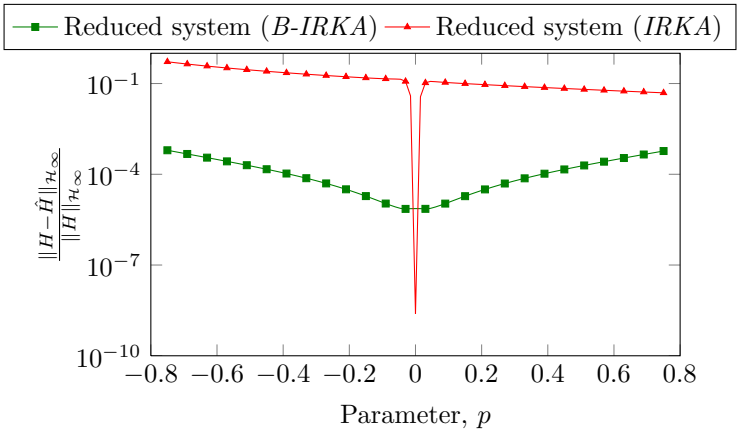


Figure 6: Relative H_∞ error versus the parameter p for the reduced linear parametric systems obtained from *B-IRKA* and *IRKA*.

The similar phenomenon, as observed in Example 5.2, can be seen in Figure 6, in particular a drop in the relative H_∞ error for *IRKA* at $p = 0$. Nevertheless, the reduced-order system, obtained by using *B-IRKA*, outperforms the one obtained by using *IRKA* for a wide range of the parameter.

6 Conclusions

We have proposed an iterative algorithm for MOR of bilinear descriptor systems, having an index-2 matrix pencil. This gives rise to locally $\mathcal{H}_2$-optimal reduced-order systems on convergence, if it converges. First we have transformed the original bilinear descriptor system into

an equivalent bilinear ODE system by means of projectors. This enabled us to employ bilinear iterative rational Krylov algorithm *(B-IRKA)*. Next, in the view of implementation, we have proposed a modified *B-IRKA* which does not require the undesirable explicit computation of the spectral projector in order to compute reduced-order systems. We have also extended the Carleman bilinearization process for quadratic-bilinear descriptor systems, having an index-2 matrix pencil and only one constraint. It is shown that the bilinearized systems preserve the index-2 of the matrix pencil. Finally, we have illustrated the efficiency of the proposed *B-IRKA* using various constraint electrical circuit examples, showing that reduced-order systems, obtained by using modified *B-IRKA*, replicate the dynamics of the original system much better as compared to the reduced-order systems obtained by using linear *IRKA*.

REFERENCES

[1] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[2] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM Publications, Philadelphia, PA, 2005.

[3] P. Benner, V. Mehrmann, and D.C. Sorensen. *Dimension Reduction of Large-Scale Systems*, volume 45 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin/Heidelberg, Germany, 2005.

[4] W. H. A. Schilders, H. A. Van der Vorst, and J. Rommes. *Model Order Reduction: Theory, Research Aspects and Applications*. Springer-Verlag, Berlin, Heidelberg, 2008.

[5] P. Benner and T. Damm. Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM J. Cont. Optim.*, 49(2):686–711, 2011.

[6] Z. Bai. Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Appl. Numer. Math*, 43(1–2):9–44, 2002.

[7] P. Benner and T. Breiten. Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.*, 33(3):859–885, 2012.

[8] T. Breiten and T. Damm. Krylov subspace methods for model order reduction of bilinear control systems. *Sys. Control Lett.*, 59(10):443–450, 2010.

[9] J. R. Phillips. Projection-based approaces for model reduction of weakly nonlinear, time-varying systems. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 22(2):171–187, 2003.

[10] P. Benner and T. Breiten. Low rank methods for a class of generalized Lyapunov equations and related issues. *Numerische Mathematik*, 124(3):441–470, 2013.

[11] S. D. Shank, V. Simoncini, and D. B Szyld. Efficient low-rank solution of generalized Lyapunov equations. *Numerische Mathematik*, XY:1–16, 2014.

[12] L. Zhang and J. Lam. On H_2 model reduction of bilinear systems. *Automatica J. IFAC*, 38(2):205–216, 2002.

[13] D. A. Wilson. Optimum solution of model-reduction problem. *Proc. IEEE*, 117:1161–1165, 1970.

[14] S. Gugercin, T. Stykel, and S. Wyatt. Model reduction of descriptor systems by interpolatory projection methods. *SIAM J. Sci. Comput.*, 35(5):B1010–B1033, 2013.

[15] P. Goyal, M. I. Ahmad, and P. Benner. Krylov subspace-based model reduction for a class of bilinear descriptor systems. Preprint MPIMD/15-07, Max Planck Institute Magdeburg, 2015. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[16] P. Benner and P. Goyal. Multipoint interpolation of Volterra series and $\mathcal{H}_2$-model reduction for a family of bilinear descriptor systems. Preprint MPIMD/15-16, Max Planck Institute Magdeburg, 2015. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[17] M. I. Ahmad, P. Benner, P. Goyal, and J. Heiland. Moment-matching based model reduction for Stokes-type quadratic-bilinear descriptor systems. Preprint MPIMD/15-18, Max Planck Institute Magdeburg, 2015. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[18] A. Bruns and P. Benner. Parametric model order reduction of thermal models using the bilinear interpolatory rational Krylov algorithm. *Math. Comput. Model. Dyn. Syst.*, 21(2):103–129, 2015.

[19] W. J. Rugh. *Nonlinear System Theory*. Johns Hopkins University Press, Baltimore, 1981.

[20] P. Goyal, M. I. Ahmad, and P. Benner. Model reduction of quadratic-bilinear descriptor systems via Carleman bilinearization. In *Proc. European Control Conference*, pages 1177–1182, 2015.

[21] M. Heinkenschloss, D. C. Sorensen, and K. Sun. Balanced truncation model reduction for a class of descriptor systems with applications to the Oseen equations. *SIAM J. Sci. Comput.*, 30(2):1038–1063, 2008.

[22] C. Gu. QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 30(9):1307–1320, 2011.

[23] P. Benner and T. Breiten. On $\mathcal{H}_2$-model reduction of linear parameter-varying systems. In *Proc. Appl. Math. Mech.*, volume 11, pages 805–806, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ELECTRICAL FIELDS SIMULATION IN HETEROGENOUS DOMAINS USING THE PROPER GENERALIZED DECOMPOSITION

Chady Ghnatios[1,2], Francisco Chinesta[2] and Anais Barasinski[2]

[1] Notre Dame University-Louaize
P.O. Box 72, Zouk Mosbeh, Lebanon
e-mail: cghnatios@ndu.edu.lb

[2] Ecole Centrale Nantes
1 rue de la No, 44300 Nantes, France
e-mail: {francisco.chinesta,anais.barasinski}@ec-nantes.fr

Keywords: Proper Generalized decomposition, Electric fields, domain separation.

Abstract.

Recently microwave heating is replacing classical methods in composite manufacturing processes. In fact one of the advantages of microwave heating is achieving a volumetric heating, which can improve the final material properties. However electrical and magnetic fields propagation is complicated to simulate and understand inside a heterogeneous domain. In fact, carbon fibers are highly conductor with respect to the composite matrix. Such behavior would alter the wave propagation inside the part and reduces the desired volumetric heating effects. In this work, we simulate the propagation of electric fields inside a heterogeneous part by using the Proper Generalized Decomposition (known as PGD). In fact this method helps us achieving full 3D simulations by the complexity of few 1D simulations.

1 INTRODUCTION

Composite parts are getting popular thanks to their good mechanical properties and low weight. Aeronautic and transport industries are turning to composite materials to reduce the energy use and therefore the transportation costs.

However composite materials manufacturing processes are not fulfilling industrial needs yet. In fact, classical manufacturing techniques can't follow up the increasing industrial demand of faster and reliable production. Therefore, researchers are turning their attention to innovative production techniques such as the microwave curing of composite parts. Moreover, microwave curing can lead to better parts with mechanical properties similar to the autoclave-made parts [1], thanks to the volumetric heating achieved by this method [2]. On the other hand, volumetric curing can also shorten the total manufacturing process time [3]. Such advantages are leading the industrials to microwave heating of thermoplastic as well as thermoset polymers.

On the other hand, microwave heating relies on the propagation of an electromagnetic field inside composite materials. However, such materials are heterogeneous, and usually with a degenerated thickness dimension. Moreover, we note that carbon fibers are electrical conductors, while, in general, the composite matrix is highly insulating. Thus, the electrical field behavior in such conditions is still not mastered yet to the knowledge of the authors. Therefore a full 3D simulation is required to analyze the wave propagation inside the composite material and understand the microwave heating process. Despite the large progress in computing power during the last decade, such simulations are still difficult to perform with the required resolution using classical calculation techniques. In fact the simulation of the wave propagation in the composite material is highly dependent on the degenerated thickness direction.

Thus, model order reduction techniques appeared recently as a replacement of classical numerical methods, which allow faster simulation with acceptable accuracy [4]. The Proper Generalized Decomposition or PGD is one of these techniques which allows overcoming the curse of dimensionality by using separated representation of the domain [4, 5]. In fact this technique allows the simulation of a full 3D models with a complexity of few 2D or 1D simulations [6, 7].

The objective of this work is to simulate the electrical field propagation using a full 3D model in an heterogeneous domain, containing carbon fibers inside a thermoplastic matrix. Using the PGD, the simulation will be separated in a sequence of three 1D simulations which will lead to the full 3D solution [8]. We will be showing the results inside two different domains. In the first one, the domain dimensions are large with respect to the wavelength, while in the second one, the domain dimensions are small with respect to the wavelength.

2 SIMULATED MODEL

In this section we describe the model used in the simulation. As illustrated in figure 1, we have a matrix domain in which a conductor which has a rectangular cross section is placed. The conductor is at $\frac{3}{5}b$ from the bottom of the domain and centered in the (x, z) plane. The cross section of the conductor has the dimensions $\frac{1}{5}b \times 0.05c$.

The physical model consists of the Maxwell equations in the phasor domain, that after some

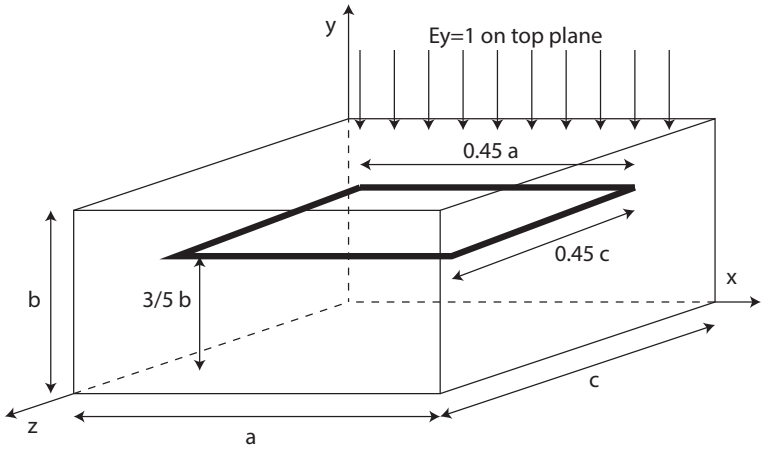


Figure 1: The model used in the simulation

manipulations and assuming an harmonic electric field, is reduced to:

$$\nabla^2 \mathbf{E} - \gamma^2 \mathbf{E} = 0 \tag{1}$$

where $\mathbf{E}$ is the electric field in the phasor domain having 3 components denoted as $\mathbf{E} = (E_x; E_y; E_z)$. While γ is defined by:

$$\gamma^2 = i \cdot w \cdot \mu \left(\sigma + i \cdot w \cdot \epsilon\right) \tag{2}$$

where i being the complex number, σ is the electrical conductivity, μ is the electrical permeability, ϵ is the permittivity and ω is the angular frequency of the wave. In our simulations we use a frequency of 2.45 GHz, which has a wavelength of about 12 cm. The boundary conditions of the simulated domain can be written by:

$$\begin{aligned}
&E_y = 0 \text{ at } y = 0;\\
&E_y = 1 \text{ at } y = b;\\
&E_x(x=0) = E_x(x=a);\\
&E_x(z=0) = E_x(z=c);\\
&\tfrac{\partial E_x}{\partial x}(x=0) = \tfrac{\partial E_x}{\partial x}(x=a);\\
&\tfrac{\partial E_x}{\partial z}(z=0) = \tfrac{\partial E_x}{\partial z}(z=c);\\
&E_z(x=0) = E_z(x=a);\\
&E_z(z=0) = E_z(z=c);\\
&\tfrac{\partial E_z}{\partial x}(x=0) = \tfrac{\partial E_z}{\partial x}(x=a);\\
&\tfrac{\partial E_z}{\partial z}(z=0) = \tfrac{\partial E_z}{\partial z}(z=c);\\
&E_y(x=0) = E_y(x=a);\\
&E_y(z=0) = E_y(z=c);\\
&\tfrac{\partial E_y}{\partial x}(x=0) = \tfrac{\partial E_y}{\partial x}(x=a);\\
&\tfrac{\partial E_y}{\partial z}(z=0) = \tfrac{\partial E_y}{\partial z}(z=c);
\end{aligned} \tag{3}$$

The boundary conditions depicted in equation 3, can be resumed by an imposed electrical field on the y direction and symmetrical boundary conditions on the x and z directions.

In table 1 we indicate the considered parameters for both the matrix and the fibers (conductor) in the domain.

Material	conductivity	permeability	permittivity
Conductor	inf	μ_0	ϵ_0
Matrix	0	μ_0	ϵ_0

Table 1: Parameters used in the simulation of the proposed model

3 USING THE PROPER GENERALIZED DECOMPOSITION

Using the Proper Generalized Decomposition, we will separate the domain into $3 \times 1D$ domains by assuming that E_x, E_y and E_z can be written as:

$$\begin{aligned} E_x &= \sum_{i=0}^{i=N} X_x^i(x) \cdot Y_x^i(y) \cdot Z_x^i(z) \\ E_y &= \sum_{i=0}^{i=N} X_y^i(x) \cdot Y_y^i(y) \cdot Z_y^i(z) \\ E_z &= \sum_{i=0}^{i=N} X_z^i(x) \cdot Y_z^i(y) \cdot Z_z^i(z) \end{aligned} \tag{4}$$

Such representation is appealing for solving 3D problems efficiently, especially for problems with rich physics in the in-plane and out-of-plane directions, while ensuring a computational complexity of few 1D resolutions [6]. Such representation allows the use of a fine mesh in the degenerated thickness direction if needed. However, classical techniques should employ a 3D extremely fine mesh in the entire domain to capture the discontinuity of the electrical field around the conductors, which involves a prohibitive number of degrees of freedom. Moreover, for the resolution of the separated 1D problems, one could use any suitable numerical technique (for example finite elements for one problem and finite differences for the second one).

The resolution procedure starts by replacing equations (4) into the weak form of the Maxwell equation (1). Afterwards, a fixed-point algorithm is used to compute the unknown functions, one at a time. All the details of the resolution algorithm can be found in [9].

4 RESULTS

4.1 Large domain

As a first step, we propose to simulate the problem detailed in section 2 on a large domain such that the wavelength is small compared to a, b or c values. Therefore we start by showing the results on a domain having the dimensions $a = 1m$; $b = 0.4m$ and $c = 1m$. We use 110 nodes along each of the x and z directions, while we use 500 nodes along the y direction. All the following figures are showing the amplitude of the electrical fields components. Figure 2 illustrated the E_y component in the domain, while figure 3 shows E_y inside the domain using 2 cuts normal to the x and z axis. Figure 4 illustrates E_y in the domain with a horizontal cut, normal to the y-axis, and passing by the conductor. On the other hand, figures 5, 7 and 6 show respectively E_x, E_y and E_z on a cut passing by $z = 0.5m$.

4.2 Small domain

In this section, we solve the same model on a domain that is relatively small compared to the wavelength, therefore we consider $a = 1mm$; $b = 0.4mm$ and $c = 1mm$. The result for E_y in a

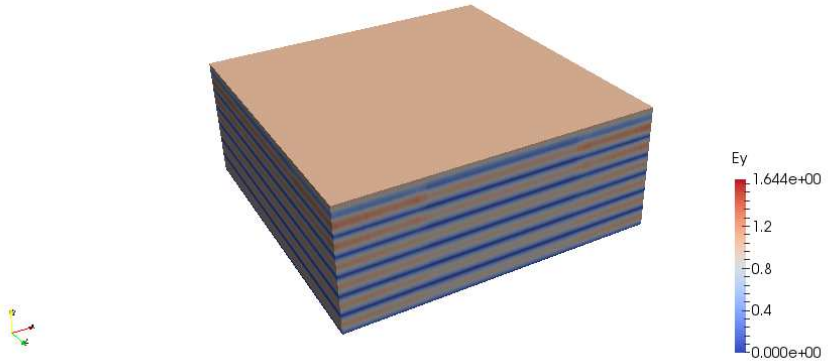


Figure 2: The electrical fields on the domain $a = 1m$; $b = 0.4m$ and $c = 1m$

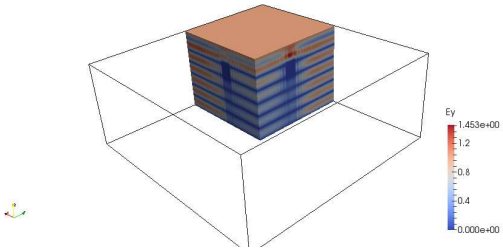


Figure 3: E_y shown inside the domain $a = 1m$; $b = 0.4m$ and $c = 1m$

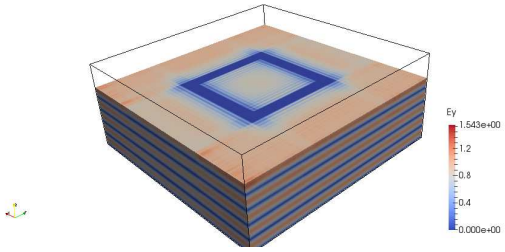


Figure 4: E_y on an horizontal cut passing by the conductor

vertical section passing by the fibers is shown in figure 8. Moreover, figures 10, 9 and 11 show respectively E_x, E_y and E_z in a section normal to the z-axis and passing by $z = 0.5 \cdot 10^{-3}m$.

5 CONCLUSION

In this work, we solved the Maxwell equations in a heterogeneous domain to simulate the electrical fields. The resolution is performed on two different domains, the first one large with respect to the wavelength, while the second one is small with respect to it.

The shown results are different in each case. In the large domain, the wavelength is small with respect to the domain dimensions therefore the electric field can oscillate inside the domain. However, in the second case, the electric field can't oscillate due to the large wavelength with respect to the dimensions of the domain.

For the first time to the knowledge of the authors we are able to simulate in full 3D the electric fields with a resolution up to more than 6 000 000 nodes (18 000 000 degrees of freedom) with about an hour on a standard laptop, core i7. Such simulation wasn't possible without the use of the PGD which allows separating the 3D domain into 3×1D domains. Therefore the computation time is reduced several orders of magnitude, from few days to few minutes. Moreover, the used resolution allows capturing the 3D effects generated by the presence of the

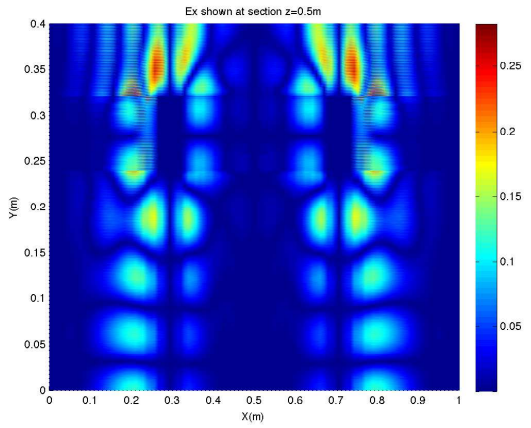


Figure 5: E_x on a section passing by $z = 0.5$ m

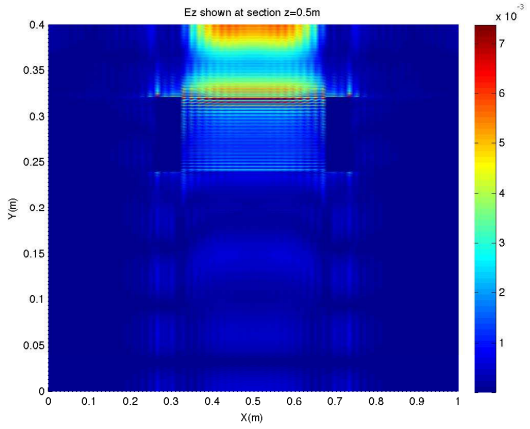


Figure 6: E_z on a section passing by $z = 0.5$ m

conductor in the domain, which changes the electrical fields around the conductor.

This present work will be generalized to a large number of fibers inside the domain, leading to effective simulation of composite materials and therefore a better understanding of the microwave heating process in both thermoplastics and thermosets.

REFERENCES

[1] M. Kwak, P. Robinson, A. Bismarck and R. Wise, Microwave curing of carbon-epoxy composites: Penetration depth and material characterisation. *Composites Part A*, **75**, 18-23, 2015.

[2] G. Hug., *Behavior analysis of carbon-epoxy laminates under high-speed loading: manufacture of the same materials by means of microwave curing for comparison, Ph.D. thesis*, Arts et Metiers ParisTech, 2005.

[3] R. Youssof, M.K. Aroua, A. Nesbitt and R.J. Day, Curing of polymeric composites using microwave resin transfer moulding (RTM). *Journal of Engineering Science and Technology*, **2**(2), 51-163, 2007.

[4] C. Ghnatios, F. Masson, A. Huerta, E. Cueto, F. Chinesta, Proper Generalized Decomposition Based Dynamic Data-Driven of Thermal Processes. *Computer Methods in Applied Mechanics and Engineering*, **213-216**, 29-41, 2012.

[5] C. Ghnatios, F. Chinesta and C. Binetruy, 3D Modeling of squeeze flows occurring in composite laminates. *International Journal of Material Forming*, **8**(1), 73-83, 2015.

[6] F. Chinesta, A. Ammar, E. Cueto, Recent Advances And New Challenges In the Use Of The Proper Generalized Decomposition For Solving Multidimensional Models. *Archives of Computational Methods in Engineering*, **17**, 327-350, 2009.

[7] B. Bognet, A. Leygue, F. Chinesta, A. Poitou and F. Bordeu, Advanced simulation of models defined in plate geometries: 3D solutions with 2D computational complexity. *Computer Methods in Applied Mechnanics and Engineering*, **201**, 1-12, 2012.

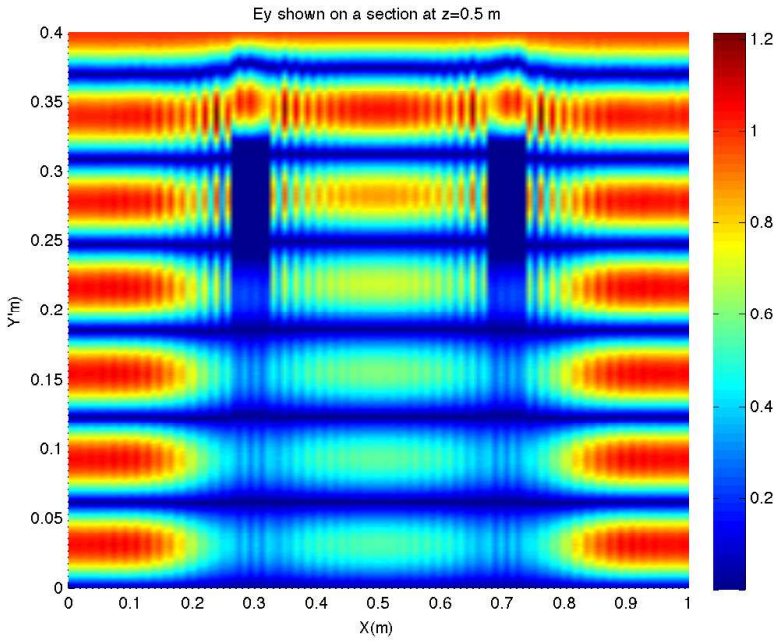


Figure 7: E_y on a section passing by $z = 0.5$ m

[8] C. Ghnatios, F. Chinesta, E. Cueto, A. Leygue, A. Poitou, P. Breitkopf, P. Villon, Methodological approach to efficient modeling and optimization of thermal processes taking place in a die: Application to pultrusion. *Composite Part A*, **42**, 1169-1178, 2011.

[9] C. Ghnatios., *Simulation anvancee des problemes thermiques rencontres lors de la mise en forme des composites, Ph.D. thesis*, Ecole Centrale Nantes, 2012.

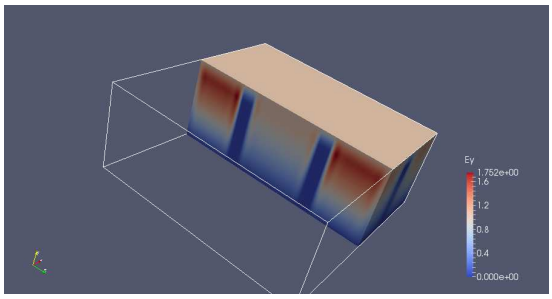


Figure 8: The electrical field E_y in the domain $a = 1 \times 10^{-3}m$; $b = 0.4 \times 10^{-3}m$ and $c = 1 \times 10^{-3}m$, shown on a section normal to z

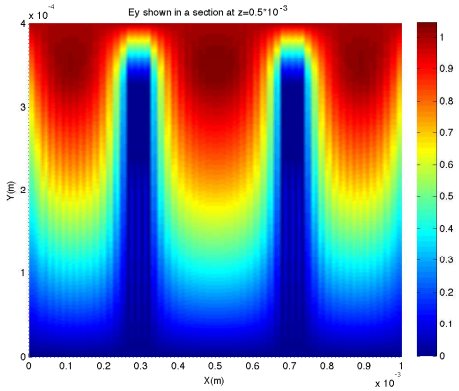


Figure 9: E_y on a section passing by $z = 0.5 \times 10^{-3}$ m

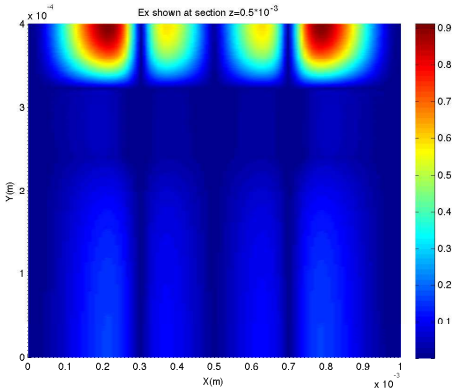


Figure 10: E_x on a section passing by $z = 0.5 \times 10^{-3}$ m

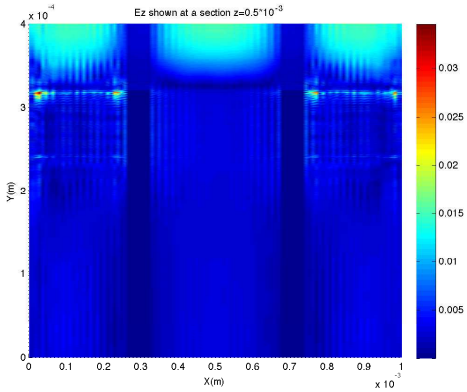


Figure 11: E_z on a section passing by $z = 0.5 \times 10^{-3}$ m

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

MODEL ORDER REDUCTION IN STRUCTURAL DYNAMICS

Raul Rodriguez Sanchez, Martin Buchschmid and Gerhard Müller

Chair of Structural Mechanics, Technische Universität München
Arcissstraße 21, 80333, Munich, Germany
e-mail: {raul.rodriguez, martin.buchschmid, gerhard.mueller}@tum.de

Keywords: Model Order Reduction, Modal Analysis, Guyan Reduction, Moment Matching, Krylov Subspace, Frequency Response.

Abstract. *Frequency response analysis in structural dynamics usually requires solving large dynamical systems of the form* $(-\omega^2\mathbf{M} + i\omega\mathbf{D} + \mathbf{K})\mathbf{u}(\omega) = \mathbf{f}(\omega)$*, which result from a FE discretization. A straightforward solution of big systems requires a high computational cost; therefore several Model Order Reduction (MOR) techniques have been developed in the last decades to obtain faster and efficient results. Between them interpolatory approaches have gained importance for solving second order dynamical systems. This work presents and compares ten MOR techniques which are suitable for structural dynamics problems. These are: Guyan-Irons Reduction, Improved Reduction System, Dynamic Reduction, Real Modal Analysis, Complex Modal Analysis, Craig-Bampton Method, and Interpolatory MOR methods like Multi-point Padé Approximation, the Krylov-based Galerkin Projection and the Derivative-based Galerkin Projection. A brief summary of the theoretical background is presented for each method. A first numerical example shows the applicability for damped systems and a second example shows suitability of the Interpolatory MOR methods for industrial applications, using data from a commercial FE software (ANSYS®).*

1 INTRODUCTION

The equation of motion resulting from a finite element discretization of a mechanical system is given by

$$\mathbf{M}\ddot{\boldsymbol{u}}(t) + \mathbf{D}\dot{\boldsymbol{u}}(t) + \mathbf{K}\boldsymbol{u}(t) = \boldsymbol{f}(t) \tag{1}$$

Applying a Fourier transformation to eq. (1), with $\mathbf{u}(\omega) = \mathcal{F}\big(\boldsymbol{u}(t)\big)$ and $\mathbf{f}(\omega) = \mathcal{F}\big(\boldsymbol{f}(t)\big)$ leads to

$$(-{\omega_j}^2\mathbf{M} + i\omega_j\mathbf{D} + \mathbf{K})\mathbf{u}(\omega_j) = \mathbf{f}(\omega_j) \quad j = 1, 2, 3...n_\omega \tag{2}$$

which is the equation of motion in the frequency domain, where $\mathbf{M}$, $\mathbf{D}$, $\mathbf{K} \in \mathbb{R}^{N\times N}$ are the mass, damping and stiffness matrix respectively, $\mathbf{f} \in \mathbb{R}^{N\times 1}$ is the load vector (force vector) and $\mathbf{u} \in \mathbb{R}^{N\times 1}$ is the displacement vector. n_ω is the number of times that $\mathbf{u}(\omega)$ is computed in eq. (2) in the scope of frequency response analysis. If n_ω and N are large numbers solving eq.(2) for $\mathbf{u}(\omega_j)$ becomes computationally expensive, in these cases one may search for a reduced-order model which would lead to a lower computational time. Such a reduced-order model is achieved using a suitable MOR technique. The main idea of MOR techniques is to find a vector space spanned by the columns of $\mathbf{V} \in \mathbb{C}^{N\times n_r}$, with $n_r \ll N$, which maps a reduced set of degrees of freedom (dofs) $\mathbf{u}_r \in \mathbb{C}^{n_r\times 1}$ into the original set of degrees of freedom $\mathbf{u}$, such that

$$\mathbf{u} \approx \tilde{\mathbf{u}} = \mathbf{V}\mathbf{u}_r \tag{3}$$

where $\tilde{\mathbf{u}}$ is the approximation of $\mathbf{u}$. In the following the approximation will not be labeled anymore. By substituting the approximation of $\mathbf{u}$ given by eq. (3) in eq. (2), there exists an error $\mathbf{e}_r$ defined by

$$\mathbf{e}_r = (-\omega^2\mathbf{M} + i\omega\mathbf{D} + \mathbf{K})\mathbf{V}\mathbf{u}_r - \mathbf{f} \tag{4}$$

This error is not contained in the Ran($\mathbf{V}$), therefore it is said to be in the null space of $\mathbf{V}$, i.e. the transpose of $\mathbf{V}$ is orthogonal to $\mathbf{e}_r$,

$$\mathbf{V}^H\big((-\omega^2\mathbf{M} + i\omega\mathbf{D} + \mathbf{K})\mathbf{V}\mathbf{u}_r - \mathbf{f}\big) = 0. \tag{5}$$

Where H is used for the conjugate transpose of a matrix. Eq. (5) is equivalent to performed Galerkin projection [14] or applied the least square method to the overestimated system of equations resulting from substituting eq. (3) in eq. (2). From eq. (5) the following reduced order model is achieved:

$$(-{\omega_j}^2\mathbf{M}_r + i\omega_j\mathbf{D}_r + \mathbf{K}_r)\mathbf{u}_r(\omega_j) = \mathbf{f}_r(\omega_j) \tag{6}$$

where $\mathbf{M}_r$, $\mathbf{D}_r$, $\mathbf{K}_r \in \mathbb{C}^{n_r\times n_r}$, $\mathbf{f}_r \in \mathbb{C}^{n_r\times 1}$ are defined by

$$\begin{aligned} &\mathbf{M}_r = \mathbf{V}^H\mathbf{M}\mathbf{V} \quad \mathbf{D}_r = \mathbf{V}^H\mathbf{D}\mathbf{V} \quad \mathbf{K}_r = \mathbf{V}^H\mathbf{K}\mathbf{V} \\ &\mathbf{f}_r = \mathbf{V}^H\mathbf{f} \end{aligned} \tag{7}$$

For the construction of that matrix $\mathbf{V}$ different approaches have been proposed in the field of structural mechanics. A classical approach to construct $\mathbf{V}$ is using a modal decomposition; this idea was in the mathematics community since the 18th century and it uses the superposition principle, which was stated by Bernoulli in 1753 [1]. With the development of the finite element method (FEM) the modal decomposition has been extensively used to reduce large numerical models leading to a set of decoupled equations, whose solutions are straightforward. In 1965 Guyan and Irons presented the first condensation method, Guyan-Irons Reduction [2, 3]. A

new paradigm was introduced in 1968 by Craig and Bampton [6], the Craig-Bampton method which introduced the concept of sub structuring, has allowed the achievement of computing large complex systems and is suitable for parallelization. A similar procedure to that of the Guyan-Irons reduction but for dynamical analysis was introduced in 1978 by Leung [4], the dynamic condensation (Dynamic Reduction) which allows to reduce a system about a frequency of interest. In 1991 Blair et. al. [19] presented an iterative process to the already established Improved Reduction System method presented by O'Callahan [17] in 1989. From system and control field the moment matching method, i.e. the Krylov Subspace method was reformulated in 1989 by Craig and Su [7] to solve second order dynamical systems, thus the reduced-order model preserves the second order form of the original model, avoiding the structural dynamics model be transformed into the first order formulation (state space form), which destroys the meaning of the original model. It turns out that the Krylov subspace method results in a very efficient method in computational cost and its parallelization could be easily achieved. In this contribution ten MOR techniques are reviewed and classified as follows

1. Modal Decomposition Methods
    - Real Modal Analysis (RMA)
    - Complex Modal Analysis (CMA)
2. Condensation Methods
    - Guyan-Irons Reduction (GR)
    - Improved Reduction System (IRS)
    - Dynamic Reduction (DR)
3. Component Mode Synthesis Methods
    - Craig-Bampton Method (CB)
4. Interpolatory Methods
    - Multi-point Padé Approximation (mP)
    - Krylov-based Galerkin Projection (KGP)
    - Derivative-based Galerkin Projection (DGP)

The general procedure of a frequency response analysis using MOR techniques is summarized as follows:

1. Define a frequency interval $\Delta_\omega = [\omega_L\ \omega_R]$
2. Define a set of reduced dofs or generalized coordinates $\mathbf{u}_r$ and construct the basis for the matrix $\mathbf{V}$, see secs. 2 - 5, such that $\mathbf{V}\mathbf{u}_r$ approximates the physical coordinates $\mathbf{u}$.
3. Compute the reduced force vector and the reduced matrices of eq. (7)
4. Compute $\mathbf{u}_r(\omega)$ in eq. (6) and use eq.(3) to get $\mathbf{u}$ in the interval Δ_ω.
    (a) Define n_ω frequency points in Δ_ω, those frequency points are denominated ω_j.

(b) Compute $\mathbf{u}_r(\omega_j)$ n_ω times, i.e. for each ω_j

$$\mathbf{u}_r(\omega_j) = \big(-\omega_j^2 \mathbf{M}_r + i\omega_j \mathbf{D}_r + \mathbf{K}_r \big)^{-1} \mathbf{f}_r \quad j = 1, 2, 3, ..., n_\omega \tag{8}$$

(c) Project the vector $\mathbf{u}_r$ into the Ran($\mathbf{V}$) to get the solution $\mathbf{u}(\omega_j)$ of the original model using (3):

$$\mathbf{u}(\omega_j) = \mathbf{V}\mathbf{u}_r(\omega_j) \tag{9}$$

2 MOR: Modal Decomposition Methods

This section presents an overview of the classical modal decomposition methods

- Real Modal Analysis
- Complex Modal Analysis

2.1 Real modal analysis

Real Modal Analysis uses a set of natural vibration forms (mode shapes) of the full-model as the vector space which contains $\mathbf{u}$, i.e. the displacement vector $\mathbf{u}(\omega)$ is approximated by the set of n_r natural modes using the expansion theorem [15]

$$\begin{aligned} \mathbf{u}(\omega) &= \sum_{j=1}^{n_r} \boldsymbol{\phi}_j \eta_j(\omega) \\ &= [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \boldsymbol{\phi}_3 \cdots \boldsymbol{\phi}_{n_r}]\boldsymbol{\eta} \\ &= \boldsymbol{\Phi}\boldsymbol{\eta} \end{aligned} \tag{10}$$

where the modal matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times n_r}$ and the vector of modal coordinates $\boldsymbol{\eta} \in \mathbb{R}^{n_r \times 1}$ are $\mathbf{V}$ and $\mathbf{u_r}$ in eq. (3) respectively.
The modal matrix $\boldsymbol{\Phi}$ is obtained from the free vibration problem of eq. (2), which appears when $\mathbf{f}$ and $\mathbf{D}$ are equal to zero. Then a generalized eigenvalue problem is formulated:

$$(\mathbf{K} - \omega_j^2 \mathbf{M})\boldsymbol{\phi}_j = 0 \qquad j = 1, 2, 3, ..., n_r \tag{11}$$

where the eigenvalue $\lambda_j = \omega^2$ is the jth natural frequency of the system and the eigenvector $\boldsymbol{\phi}_j \in \mathbb{R}^{N \times 1}$ is the jth mode shape, which is normalized for convenience. The mode shapes are mutually orthogonal to the mass and stiffness matrices, i.e.

$$\boldsymbol{\phi}_j^T \mathbf{M} \boldsymbol{\phi}_i = \quad \begin{Bmatrix} = 0 \quad \text{for} \quad j \neq i \quad \text{orthogonality} \\ = m_j \quad \text{for} \quad j = i \end{Bmatrix} \tag{12}$$

$$\boldsymbol{\phi}_j^T \mathbf{K} \boldsymbol{\phi}_i = \quad \begin{Bmatrix} = 0 \quad \text{for} \quad j \neq i \quad \text{orthogonality} \\ = k_j \quad \text{for} \quad j = i \end{Bmatrix} \tag{13}$$

If the damping matrix $\mathbf{D}$ is defined to be proportional to the mass and stiffness matrices, the mode shapes are also mutually orthogonal to the damping matrix $\mathbf{D}$. Thus, due to the orthogonality property of the mode shapes of the modal matrix $\boldsymbol{\Phi}$, the reduced-order model given by eq. (6) results in a set of n_r decoupled equations of the form

$$(-\omega^2 m_j + i\omega c_j + k_j)\eta_j(\omega) = \boldsymbol{\phi}_j^T \mathbf{f}(\omega) \quad j = 1, 2, 3, \ldots, n_r \tag{14}$$

2.2 Complex Modal Analysis

In case of non-proportional damping a reduced decoupled system is obtained using a complex modal analysis, i.e. using complex eigenvectors and eigenvalues. Therefore the second order ordinary differential equations given by eq. (1) is expressed as a 2N-dofs first order differential equations, i. e. the state-space formulation

$$\dot{\boldsymbol{z}}(t) = \mathbf{A}\boldsymbol{z}(t) + \mathbf{B}\boldsymbol{f}(t) \tag{15}$$

where the state vector $\boldsymbol{z} \in \mathbb{R}^{2N\times 1}$, the state matrix $\mathbf{A} \in \mathbb{R}^{2N\times 2N}$, the input matrix $\mathbf{B} \in \mathbb{R}^{2N\times N}$, and the input vector $\boldsymbol{f} \in \mathbb{R}^{2N\times 1}$ are defined by

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{u} \\ \dot{\boldsymbol{u}} \end{bmatrix}; \quad \mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{D} \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{bmatrix} \tag{16}$$

Considering the free vibration problem, i.e. the external force vector $\boldsymbol{f}$ equal to zero and using a solution of the form $\boldsymbol{z} = \sum_{k=1}^{n_r} e^{\lambda_k}\hat{\boldsymbol{\phi}}_{\boldsymbol{k}}$ leads to

$$\lambda_k\hat{\boldsymbol{\phi}}_k = \mathbf{A}\hat{\boldsymbol{\phi}}_k \qquad k = 1, 2, ..., n_r \tag{17}$$

which is the standard eigenvalue problem. $\mathbf{A}$ is a real and nonsymmetric matrix, therefore the $2N$ eigenvalues must either be real or they must occur in complex conjugate pairs and the modal matrix is composed by N eigenvectors and their N complex conjugates. The right eigenvectors (eigenvectors of $\mathbf{A}$) $\hat{\boldsymbol{\phi}}_j$ and $\hat{\boldsymbol{\phi}}_k$ are not orthogonal with respect to the matrix $\mathbf{A}$. Nevertheless the eigenvalues of $\mathbf{A}$ and the eigenvalues of $\mathbf{A}^T$ are the same. Furthermore the right eigenvectors $\hat{\boldsymbol{\phi}}_k$, with $k = 1, 2, ..., n_r$, are biorthogonal with respect to the matrix $\mathbf{A}$ to the left eigenvectors $\hat{\boldsymbol{\psi}}_j$ (eigenvectors of $\mathbf{A}^T$), with $j = 1, 2, ..., n_r$, [16]. For convenience the normalization of the eigenvectors is done such that

$$\hat{\boldsymbol{\psi}}_j^T\hat{\boldsymbol{\phi}}_k = \delta_{kj}, \quad j, k = 1, 2, ..., n_r \tag{18}$$

holds true. Then n_r right eigenvectors of $\mathbf{A}$ are used as a basis for a vector space which contains $\boldsymbol{z}$ approximately, i.e. the expansion theorem permits to represent the state vector $\boldsymbol{z}$ as a combination of the right eigenvectors of $\mathbf{A}$

$$\boldsymbol{z} = \sum_{k=1}^{n_r} \hat{\boldsymbol{\phi}}_k\xi_k(t) \tag{19}$$

or in the frequency domain

$$\mathbf{z} = \sum_{k=1}^{n_r} \hat{\boldsymbol{\phi}}_k\hat{\xi}_k(i\omega) \tag{20}$$

where $\hat{\xi}_k(i\omega) = \mathcal{F}\big(\xi_k(t)\big)$. Taking Fourier transform of eq. (15) and using eq.(20) leads to

$$i\omega\hat{\xi}_k(i\omega) = \lambda_k\hat{\xi}_k(\omega) + \hat{\boldsymbol{\psi}}^T\mathbf{B}\mathbf{f}(i\omega) \tag{21}$$

or

$$\hat{\xi}_k(i\omega) = \frac{\hat{\boldsymbol{\psi}}^T\mathbf{B}\mathbf{f}(i\omega)}{i\omega - \lambda_k} \tag{22}$$

where the set of modal coordinates $\hat{\boldsymbol{\xi}} = [\hat{\xi}_1, \hat{\xi}_2, .., \hat{\xi}_{n_r}]^T \in \mathbb{C}^{n_r\times 1}$, and the set of right eigenvectors $\hat{\boldsymbol{\Phi}} = [\hat{\boldsymbol{\phi}}_1, \hat{\boldsymbol{\phi}}_2, ..., \hat{\boldsymbol{\phi}}_{n_r}] \in \mathbb{C}^{2N\times n_r}$ are the reduced set of dofs $\mathbf{u}_r$ and the matrix $\mathbf{V}$ in eq. (3) respectively and the set of left eigenvectors $\hat{\boldsymbol{\Psi}} = [\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2, ..., \hat{\boldsymbol{\psi}}_{n_r}] \in \mathbb{C}^{2N\times n_r}$ is $\mathbf{V}^H$ in eq. (5) and (7).

3 MOR: Condensation Methods

This section addresses the formulation of the following condensation methods:

- Guyan-Irons Reduction
- Improved Reduction System
- Dynamic Reduction

The main approach of the condensation methods is to have a linear dependence between dofs, i.e. if a set of *dependent* dofs, $\mathbf{u}_d \in \mathbb{R}^{(N-n_r)\times 1}$, can be expressed as linear dependent of a set of *active* dofs, $\mathbf{u}_a \in \mathbb{R}^{n_r \times 1}$, then a condensation method can be applied to the system, which leads to a reduced-order model. Assuming the above statement is valid for a structural system, the total displacement vector is divided in two groups:

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_a \end{bmatrix} \tag{23}$$

then the linear relation between $\mathbf{u}_a$ and $\mathbf{u}_d$ is created through the linear transformation matrix $\mathbf{T}$ as follows:

$$\mathbf{u}_d = \mathbf{T}\mathbf{u}_a \tag{24}$$

Thus, the total displacement vector is defined by

$$\mathbf{u} = \begin{bmatrix} \mathbf{T} \\ \mathbf{I} \end{bmatrix} \mathbf{u}_a \tag{25}$$

where $[\mathbf{T} \;\; \mathbf{I}]^T$ and the set of active dofs $\mathbf{u}_a$ are $\mathbf{V}$ and the reduced vector of dofs $\mathbf{u}_r$ in eq. (3) respectively.

3.1 Guyan-Irons Reduction

The Guyan-Iron Reduction [2, 3] neglects the inertial and the damping contributions in eq. (2), i.e. it considers just the static case, for this reason it is called static condensation. Then the static equation system is divided as follows

$$\begin{bmatrix} \mathbf{K}_{dd} & \mathbf{K}_{da} \\ \mathbf{K}_{ad} & \mathbf{K}_{aa} \end{bmatrix} \begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_a \end{bmatrix} = \begin{bmatrix} \mathbf{f}_d \\ \mathbf{f}_a \end{bmatrix} \tag{26}$$

The linear relation between dofs given by eq. (24) is obtained by solving for $\mathbf{u}_d$ in the first row of eq. (26) and is given by

$$\mathbf{u}_d = \mathbf{T}_g \mathbf{u}_a + \mathbf{u}_e \tag{27}$$

where $\mathbf{T}_g$ is the Guyan linear transformation matrix defined by

$$\mathbf{T}_g = -\mathbf{K}_{dd}^{-1}\mathbf{K}_{da} \tag{28}$$

and $\mathbf{u}_e = \mathbf{K}_{dd}^{-1}\mathbf{f}_d$ is the Guyan Reduction-error which would emerge if $\mathbf{f}_d$ is not the zero vector. It is also interpreted as a correction factor which guaranties that the exact solution of the full-order model is recovered when static systems are considered. Thus the displacement vector $\mathbf{u}$ can be written as a function of the active dofs $\mathbf{u}_a$:

$$\begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_a \end{bmatrix} = \begin{bmatrix} -\mathbf{K}_{dd}^{-1}\mathbf{K}_{da} \\ \mathbf{I} \end{bmatrix} \mathbf{u}_a + \begin{bmatrix} \mathbf{K}_{dd}^{-1}\mathbf{f}_d \\ \mathbf{0} \end{bmatrix} \tag{29}$$

Usually, the set of depended dofs are the force-free ones, i.e. $\mathbf{f}_d = \mathbf{0}$. From (29) the Guyan projection matrix is given by

$$\mathbf{V}_g = \begin{bmatrix} \mathbf{T}_g \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} -\mathbf{K}_{dd}^{-1}\mathbf{K}_{da} \\ \mathbf{I} \end{bmatrix} \tag{30}$$

$\mathbf{V}_g$ and $\mathbf{u}_a$ are the matrix $\mathbf{V}$ and the reduced set $\mathbf{u}_r$ in eq. (3) respectively. Therefore the reduced-order model, eq. (6), is given by:

$$(-\omega^2\mathbf{M}_g + i\omega\mathbf{D}_g + \mathbf{K}_g)\mathbf{u}_a = \mathbf{f}_g \tag{31}$$

Applying equations given by (7) the following Guyan reduced matrices are obtained:

$$\begin{aligned}
\mathbf{M}_g = \mathbf{V}_g^T\mathbf{M}\mathbf{V}_g &= \begin{bmatrix} (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{M}_{dd} & \mathbf{M}_{da} \\ \mathbf{M}_{ad} & \mathbf{M}_{aa} \end{bmatrix} \begin{bmatrix} (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) \\ \mathbf{I} \end{bmatrix} \\
&= \mathbf{M}_{aa} + (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T\mathbf{M}_{dd}(-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) + \mathbf{M}_{ad}(-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) + (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T\mathbf{M}_{da} \\
&= \mathbf{M}_{aa} + \mathbf{T}_g^T\mathbf{M}_{dd}\mathbf{T}_g + \mathbf{M}_{ad}\mathbf{T}_g + \mathbf{T}_g^T\mathbf{M}_{da}
\end{aligned} \tag{32}$$

$$\begin{aligned}
\mathbf{D}_g = \mathbf{V}_g^T\mathbf{D}\mathbf{V}_g &= \begin{bmatrix} (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{D}_{dd} & \mathbf{D}_{da} \\ \mathbf{D}_{ad} & \mathbf{D}_{aa} \end{bmatrix} \begin{bmatrix} (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) \\ \mathbf{I} \end{bmatrix} \\
&= \mathbf{D}_{aa} + (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T\mathbf{D}_{dd}(-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) + \mathbf{D}_{ad}(-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) + (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T\mathbf{D}_{da} \\
&= \mathbf{D}_{aa} + \mathbf{T}^T\mathbf{D}_{dd}\mathbf{T}_g + \mathbf{D}_{ad}\mathbf{T}_g + \mathbf{T}_g^T\mathbf{D}_{da}
\end{aligned} \tag{33}$$

$$\begin{aligned}
\mathbf{K}_g = \mathbf{V}_g^T\mathbf{K}\mathbf{V}_g &= \begin{bmatrix} (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{K}_{dd} & \mathbf{K}_{da} \\ \mathbf{K}_{ad} & \mathbf{K}_{aa} \end{bmatrix} \begin{bmatrix} (-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) \\ \mathbf{I} \end{bmatrix} \\
&= \mathbf{K}_{aa} + (\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T\mathbf{K}_{dd}\mathbf{K}_{dd}^{-1}\mathbf{K}_{da} + \mathbf{K}_{ad}(-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da}) - (\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})^T\mathbf{K}_{da} \\
&= \mathbf{K}_{aa} + \mathbf{T}_g^T\mathbf{K}_{da}
\end{aligned} \tag{34}$$

$$\mathbf{f}_g = \mathbf{V}_g^T\mathbf{f} = \begin{bmatrix} \mathbf{T}_g^T & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f}_d \\ \mathbf{f}_a \end{bmatrix} = \mathbf{T}_g^T\mathbf{f}_d + \mathbf{f}_a = \mathbf{f}_a \tag{35}$$

Usually, to save computational time only the first term of the right hand side of eqs. (32) and (33) are considered to define $\mathbf{M}_{aa}$ and $\mathbf{D}_{aa}$ respectively.

3.2 Improved Reduction System (IRS)

In the derivation of Guyan-Irons transformation matrix the inertial terms of eq. (2) were excluded, in the IRS [17] procedure they are considered to have an *improved* solution. Then (2) is subdivided in blocks having the dependent set of dofs force-free, i.e. $\mathbf{f}_d = \mathbf{0}$:

$$\left(-\omega^2 \begin{bmatrix} \mathbf{M}_{dd} & \mathbf{M}_{da} \\ \mathbf{M}_{ad} & \mathbf{M}_{aa} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{dd} & \mathbf{K}_{da} \\ \mathbf{K}_{ad} & \mathbf{K}_{aa} \end{bmatrix} \right) \begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_a \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_a \end{bmatrix} \tag{36}$$

The first row of (36) is given by

$$(\mathbf{K}_{dd} - \omega^2\mathbf{M}_{dd}+)\mathbf{u}_d + (\mathbf{K}_{da} - \omega^2\mathbf{M}_{da})\mathbf{u}_a = \mathbf{0} \tag{37}$$

Therefore the linear transformation defined by eq. (24) is given by

$$\mathbf{u}_d = -(\mathbf{K}_{dd} - \omega^2 \mathbf{M}_{dd})^{-1}(\mathbf{K}_{da} - \omega^2 \mathbf{M}_{da})\mathbf{u}_a \tag{38}$$

$(\mathbf{K}_{dd} - \omega^2 \mathbf{M}_{dd})^{-1}$ can be approximated by the first three terms of the Taylor series about $\omega_0 = 0$, i.e. using the binomial series [18] until the third term

$$\begin{aligned}
\mathbf{u}_d &= -(\mathbf{K}_{dd}^{-1} + \omega^2 \mathbf{K}_{dd}^{-1}\mathbf{M}_{dd}\mathbf{K}_{dd}^{-1})(\mathbf{K}_{da} - \omega^2 \mathbf{M}_{da})\mathbf{u}_a \\
&= -(\mathbf{K}_{dd}^{-1} + \omega^2 \mathbf{K}_{dd}^{-1}\mathbf{M}_{dd}\mathbf{K}_{dd}^{-1})(\mathbf{K}_{da} - \omega^2 \mathbf{M}_{da})\mathbf{u}_a \\
&= -\big(\mathbf{K}_{dd}^{-1}\mathbf{K}_{da} - \omega^2 \mathbf{K}_{dd}^{-1}\mathbf{M}_{da} + \omega^2 \mathbf{K}_{dd}^{-1}\mathbf{M}_{dd}\mathbf{K}_{dd}^{-1}\mathbf{K}_{da} - \omega^4 \mathbf{K}_{dd}^{-1}\mathbf{M}_{dd}\mathbf{K}_{dd}^{-1}\mathbf{M}_{da}\big)\mathbf{u}_a \\
&= \big(- \mathbf{K}_{dd}^{-1}\mathbf{K}_{da} + \mathbf{K}_{dd}^{-1}\big(\mathbf{M}_{da} + \mathbf{M}_{dd}(-\mathbf{K}_{dd}^{-1}\mathbf{K}_{da})\big)\omega^2 + \mathcal{O}(\omega^4)\big)\mathbf{u}_a
\end{aligned} \tag{39}$$

The frequency dependency in eq. (39) is eliminated using the following expression which comes from eq. (31) in case of free vibration of the reduced-Guyan model:

$$\omega^2 \mathbf{u}_a = \mathbf{M}_g^{-1}\mathbf{K}_g \mathbf{u}_a \tag{40}$$

Substituting eq. (40) in eq. (39), the set $\mathbf{u}_d$ is given by

$$\mathbf{u}_d = \mathbf{T}_{irs}\mathbf{u}_a \tag{41}$$

where $\mathbf{T}_{irs}$ is given by

$$\mathbf{T}_{irs} = \mathbf{T}_g + \boldsymbol{\tau}_{irs} \tag{42}$$

where $\mathbf{T}_g$ is given by eq. (28) and the inertial-corrective factor $\boldsymbol{\tau}_{irs}$ is given by

$$\boldsymbol{\tau}_{irs} = \mathbf{K}_{dd}^{-1}(\mathbf{M}_{da} + \mathbf{M}_{dd}\mathbf{T}_g)\mathbf{M}_g^{-1}\mathbf{K}_g \tag{43}$$

Thus the IRS projection matrix is given by:

$$\mathbf{V}_{irs} = \begin{bmatrix} \mathbf{T}_{irs} \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\tau}_{irs} + \mathbf{T}_g \\ \mathbf{I} \end{bmatrix} \tag{44}$$

A better IRS projection matrix can be achieved by an iterative procedure which was introduced by Blair et al. [19]. The reduced system matrices that Guyan reduction provides are used in (43) as updating matrices to produce the IRS transformation matrix. Thus, it is possible to update again the matrices to be used in an iterative scheme. The substitution of the reduced Guyan matrices $\mathbf{M}_g^{-1}$ and $\mathbf{K}_g$ in (43) by the new updated IRS matrices results in

$$\begin{aligned}
\mathbf{T}_{irs,i} &= \mathbf{K}_{dd}^{-1}(\mathbf{M}_{da} + \mathbf{M}_{dd}\mathbf{T}_g)\mathbf{M}_{irs,i-1}^{-1}\mathbf{K}_{irs,i-1} + \mathbf{T}_g \\
&= \boldsymbol{\tau}_{irs,i} + \mathbf{T}_g
\end{aligned} \tag{45}$$

The ith IRS projection matrix is now given by

$$\mathbf{V}_{irs,i} = \begin{bmatrix} \mathbf{T}_{irs,i} \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\tau}_{irs,i} + \mathbf{T}_g \\ \mathbf{I} \end{bmatrix} \tag{46}$$

with:

$$\mathbf{M}_{irs,0}^{-1} = \mathbf{M}_g^{-1} \tag{47}$$

$$\mathbf{K}_{irs,0} = \mathbf{K}_g \tag{48}$$

The reduced-order model, eq. (6), using the IRS projection matrix defined in eq. (46) is given by

$$(-\omega^2 \mathbf{M}_{irs,i} + i\omega \mathbf{D}_{irs,i} + \mathbf{K}_{irs,i})\mathbf{u}_a = \mathbf{f}_{irs,i} \tag{49}$$

These reduced matrices and vector are computed analogously to eqs. (32) - (35).

3.3 Dynamic Reduction

The dynamic reduction [4] is obtained considering the undamped version of eq. (2) which is given by

$$(-\omega^2\mathbf{M} + \mathbf{K})\mathbf{u}(\omega) = \mathbf{f}(\omega) \tag{50}$$

Subdividing eq. (50) in blocks yields

$$\begin{bmatrix} \mathbf{Z}_{dd} & \mathbf{Z}_{da} \\ \mathbf{Z}_{ad} & \mathbf{Z}_{aa} \end{bmatrix} \begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_a \end{bmatrix} = \begin{bmatrix} \mathbf{0}_d \\ \mathbf{f}_a \end{bmatrix} \tag{51}$$

where

$$\mathbf{Z}(\omega) = -\omega^2\mathbf{M} + \mathbf{K} \tag{52}$$

Solving the first row of (51) for $\mathbf{u}_d$ yields

$$\mathbf{u}_d = -\mathbf{Z}_{dd}(\omega)^{-1}\mathbf{Z}_{da}(\omega)\mathbf{u}_a \tag{53}$$

Thus the displacement vector $\mathbf{u}$ can be written as a function of the active dofs $\mathbf{u}_a$ and a chosen frequency ω as follows:

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_a \end{bmatrix} \boldsymbol{u}_a = \mathbf{V}_{dy}\mathbf{u}_a \tag{54}$$

where

$$\mathbf{V}_{dy} = \begin{bmatrix} \mathbf{T}_{dy} \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} -\mathbf{Z}_{dd}^{-1}(\omega)\mathbf{Z}_{da}(\omega) \\ \mathbf{I} \end{bmatrix} \tag{55}$$

$\mathbf{u}_a$ and $\mathbf{V}_{dy}$ are the reduced vector $\mathbf{u}_r$ and the matrix $\mathbf{V}$ in eq. (3) respectively. Therefore the reduced-order model , eq. (6), is given by

$$(-\omega^2\mathbf{M}_{dy} + i\omega\mathbf{D}_{dy} + \mathbf{K}_{dy})\mathbf{u}_a = \mathbf{f}_{dy} \tag{56}$$

The reduced matrices and reduced force vector are obtained using eq. (7) analogously to eqs. (32)-(35).

4 MOR: Component-Mode Synthesis Methods

The component-mode synthesis approach consists in the subdivision of a mechanical model into substructures using the following classification for the dofs [15], see fig. 1:
$\mathcal{I}$: interior coordinates (i as subscript in matrix notation).
$\mathcal{R}$: rigid-body coordinates (r as subscript in matrix notation).
$\mathcal{E}$: excess boundary coordinates (e as subscript in matrix notation).
$\mathcal{B} = \mathcal{R} + \mathcal{E}$: boundary coordinates ($b$ as subscript in matrix notation) with the assumption that dofs with external loads are assumed to be boundary dofs.

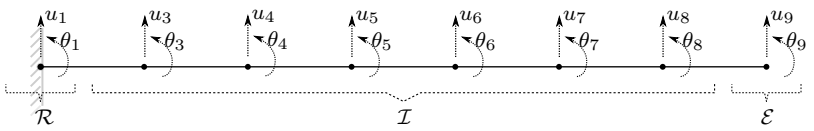


Figure 1: FE discretization of a catiliver beam.

The equation of motion of an undamped sub-structure, identified by the superscript α, is given by

$$\mathbf{M}^{(\alpha)}\ddot{\boldsymbol{u}}^{(\alpha)} + \mathbf{K}^{(\alpha)}\boldsymbol{u}^{(\alpha)} = \boldsymbol{f}^{(\alpha)} + \boldsymbol{r}^{(\alpha)} \tag{57}$$

where $\mathbf{M}^{\alpha}$, $\mathbf{K}^{\alpha}$, $\boldsymbol{u}^{\alpha}$ and $\boldsymbol{f}^{\alpha}$ are the mass matrix, stiffness matrix, displacement vector and force vector at the component label α respectively. $\boldsymbol{r}^{\alpha}$ is a force vector which contains the reaction forces on the component due to its connection to adjacent components at the boundary dofs.

Component Modes

Eq. (57) may be written in the following partitioned form

$$\begin{bmatrix} \mathbf{M}_{ii} & \mathbf{M}_{ib} \\ \mathbf{M}_{bi} & \mathbf{M}_{bb} \end{bmatrix} \begin{bmatrix} \ddot{\boldsymbol{u}}_i \\ \ddot{\boldsymbol{u}}_b \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{ii} & \mathbf{K}_{ib} \\ \mathbf{K}_{bi} & \mathbf{K}_{bb} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}_b \end{bmatrix} = \begin{bmatrix} \mathbf{0}_i \\ \boldsymbol{f}_b + \boldsymbol{r}_b \end{bmatrix} \tag{58}$$

Fixed-Interface Normal Modes

The fixed interfaced normal modes are a set of n_i modes shapes resulting by solving the following eigenvalue problem of the internal dofs of eq. (58), while the boundary dofs are fixed, see fig. 2,

$$\left[\mathbf{K}_{ii} - \omega_k^2 \mathbf{M}_{ii}\right] \boldsymbol{\phi}_k = 0 \qquad k = 1, 2, \ldots, n_i \tag{59}$$

The eigenvectors $\boldsymbol{\phi}_k$ are gathered in the internal modal matrix $\boldsymbol{\Phi}_{ii} \in \mathbb{R}^{n_i \times 1}$. Therefore the fixed interfaced normal modes for a substructure α is given by

$$\boldsymbol{\Phi}_i = \begin{bmatrix} \boldsymbol{\Phi}_{ii} \\ \mathbf{0}_{bi} \end{bmatrix} \tag{60}$$

where $\boldsymbol{\Phi}_i \in \mathbb{R}^{N_\alpha \times 1}$. N_α is the total number of dofs of the structure or substructure.

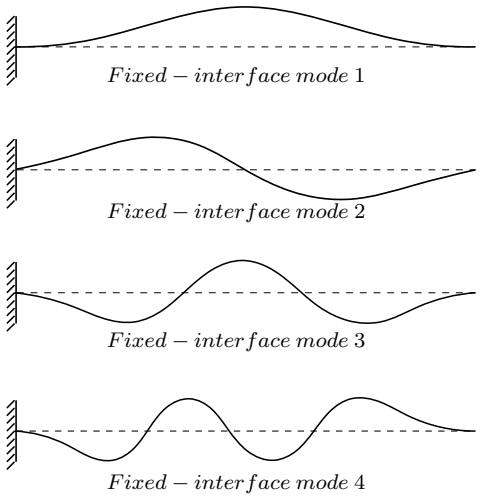


Figure 2: Four component fixed-interface normal modes of a cantiliver beam.

Interface Constraint Modes

Interface constraint modes result when a unit displacement is applied in one dof in the set of boundary dofs, while the remaining dofs of that set are restrained, and conserving the remaining dofs force-free, see fig. 3. The constraint modes are defined by solving the following static equation for $[\boldsymbol{\Psi}_{ib} \;\; \mathbf{I}_{bb}]^T$:

$$\begin{bmatrix} \mathbf{K}_{ii} & \mathbf{K}_{ib} \\ \mathbf{K}_{bi} & \mathbf{K}_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{\Psi}_{ib} \\ \mathbf{I}_{bb} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{ib} \\ \mathbf{R}_{bb} \end{bmatrix} \tag{61}$$

Thus the interface constraint mode matrix $\mathbf{\Psi}_b$ is given by

$$\mathbf{\Psi}_b = \begin{bmatrix} \mathbf{\Psi}_{ib} \\ \mathbf{I}_{bb} \end{bmatrix} = \begin{bmatrix} -\mathbf{K}_{ii}^{-1}\mathbf{K}_{ib} \\ \mathbf{I}_{bb} \end{bmatrix} \tag{62}$$

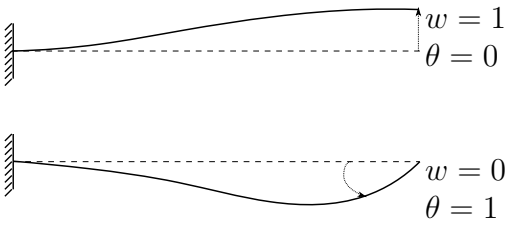


Figure 3: Interface-constraint modes in the boundary dofs of a cantilever beam.

Those constraint modes are stiffness-orthogonal to all of the fixed-interface normal modes [15].

4.1 The Craig-Bampton Method

The Craig-Bampton transformation [6, 8] consists of transforming a set of physical coordinates made up of internal dofs $\boldsymbol{u}_i$ and dofs at the boundary $\mathbf{u}_b$ in terms of a hybrid set which is compound of a set of modal coordinates at the interior $\boldsymbol{\eta}_i$ and physical coordinates $\boldsymbol{u}_b$, at the boundary. The Craig-Bampton hybrid coordinates $\boldsymbol{u}_{cb} = [\boldsymbol{\eta}_i \;\; \boldsymbol{u}_b]^T$ are then related to the physical coordinates $\boldsymbol{u} = [\boldsymbol{u}_i \;\; \boldsymbol{u}_b]^T$ as follows

$$\begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}_b \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_i & \mathbf{\Psi}_b \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_i \\ \boldsymbol{u}_b \end{bmatrix} \tag{63}$$

where $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_b$ are given by eqs. (60) and (62) respectively. Substituting those equations in (63)

$$\begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}_b \end{bmatrix} = \mathbf{V}_{cb}\mathbf{u}_{cb} \tag{64}$$

where

$$\mathbf{V}_{cb} = \begin{bmatrix} \mathbf{\Phi}_{ii} & -\mathbf{K}_{ii}^{-1}\mathbf{K}_{ib}] \\ \mathbf{0}_{bi} & \mathbf{I}_{bb} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_{ii} & \mathbf{\Psi}_{ib} \\ \mathbf{0}_{bi} & \boldsymbol{I}_{bb} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_i & \mathbf{\Psi}_c \end{bmatrix} \tag{65}$$

is the Craig-Bampton projection matrix. $\mathbf{V}_{cb}$ and $\mathbf{u}_{cb}$ are the $\mathbf{V}$ matrix and the $\mathbf{u}_r$ vector in eq. (3) respectively. After defining the Craig-Bampton Transformation Matrix above, it follows the transformation of the equation of motion from the set of physical coordinates to a set of coordinates consisting of physical coordinates at the subset of the boundary dofs and modal coordinates at interior dofs.
Using the the Craig-Bampton projection matrix eq. (65) in eq. (7) the reduced order model in eq. (6) is obtained [8] and given by

$$\left(-\omega^2 \begin{bmatrix} \mathbf{m} & \mathbf{M}_{kb} \\ \mathbf{M}_{bk} & \mathbf{M}_{bb} \end{bmatrix} + i\omega \begin{bmatrix} \mathbf{c} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{k} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{bb} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{\eta} \\ \mathbf{u}_b \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_b \end{bmatrix} \tag{66}$$

where

$$\begin{aligned}
\mathbf{m} &= \Phi_i^T \mathbf{M} \Phi_i = \Phi_{ii}^T \mathbf{M}_{ii} \Phi_{ii} \\
\mathbf{M}_{ib} &= \Phi_i^T \mathbf{M} \Psi_c = \Phi_{ii}^T \Big[\mathbf{M}_{ib} + \mathbf{M}_{ii} \Psi_{ib} \Big] \\
\mathbf{M}_{bi} &= \Psi_c^T \mathbf{M} \Phi_i = \Big[\mathbf{M}_{bi} + \Psi_{ib}^T \mathbf{M}_{ii} \Big] \Phi_{ii} \\
\mathbf{M}_{bb} &= \Psi_c^T \mathbf{M} \Psi_c = \mathbf{M}_{bb} + \mathbf{M}_{bi} \Psi_{ib} + \Psi_{ib}^T \mathbf{M}_{bi} + \Psi_{ib}^T \mathbf{M}_{ii} \Psi_{ib} \\
\mathbf{k} &= \Phi_i^T \mathbf{K} \Phi_i = \Phi_{ii}^T \mathbf{K}_{ii} \Phi_{ii} = \mathbf{m} \omega_E^2 \\
\mathbf{K}_{ib} &= \Phi_i^T \mathbf{K} \Psi_c = \Phi_{ii}^T \Big[\mathbf{K}_{ib} - \mathbf{K}_{ib} \Big] = 0 \\
\mathbf{K}_{bi} &= \Psi_c^T \mathbf{K} \Phi_i = \Big[\mathbf{K}_{bi} - \mathbf{K}_{ib} \Big] \Phi_{ii} \ = 0 \\
\mathbf{K}_{bb} &= \Psi_c^T \mathbf{K} \Psi_c = \mathbf{K}_{bb} - \mathbf{K}_{bi} \mathbf{K}_{ii}^{-1} \mathbf{K}_{ib} \\
\mathbf{c} &= \Phi_i^T \mathbf{D} \Phi_i = \Phi_{ii}^T \mathbf{D}_{ii} \Phi_{ii} = 2\zeta \mathbf{m} \omega_E
\end{aligned}$$

The mass matrix $\mathbf{M}_{bb}$ and the stiffness matrix $\mathbf{K}_{bb}$ are reduced to the boundary nodes in the same way as they are reduced using the Guyan-Irons reduction. $\mathbf{m}$, $\mathbf{c}$ and $\mathbf{k}$ are the modal mass, modal damping and modal stiffness matrix of the interior dofs respectively. $\Psi_c^T \mathbf{D} \Psi_c$, $\Psi_c^T \mathbf{D} \Phi_i$ and $\Phi_i^T \mathbf{D} \Psi_c$ are considered equal to $\mathbf{0}$. Thus only the submatrix $2\zeta \mathbf{m} \omega_E$ has significance [8].

5 MOR: Interpolatory Methods

The standard formulation of the second order dynamical system, (2), in the frequency domain is given by

$$\begin{aligned}
(s^2 \mathbf{M} + s\mathbf{D} + \mathbf{K}) \mathbf{u}(s) &= \mathbf{B} \mathbf{f} \mathrm{f}(s) \\
\mathbf{y}(s) &= \mathbf{C} \mathbf{u}(s)
\end{aligned} \tag{67}$$

where $s = i\omega$, $\mathbf{B}$ and $\mathbf{C}$ can be considered equal to the identity matrix $\mathbf{I}$ and the function $\mathrm{f}(s)$ equal to one in case of a response having the same form of the excitation force, i.e. the input vector is constant. The transfer function of eq. (67) is defined by

$$\mathcal{H} = \mathbf{C}(s^2 \mathbf{M} + s\mathbf{D} + \mathbf{K})^{-1} \mathbf{B} \mathbf{f} \tag{68}$$

The main approach of an interpolatory MOR method is to match the first ν_p terms of the Taylor expansion of eq. (68) about ω_p with $p = 0, 1, 2, ..., P$, i.e. about P frequency points in a frequency interval $\Delta_\omega = [\omega_L,\ \omega_R]$, which is the same as to find an approximated vector $\mathbf{u}_\nu$ which matches the first ν terms of the Taylor expansion of the vector $\mathbf{u}$ about the same ω_p frequency points in the context of structural mechanics problems, i.e.

$$\mathbf{u}(s) = \mathbf{u}_\nu(s) + \mathcal{O}\Bigg(\prod_{p=1}^{P} (\Delta s_p)^{\nu_p} \Bigg) \tag{69}$$

where $\Delta s_p = s - s_p$ and ν_p is the order of the Taylor series expansion at the interpolation frequency ω_p. Thus $\mathbf{u}_\nu^{(k)}(s_p) = \frac{d^k}{ds^k}\big(\mathbf{u}(s)\big)$. Therefore the main task for interpolatory MOR methods is to find a vector space that contains the first ν moments (see, sub-sec. 5.2) of $\mathcal{H}$. In the case of the multi-point Padé approximation such an approach is simplified to find the coefficients of the Padé approximation $[L/M]$ of a single dof, which implies to match the first $L + M + 1$ moments of the transfer function, eq. (68). Due to the matching process and the analogy to the Padé approximant those methods are also known as moment matching methods or Padé type methods [9, 10]. Here a similar terminology as in [14] was adopted.

5.1 Multi-Padé approximation

Every dof $\mathrm{u}_j(s)$, $j = 1, 2, 3, ..., N$, of the displacement vector can be approximated by a Padé approximant about a frequency point ω_0. The Padé approximant of the jth dof has the form

$$\begin{aligned} u_j(\omega) = [L/M] &= \frac{p_0 + p_1\Delta\omega_0 + \cdots + p_L\Delta\omega_0^L}{1 + q_1\Delta\omega_0 + \cdots + q_M\Delta\omega_0^M} + \mathcal{O}(\Delta\omega_0^{L+M+1}) \\ &= \frac{P_j(\omega)}{Q_j(\omega)} + \mathcal{O}(\Delta\omega_0^{L+M+1}) \end{aligned} \tag{70}$$

where $\omega = \omega_0 + \Delta\omega$, $P_j(\omega)$ and $Q_j(\omega)$ is a polynomial of degree L and M respectively. By Cross-multiplying eq. (70) and disregarding the error of the right hand side leads to

$$u_j(\omega)Q_j(\omega) = P_j(\omega) \tag{71}$$

Successively differentiating eq. (71) with respect to ω the following equation is obtained [13]

$$\sum_{r=0}^{k}\left(\frac{k!}{r(k-r)!}u_j^{(k-r)}\sum_{m=r}^{M}\frac{m!}{(m-r)!}q_m\Delta\omega^{m-r}\right) - \sum_{l=k}^{L}\frac{l!}{(l-k)!}p_l\Delta\omega^{l-k} = 0 \qquad k = 0, 1, \ldots, L+M \tag{72}$$

with

$$\begin{aligned} q_l &= 0 \quad if \quad r > M \\ p_k &= 0 \quad if \quad k > L \end{aligned} \tag{73}$$

where $u_j^k = \left(\frac{d^k}{d\omega^k}\big(\mathrm{u}(\omega_0)\big)\right)_j$ and $\Delta\omega = \omega - \omega_0$. Then the coefficients of the polynomials P_j and Q_j can be computed if u_j and its first $L + M$ derivatives are known, i.e. the Padé approximant given by eq.(70) matches the first $L + M + 1$ terms of the corresponding Taylor series about ω_0, therefore the Padé approximant matches the first $L + M + 1$ terms (moments) of a power expansion of the transfer function given by eq. (68) [9].

It is said to be a multi-point Padé Approximation [13] if P frequency points are defined and interpolated in the frequency interval of interest, Δ_ω, where each frequency is denoted by ω_p with $p = 1, 2, 3, ..., P$, using

$$\sum_{r=0}^{k}\left(\frac{k!}{r!(k-r)!}u_{j,\omega_p}^{(k-r)}\sum_{m=r}^{M}\frac{m!}{(m-r)!}q_m\Delta\omega_p^{m-r}\right) - \sum_{l=k}^{L}\frac{l!}{(l-k)!}p_l\Delta\omega_p^{l-k} = 0 \qquad k = 0, 1, \ldots, \left\lceil\frac{L+M+1}{P}\right\rceil - 1 \tag{74}$$

where $\lceil\ \rceil$ is the ceiling function and $u_{j,\omega_p}^{(k)}$ is the k derivative of the jth dof of the displacement vector $\mathbf{u}$ evaluated at ω_p and is defined by

$$u_{j,\omega_p}^{(k)} = (\mathbf{u}_{\omega_p}^{(k)})_j = \begin{cases} \mathbf{K}_\omega^{-1}\mathbf{f} & k = 0 \\ \mathbf{K}_\omega{}^{-1}\big((2\omega_p\mathbf{M} - i\mathbf{D})\mathbf{u}_{\omega_p} + \mathbf{f}^{(1)}\big) & k = 1 \\ \mathbf{K}_\omega^{-1}\big(k\mathbf{M}(2\omega_p\mathbf{u}_{\omega_p}^{(k-1)} + (k-1)\mathbf{u}_{\omega_p}^{(k-2)} - ik\mathbf{D}\mathbf{u}_{\omega_p}^{(k-1)}\big) & k \geq 2 \end{cases} \tag{75}$$

where $\Delta\omega_p = \omega_p - \omega_0$ and $\mathbf{K}_\omega = (-\omega_p^2\mathbf{M} + i\omega\mathbf{D} + \mathbf{K})$. Eq. (75) implies to solve P times a generalized impedance problem with $\left\lceil\frac{L+M+1}{P}\right\rceil$ different right hand sides.

From eq. (75) it is possible to obtain the coefficients of the Padé polynomials for each component u_j of $\mathbf{u}$, those coefficients are collected in the vector $\mathbf{x} = [p_0\, p_1\, p_2\, ...\, p_L\, q_1\, q_2\, ...\, q_M]^T$ and are found by solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ which is formed as follows:

$$\begin{bmatrix} \begin{bmatrix} \cdots & \mathbf{A}_1 & \cdots \end{bmatrix}_{[n_\nu \times n_P]} \\ \vdots \\ \begin{bmatrix} \cdots & \mathbf{A}_p & \cdots \end{bmatrix}_{[n_\nu \times n_P]} \\ \vdots \\ \begin{bmatrix} \cdots & \mathbf{A}_P & \cdots \end{bmatrix}_{[n_\nu \times n_P]} \end{bmatrix}_{[n_P \times n_P]} \qquad \mathbf{x}_{[n_P \times 1]} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_p \\ \vdots \\ \mathbf{b}_P \end{bmatrix}_{[n_P \times 1]} \tag{76}$$

where $\mathbf{A}_p \in \mathbb{C}^{\lceil \frac{(L+M+1)}{P} \rceil \times (L+M+1)}$, $\mathbf{b}_p \in \mathbb{C}^{\lceil \frac{(L+M+1)}{P} \times 1}$, $p = 1, 2, \ldots, P$, $n_\nu = \lceil \frac{L+M+1}{P} \rceil$ and $n_P = L + M + 1$. $\mathbf{A}$ is a symmetric matrix but ill-conditioned, thus this approach can be used only when few interpolation points are considered [9].

$\mathbf{A}_p$ and $\mathbf{b}_p$ are computed by expanding eq. (74) for $k = 0, 1, \ldots, \left\lceil \frac{L+M+1}{P} \right\rceil - 1$, i.e the row $k+1$ of $\mathbf{A}_p$ is an equation given by

$$\begin{aligned} -\tfrac{(k)!}{0!} p_k - \tfrac{(k+1)!}{1!} (\Delta\omega_p)^1 p_{k+1} - \tfrac{(k+2)!}{2!} (\Delta\omega_p)^2\, p_{k+2} - \cdots - \tfrac{L!}{(L-k)!} (\Delta\omega_p)^{L-k}\, p_L \\ + \big(\textstyle\sum_{r=0}^{k} \tfrac{k!}{r!(k-r)!} \tfrac{1!}{(1-r)!} (\Delta\omega)^{1-r} u^{(k-r)} \big) q_1 \\ + \big(\textstyle\sum_{r=0}^{k} \tfrac{k!}{r!(k-r)!} \tfrac{2!}{(2-r)!} (\Delta\omega)^{2-r} u^{(k-r)} \big) q_2 \\ \cdots + \big(\textstyle\sum_{r=0}^{k} \tfrac{k!}{r!(k-r)!} \tfrac{M!}{(M-r)!} (\Delta\omega)^{M-r} u^{(k-r)} \big) q_M = u_p^{(k)} \end{aligned} \tag{77}$$

Then the coefficients of the left hand side of eq. (77) are placed in the row $k+1$ of $\mathbf{A}_p$, and the right hand side is placed at the $k+1$ position of the vector $\mathbf{b}_p$. It permits to express eq. (74) in a matrix notation of the form $[\mathbf{A}_p]_{[n_\nu \times n_P]} [\mathbf{x}]_{[n_P \times 1]} = [\mathbf{b}_p]_{[n_P \times 1]}$ as follows

$$\begin{bmatrix} -1 & -(\Delta\omega_p) & -(\Delta\omega_p)^2 & \cdots & -(\Delta\omega_p)^L & (\Delta\omega_p) u_p & (\Delta\omega_p)^2 u_p & \cdots & (\Delta\omega_p)^M u_p \\ 0 & -1 & -2(\Delta\omega_p) & \cdots & L(\Delta\omega_p)^{L-1} & \left(u_p^{(0)} + (\Delta\omega_p) u_p^{(1)}\right) & \left(2(\Delta\omega_p) u_p^{(0)} + (\Delta\omega_p)^2 u_p^{(1)}\right) & \cdots & \left(M(\Delta\omega_p)^{M-1} U_p^{(0)} + (\Delta\omega_p)^M u_p^{(1)}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_L \\ q_1 \\ q_2 \\ \vdots \\ q_M \end{bmatrix} = \begin{bmatrix} u_p^{(0)} \\ u_p^{(1)} \\ \vdots \end{bmatrix} \tag{78}$$

Thus P small systems of equations, which share the same unknown vector $\mathbf{x}$, of the form given by eq.(78) are assembled, i.e they are gathered in the assembled-system of equations given by eq.(76), which permits to compute the coefficients of the multi-point Padé approximation with a very low computational cost.

5.2 Krylov-based Galerkin projection Method

First Order approach

The second order dynamical system in eq. (67) is reformulated in the standard first order form, see sub-sec. 2.2, and given by

$$\begin{aligned} s\, \mathbf{z}(s) &= \mathbf{A}\mathbf{z}(s) + \mathbf{B}\mathbf{f}(s) \\ \mathbf{y}(s) &= \mathbf{C}\mathbf{z}(s) \end{aligned} \tag{79}$$

Where $\mathbf{z} \in \mathbb{R}^{2N\times 1}$, $\mathbf{A} \in \mathbb{R}^{2N\times 2N}$ and $\mathbf{B} \in \mathbb{R}^{2N\times N}$ are the state vector, the state matrix and the input matrix respectively and are defined by

$$\mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \dot{\mathbf{u}} \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{D} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1}\mathbf{f} \end{bmatrix} \tag{80}$$

The transfer matrix of (79) is given by

$$\mathbf{h}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \tag{81}$$

If $\mathbf{A}$ is non singular, the transfer matrix $\mathbf{h}(s)$ can be approximated about $s_0 = 0$ using the Taylor series as follows

$$\mathbf{h}(s) = \mathbf{C}\big((-\mathbf{A}^{-1}\mathbf{B}) + (\mathbf{A}^{-1})(-\mathbf{A}^{-1}\mathbf{B})s + \cdots + (\mathbf{A}^{-1})^i(-\mathbf{A}^{-1}\mathbf{B})s^i + \cdots\big) \tag{82}$$

For convenience one may define

$$\boldsymbol{\mathcal{H}} = \mathbf{A}^{-1} = \begin{bmatrix} -\mathbf{K}^{-1}\mathbf{D} & -\mathbf{K}^{-1}\mathbf{M} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \mathbf{b} = \mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{K}^{-1}\mathbf{f} \\ \mathbf{0} \end{bmatrix} \tag{83}$$

and

$$\boldsymbol{b}_0 = \mathbf{K}^{-1}\mathbf{f} \quad \boldsymbol{\mathcal{H}}_{11} = -\mathbf{K}^{-1}\mathbf{D} \quad \boldsymbol{\mathcal{H}}_{12} = -\mathbf{K}^{-1}\mathbf{M} \tag{84}$$

The resulting non-negative coefficients of the series (82) are said to be the system's *moments*

$$\mathbf{m}_i = \mathbf{C}(\mathbf{A}^{-1})^i(\mathbf{A}^{-1}\mathbf{B}) = \mathbf{C}(\boldsymbol{\mathcal{H}})^i\mathbf{B}, \quad i = 0, 1, 2, \ldots \tag{85}$$

where $\mathbf{m}_i \in \mathbb{C}^{2N\times 1}$ is the ith moment of eq. (81). It is clear from eq. (85) that $\mathbf{z}(s)$ of (80) is contained in a vector space spanned by the columns of the Krylov subspace $\mathcal{K}_\nu(\boldsymbol{\mathcal{H}},\ \mathbf{b})$ with starting vector $\mathbf{b}$ and matrix $\boldsymbol{\mathcal{H}}$. A simplification of the dimension of the vectors in the Krylov subspace can be done if one notes that

$$\mathbf{m}_0 = \mathbf{b} = \begin{bmatrix} \mathbf{K}^{-1}\mathbf{f} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_0 \\ \mathbf{0} \end{bmatrix} \qquad \mathbf{m}_2 = \boldsymbol{\mathcal{H}}^2\mathbf{b} = \begin{bmatrix} \boldsymbol{\mathcal{H}}_{11}\boldsymbol{b}_1 + \boldsymbol{\mathcal{H}}_{12}\boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_2 \\ \boldsymbol{b}_1 \end{bmatrix}$$

$$\mathbf{m}_1 = \boldsymbol{\mathcal{H}}\mathbf{b} = \begin{bmatrix} \boldsymbol{\mathcal{H}}_{11}\boldsymbol{b}_0 \\ \boldsymbol{b}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_0 \end{bmatrix} \qquad \mathbf{m}_3 = \boldsymbol{\mathcal{H}}^3\mathbf{b} = \begin{bmatrix} \boldsymbol{\mathcal{H}}_{11}\boldsymbol{b}_2 + \boldsymbol{\mathcal{H}}_{12}\boldsymbol{b}_1 \\ \boldsymbol{b}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_3 \\ \boldsymbol{b}_2 \end{bmatrix}$$

therefore the displacement part $\mathbf{u}(s)$ of the vector $\mathbf{z}(s)$ is contained in the column space spanned by $\{\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, ...\}$, i,e. by the Krylov subspace $\mathcal{K}_\nu(\boldsymbol{\mathcal{H}}_{11},\ \boldsymbol{\mathcal{H}}_{12};\ b_0)$,

$$\mathbf{u}(\omega) \in \mathcal{K}_\nu(-\mathbf{K}^{-1}\mathbf{D},\ -\mathbf{K}^{-1}\mathbf{M};\ \mathbf{K}^{-1}\mathbf{f}) \tag{86}$$

which holds true for low frequency range, because the interpolation point s_0 was chosen equal to zero.

It remains to show how to get the Krylov subspace for $s \neq 0$, such that the matrix $\mathbf{V}$ contains information of many points, i.e. a interpolatory approach over the frequency interval Δ_ω. It is achieved by noting that the moments of $\mathbf{h}(s)$ about $s_0 \neq 0$ are equal to the moments of $\mathbf{h}(s+s_0)$ about zero [12], i.e the transfer function of eq. (81) is given by

$$\begin{aligned} \mathbf{h}(s + s_0) &= \mathbf{C}((s + s_0)^2\mathbf{M} + (s + s_0)\mathbf{D} + \mathbf{K})^{-1}\mathbf{B} \\ &= \mathbf{C}(s^2\mathbf{M} + s(2s_0\mathbf{M} + \mathbf{D}) + (s_0^2\mathbf{M} + (s_0\mathbf{D} + \mathbf{K}))^{-1}\mathbf{B} \end{aligned} \tag{87}$$

For convenience one may define

$$\mathbf{D}_s = 2s_0\mathbf{M} + \mathbf{D} \;\text{ and }\; \mathbf{K}_s = s_0^2\mathbf{M} + s_0\mathbf{D} + \mathbf{K} \tag{88}$$

$\mathbf{D}_s$ and $\mathbf{K}_s$ is the generalized damping and the dynamic impedance respectively. The substitution of $\mathbf{D}$ and $\mathbf{K}$ in the Krylov subspace in eq. (86) by the definitions given by eq. (88) leads to

$$\mathcal{K}_\nu(-\mathbf{K}_s^{-1}\mathbf{D}_s,\ -\mathbf{K}_s^{-1}\mathbf{M}; \mathbf{K}_s^{-1}\mathbf{f}) \tag{89}$$

which approximates the solution $\mathbf{u}$ for any s_0. Using $s = i\omega$ and defining P frequencies as interpolation points in Δ_ω a vector space $\mathbf{V}$ may be defined by

$$\begin{aligned}\mathbf{V}_{[N\times 3P]} = \bigcup_{p=1}^{P} \mathcal{K}_{\nu_p}\Big(&-(-\omega_p^2\mathbf{M} + i\omega_p\mathbf{D} + \mathbf{K})^{-1}(2i\omega\mathbf{M} + \mathbf{D}),\\ &-(-\omega_p^2\mathbf{M} + i\omega_p\mathbf{D} + \mathbf{K})^{-1}\mathbf{M};\ (-\omega_p^2\mathbf{M} + i\omega_p\mathbf{D} + \mathbf{K})^{-1}\mathbf{f}\Big)\end{aligned} \tag{90}$$

where ω_p is the interpolation frequency, with $p = 1, 2, ..., P$ and ν_p is the number of matched terms about ω_p of the Taylor expansion of the transfer matrix given by eq.(81). If the dynamic damping is neglected eq. (90) simplifies to

$$\mathbf{V}_{[N\times 3P]} = \bigcup_{p=1}^{P} \mathcal{K}_\nu\big(-(-\omega_p^2\mathbf{M} + i\omega_p\mathbf{D} + \mathbf{K})^{-1}\mathbf{M}, (-\omega_p^2\mathbf{M} + i\omega_p\mathbf{D} + \mathbf{K})^{-1}\mathbf{f}\big) \tag{91}$$

which is the suitable form for proportional (Rayleigh damping and modal damping) damped systems.

Second order approach

The transfer matrix of the second order dynamical system of eq. (67) is given by eq. (68) and rewritten here for convenience

$$\mathcal{H}(s) = \mathbf{C}(s^2\mathbf{M} + s\mathbf{D} + \mathbf{K})^{-1}\mathbf{B}\mathbf{f} = \mathbf{C}\mathbf{K}_s\mathbf{B}\mathbf{f} \tag{92}$$

The Taylor expansion of eq. (92) is given by

$$\mathcal{H}(s_0 \neq 0) = \sum_k^\infty \frac{1}{k!}\frac{\mathrm{d}^{(k)}}{ds^{(k)}}\big(\mathcal{H}(s_0)\big)(\Delta s)^k \tag{93}$$

therefore the moments of eq. (93) are

$$\mathbf{m}_k = \frac{1}{k!}\frac{\mathrm{d}^{(k)}}{ds^{(k)}}\big(\mathcal{H}(s_0)\big) \quad k = 0, 1, 2, 3, ..., \nu \tag{94}$$

By solving the first ν moments, eq. (94), of the Taylor series given by eq. (93) and considering $\mathbf{C}$ and $\mathbf{B}$ equal to the identity matrix it is possible to group together their terms such that a Krylov subspace of the form given by (89) is obtained. It is illustrated by solving the first four

moments, using eq. (94), of the Taylor series about s_0, i.e.

$$\mathbf{m}_0 = \mathbf{K}_s^{-1}\mathbf{f} = \boldsymbol{b}_0 \tag{95}$$

$$\mathbf{m}_1 = -\mathbf{K}_s^{-2}(2s_0\mathbf{M} + \mathbf{D})\mathbf{f} = (-\mathbf{K}_s^{-1}\mathbf{D}_s)\boldsymbol{b}_0 \tag{96}$$

$$\begin{aligned}\mathbf{m}_2 &= \mathbf{K}_s^{-3}(2s_0\mathbf{M} + \mathbf{D})^2\mathbf{f} - \mathbf{K}_s^{-2}(\mathbf{M})\mathbf{f} \\ &= (-\mathbf{K}_s^{-1}\mathbf{D}_s)(-\mathbf{K}_s^{-1}\mathbf{D}_s)\mathrm{b}_0 + (-\mathbf{K}_s^{-1}\mathbf{M})\boldsymbol{b}_0\end{aligned} \tag{97}$$

$$\begin{aligned}\mathbf{m}_3 &= -\mathbf{K}_s^{-4}(2s_0\mathbf{M} + \mathbf{D})^3\mathbf{f} + \frac{8}{6}\mathbf{K}_s^{-3}(2s_0\mathbf{M} + \mathbf{D})(\mathbf{M})\mathbf{f} \\ &\quad + \frac{4}{6}\mathbf{K}_s^{-3}(2s_0\mathbf{M} + \mathbf{D})(\mathbf{M})\mathbf{f} \\ &= (-\mathbf{K}_s^{-1}\mathbf{D}_s)(-\mathbf{K}_s^{-1}\mathbf{D}_s)(-\mathbf{K}_s^{-1}\mathbf{D}_s)\mathrm{b}_0 + (-\mathbf{K}_s^{-1}\mathbf{D}_s)(-\mathbf{K}_s^{-1}\mathbf{M})\boldsymbol{b}_0 \\ &\quad + (-\mathbf{K}_s^{-1}\mathbf{M})(-\mathbf{K}_s^{-1}\mathbf{D}_s)\boldsymbol{b}_0\end{aligned} \tag{98}$$

These show that the moments $\mathbf{m}_0, \mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_\nu$ of the second order dynamical system given by eq. (67) span the Krylov subspace defined by eq. (89) and vice versa.

5.3 Derivative-based Galerkin projection Method

The transfer function of the original second order dynamical system of eq. (67) can be redefined as

$$\mathcal{H}(s) = \mathbf{C}(s^2\mathbf{M} + s\mathbf{D} + \mathbf{K})^{-1}\mathbf{B} = \mathbf{C}\mathbf{K}_\omega^{-1}\mathbf{B} \tag{99}$$

$\mathbf{f}\mathrm{f}(s)$ of eq. (67) is interpreted as a constant input with $\mathrm{f}(s) = 1$. Analogously the transfer function of a reduced second order system is defined by

$$\mathcal{H}_\nu(s) = \mathbf{C}\mathbf{V}(s^2\mathbf{M}_\mathbf{r} + s\mathbf{D}_r + \mathbf{K}_r)^{-1}\mathbf{V}^H\mathbf{B} = \mathbf{C}\mathbf{V}\mathbf{K}_{s,\nu}^{-1}\mathbf{V}^H\mathbf{B} \tag{100}$$

One searches for a $\mathcal{H}_\nu$ which acting in a vector $\mathbf{f}$ approximates the original transfer function acting on the same vector, i.e.

$$\| \mathcal{H}(s_p + \Delta s)\mathbf{f} - \mathcal{H}_\nu(s_p + \Delta s)\mathbf{f} \ \| \leq c|\Delta s|^h \tag{101}$$

where Δs is sufficient small and $h > 0$. An equivalent expression to eq. (101) is given by

$$\mathcal{H}(s)\mathbf{f} = \mathcal{H}_\nu(s)\mathbf{f} + \mathcal{O}((\Delta s)^\nu) \tag{102}$$

which implies

$$\frac{d^k}{ds^k}(\mathcal{H}) = \frac{d^k}{ds^k}(\mathcal{H}_\nu) \quad for \ \ k = 0, 1, 2, 3, ..., \nu - 1 \tag{103}$$

Thus if $\frac{d^k}{ds^k}((s_p^2\mathbf{M} + 2s_p\mathbf{D} + \mathbf{K})^{-1}\mathbf{f}) \in \mathbf{V}$ for $k = 0, 1, 2, ..., \nu - 1$ then one proofs eq. (101) as follows [11]

$$\|\mathcal{H}(s_p + \Delta s)\mathbf{f} - \mathcal{H}_\nu(s_p + \Delta s)\mathbf{f}\| = \|\mathbf{C}\mathbf{K}_s^{-1}(s_p + \Delta s)\mathbf{f} - \mathbf{C}\mathbf{V}\mathbf{K}_{s,\nu}^{-1}(s_p + \Delta s)\mathbf{V}^H\mathbf{f}\|$$

Using the skew projection $\mathbf{P}_\nu = \mathbf{V}\mathbf{K}_{s,\nu}^{-1}(s+\Delta s)\mathbf{V}^H\mathbf{K}_s(s+\Delta s)$

$$\begin{aligned}
&\|\mathbf{C}\mathbf{K}_s^{-1}(s_p+\Delta s)\mathbf{f} - \mathbf{C}\mathbf{V}\mathbf{K}_{s,\nu}^{-1}(s_p+\Delta s)\mathbf{V}^H\mathbf{f}\| \\
&= \|\mathbf{C}\mathbf{K}_s^{-1}(s_p+\Delta s)\mathbf{f} - \mathbf{C}\mathbf{P}_\nu\mathbf{K}_s^{-1}(s_p+\Delta s)\mathbf{f}\| \\
&= \|\mathbf{C}(\mathbf{I}-\mathbf{P}_\nu)\sum_{k=0}^{\infty}\frac{1}{k!}\frac{d^{(k)}}{ds^{(k)}}\big(\mathbf{K}_s^{-1}(s_p+\Delta s)\big)\Delta s^k\mathbf{f}\| \\
&= \|\mathbf{C}\sum_{k=\nu}^{\infty}\frac{1}{k!}\frac{d^{(k)}}{ds^{(k)}}\big(\mathbf{K}_s^{-1}(s_p+\Delta s)\big)\Delta s^k\mathbf{f}\| \\
&= |\Delta s|^\nu \|\mathbf{C}\sum_{k=0}^{\infty}\frac{1}{(k+\nu)!}\frac{d^{(k+\nu)}}{ds^{(k+\nu)}}\big(\mathbf{K}_s^{-1}(s_p+\Delta s)\big)\Delta s^k\mathbf{f}\| \\
&\leq c|\Delta s|^\nu
\end{aligned} \tag{104}$$

where c is chosen uniformly for all Δs sufficient small. For the second order dynamical system given by eq. (67) the matrix $\mathbf{C}$ is considered equal to the identity matrix and one is interested in the state vector $\mathbf{u}$, therefore taking the vector $\mathbf{f}$ of eq. (102) equal to the force vector, and the fact that $s = i\omega$, eq. (102) holds true if

$$\bigcup_{p=1}^{P}\text{span}\left\{\mathbf{u}(\omega_p), \frac{d}{d\omega}\big(\mathbf{u}(\omega_p)\big), \cdots, \frac{d^{\nu-1}}{d\omega^{\nu-1}}\big(\mathbf{u}(\omega_p)\big)\right\} \subset \text{Ran}(\mathbf{V}) \tag{105}$$

$\mathbf{u}$ and its derivatives are found by solving eq. (75) for $k = 0, 1, 2, ..., \nu-1$.

6 Numerical Examples

Two numerical example are addressed in this section. In the first example a frequency response analysis using the ten MOR techniques mentioned in section 1 are applied to a cantilever beam in order to evaluate the quality of the approximation of those methods for proportional and non-proportional damped systems. In the second example a medium-sized structure is investigated in order to evaluate the efficiency in computation time of the Interpolatory MOR methods. In both examples the solutions given by the MOR methods are compared with the solution of the full-order model (DM).

6.1 Cantilever Beam

A cantilever beam has been discretized in 10 finite elements, as illustrated in fig.4,

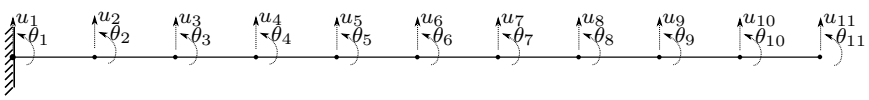


Figure 4: FE discretization of a catiliver beam.

The beam has a length of $1m$, cross section of $0.1m \times 0.1m$, modulus of elasticity of $2.068e11\frac{N}{m^2}$ and material density of 7830 $\frac{kg}{m^3}$. It is loaded in the vertical direction with a distributed load of -100 $\frac{N}{m}$. $\Delta_\omega = [1, 700]$ and $n_\omega = 700$. In the Real Modal Analysis (RMA) and the complex Modal Analysis (CMA) 20 modes shapes are considered. The condensation methods use the vertical displacement as the active set of dofs, $\mathbf{u}_a$. The Dynamic Reduction

(DR) is done about $\omega_{dy} = 50$ and the IRS reduction method uses 3 iterations. Craig-Bampton method uses 5 fixed-interface normal modes. For the Interpolatory MOR methods ω_p ={1, 30, 80, 110,300, 500, 700}. The response of the dof u_{11} will be compared for a Rayleigh-damped and non-proportional damped beam.

6.1.1 Rayleigh Damped Cantilever Beam

The Rayleigh damping matrix is defined as $\mathbf{C} = a_0\mathbf{M} + a_1\mathbf{K}$. In this example $a_0 = 0.0002$ and $a_1 = 0.0001$. The ten MOR methods are able to capture the exact solution in amplitude and phase angle, see figs. 5 and 6, even the GR, and IRS which are in general only suitable for a low frequency range.

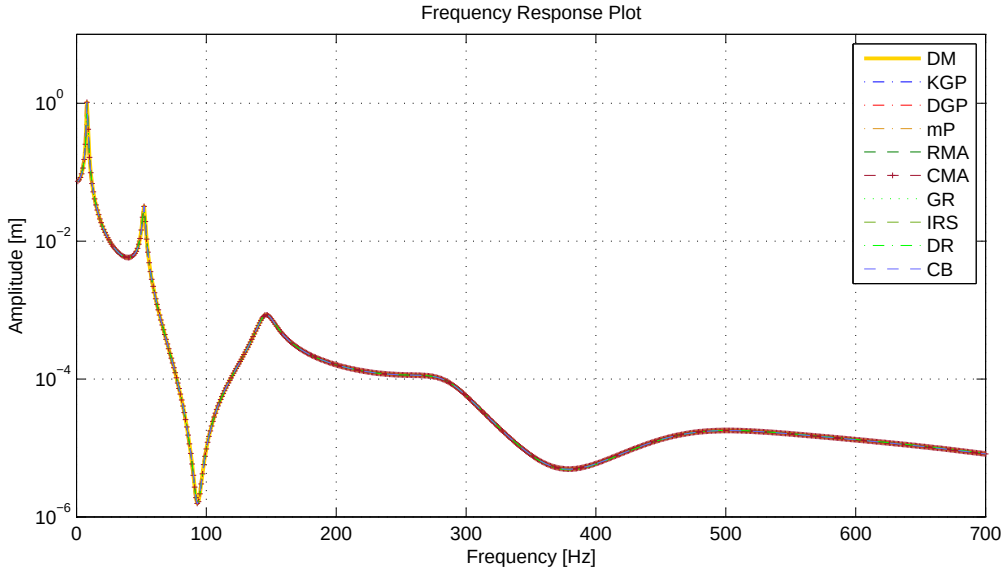


Figure 5: Amplitude of the response for u_{11} of the Rayleigh-damped cantilever beam.

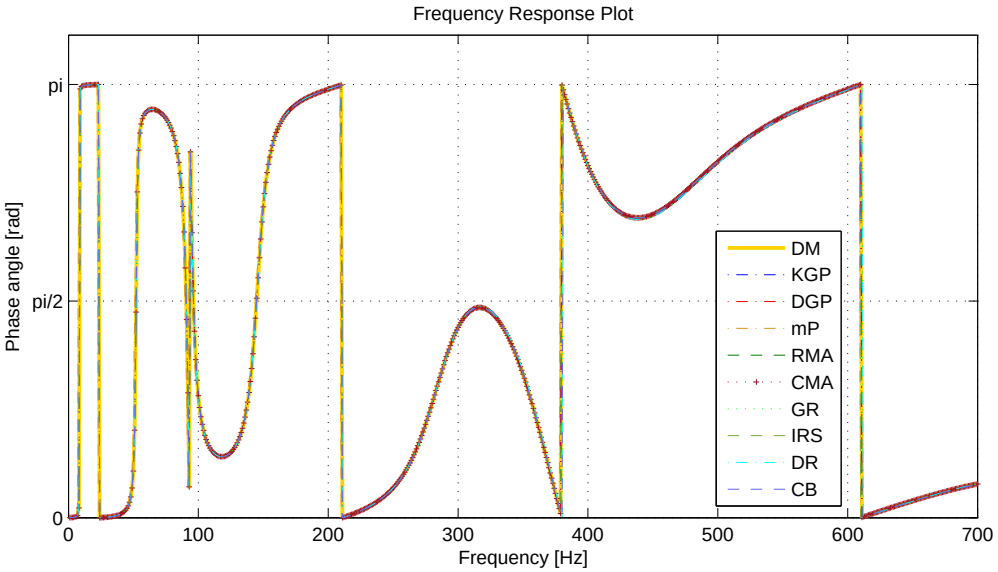


Figure 6: Phase angle of the response for u_{11} of the Rayleigh-damped cantilever beam.

6.1.2 Non-proportional Damped Cantilever Beam

For an arbitrary damped cantilever beam a random dense matrix was defined in MATLAB R2015b® using the command $randn$ with a $seed$ value of 0.1, a mean of 0.15 and a standard deviation of 0.1. That matrix does not have a physical meaning due to the fact that the matrix was not derived according to the assembling of the elements, nevertheless it is important to see how the MOR methods behave for a simple arbitrary damped system like the 10-elements cantilever beam.

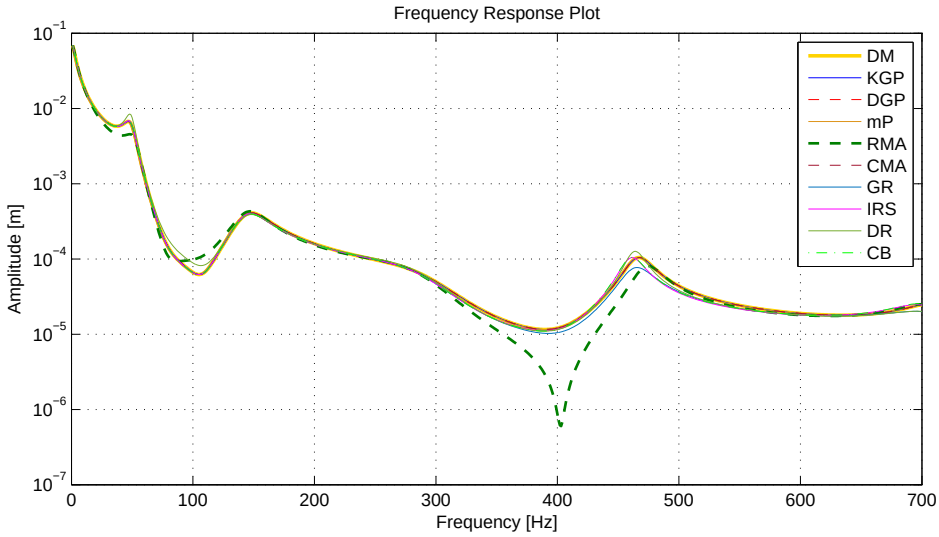


Figure 7: Amplitude of the response for u_{11} of the non-proportional damped cantilever beam.

From figs. 7 and 8 it is clear that the RMA gives a result far from the exact solution, the condensations methods and the CB method do not capture the exact solution any more as the Interpolatory MOR Methods (DGP, KGP and the mP) and the CMA do.

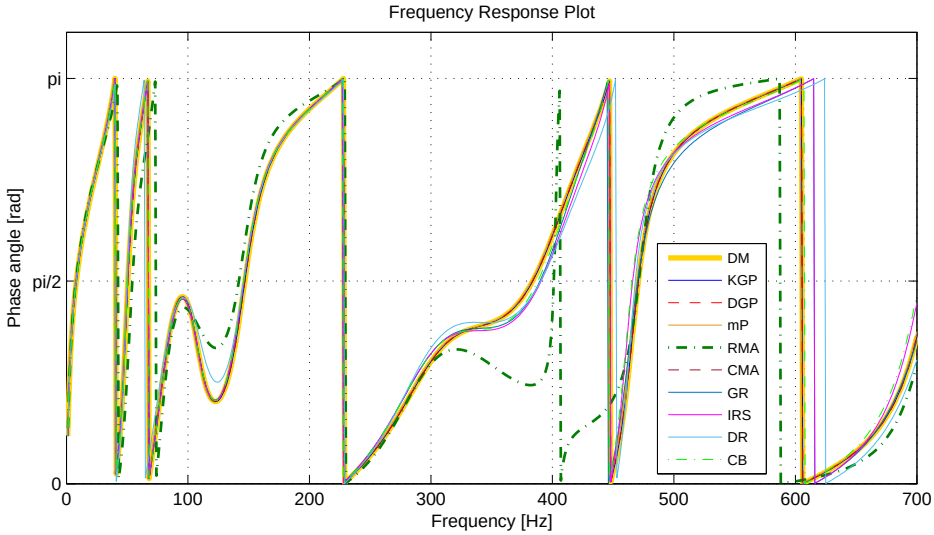


Figure 8: Phase angle of the response for u_{11} of the non-proportional damped cantilever beam.

If the damping increases, the capability of the Interpolatory MOR Methods are more evident. In figs. 9 and 10, the response of u_{11} is shown in the case of an arbitrary damping matrix, created in the same way as before, but with mean value and standard deviation equal to 0.35 and 0.2 respectively. It results in a more evident lack of approximation for the RMA, Condensation Methods and CB Method, in contrast with the Interpolatory MOR Methods and the CMA which lead to the exact solution. Nevertheless the CMA is not suitable for large systems due to the expensive computation of the complex modes, so the Interpolatory MOR Methods remain as a very attractive possibility to be used as standard MOR techniques for frequency response analysis in case of non-proportional damped systems.

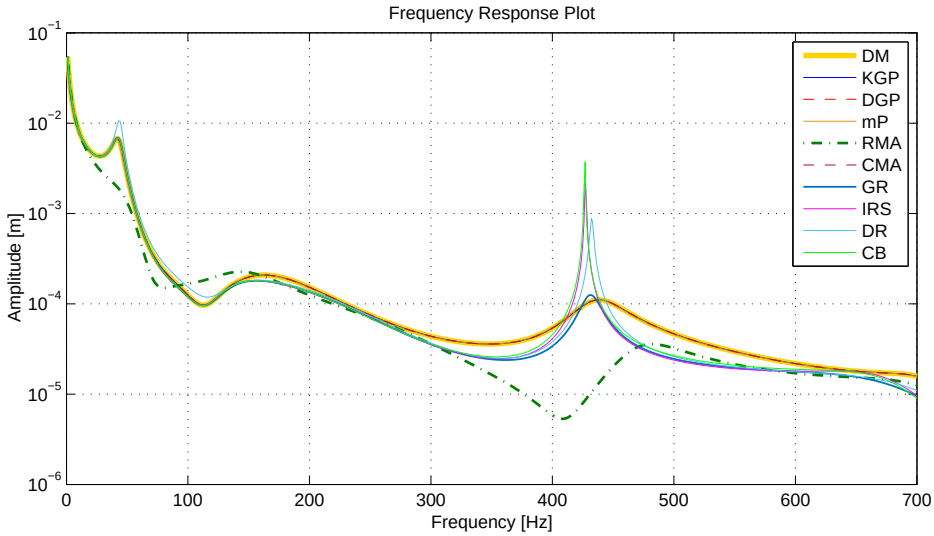


Figure 9: Amplitude of the response for u_{11} of the non-proportional damped cantilever beam

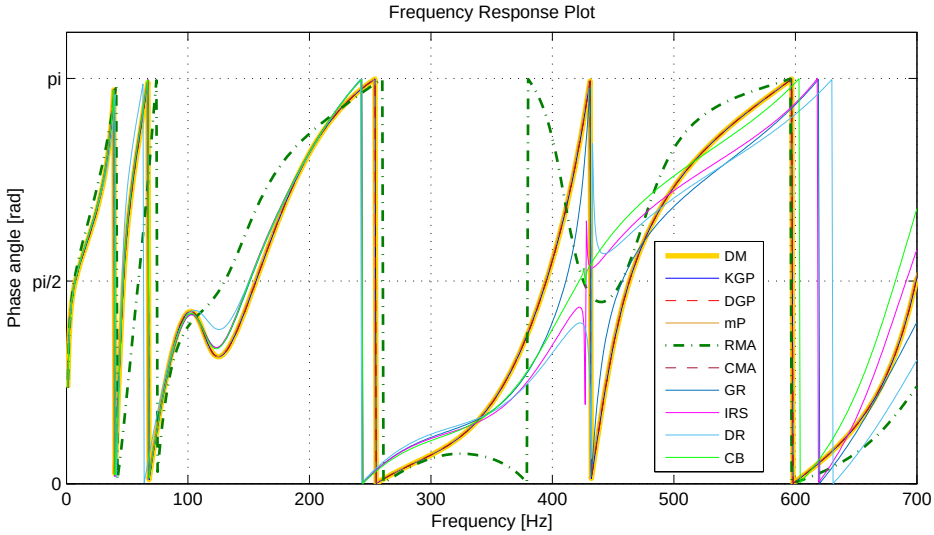


Figure 10: Phase angle of the response for u_{11} of the arbitrary damped cantilever beam

6.2 Solar Panel - Structure

The second numerical example is a solar panel modeling with ANSYS®, see fig. 11. It is a $4\ m \times 12\ m$ structure, built out of beam and shell elements. The beam188 element was used for the frame, with an elastic modulus of $2 \times 10^{11}\ \frac{N}{m^2}$, a Poisson's ratio of 0.3 and mass density of $7850\ \frac{kg}{m^3}$. For the panel the shell181 element was used, with an elastic modulus of $0.7 \times 10^{11}\ \frac{N}{m^2}$, a Poisson's ratio of 0.3 and mass density of $2500\ \frac{kg}{m^3}$. This structure was loaded with a distributed load of $50N$ in vertical direction (Y direction of the global coordinate system). In total, the solar panel contains 9582 dofs, 3 nodes are fixed, i.e. 18 dofs, and has 33 coupling dofs, thus the equation system to be solved contains 9531 unknowns (displacements and rotations).

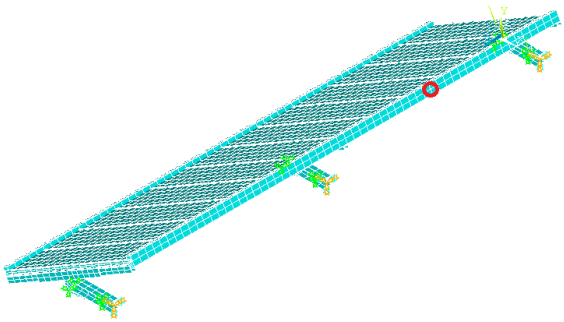


Figure 11: Solar panel modeling in ANSYS®.

For medium- and large-scale structures without going into the sub structuring approach (CB method) the RMA is extensively used in frequency response analysis. The computation of the set of mode shapes is the time-demanding task in the RMA but leads to a reduced-decouple equation system. RMA can be also applied over a large frequency interval $\Delta\omega$, unlike condensation methods which offer good approximation only in a narrow $\Delta\omega$ near to $\omega = 0$ or to the chosen dynamic frequency ω_{dy} in case of DR. For those reasons the computational time of the

Interpolarory MOR are compared to that of the RMA and the DM solving a non-damped solar panel - structure.

6.2.1 Interpolatory MOR Methods and RMA

For this first analysis $\Delta_\omega = [1, 100]$, $n_\omega = 100$, $P = 16$, and ω_p={ 5, 10, 15, 20, 25, 30, 40, 50, 55, 60, 65, 70, 85, 90, 95, 100 }. mP uses $P = 10$ and ω_p={1, 15, 30, 40, 50, 60, 75, 85, 95, 100}. RMA number of modes is equal to 100.

The computation time required for this analysis is shown in tab. 1 and the amplitude of the frequency response for dof UY of the node 153, marked with a circle over the upper beam in fig. 11, is depicted in fig. 12.

$Method$	$c.t.[s]$	$c.t.parallel[s]$
DM	10.330	5.231
mP	3.650	2.153
KGP	3.840	2.560
DGP	3.900	2.249
RMA	1.820	2.316

Table 1: Computation time using DM, mP, KGP, DGP and RMA.

The comparison in computation time (c.t.) shows a faster solution for the parallel implementation of Interpolatory MOR Methods, i.e. the KGP and DGP are 4.5 times faster than the DM runs in the conventional way. The RMA provides also a fast computation for this problem and overlaps the exact solution as KGP and DGP. The mP computation time in tab. 12 is just for the computation of the response of a single dof and fig. 12 shows that the mP-response is slightly shifted from the exact one. For the mP approximation, it is recommended to used a mesh with few interpolation points, e.g. for this example $P < 10$, due to the fact that the matrix $\mathbf{A}$ in eq. (76) is ill-conditioned for large P [9], which yields to wrong coefficients and therefore wrong approximation of the response.

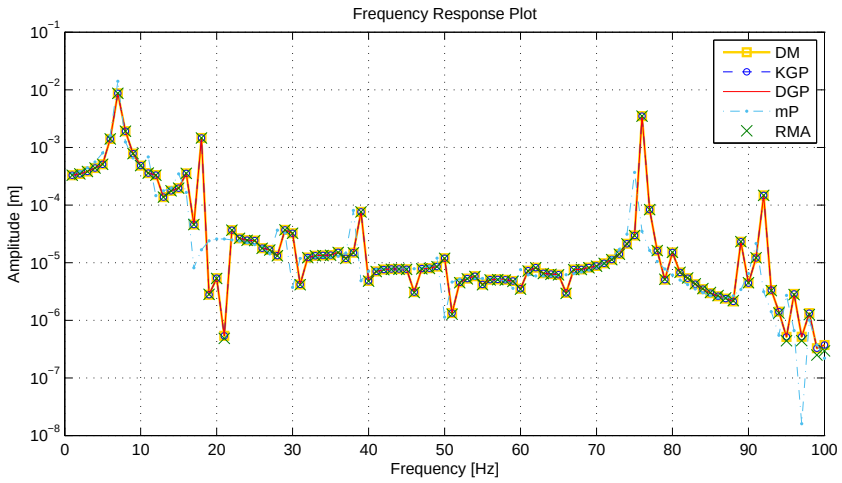


Figure 12: Amplitude of the response using DM, KGP, DGP and RMA.

6.2.2 Interpolatory MOR Methods

In this second investigation $\Delta_\omega = [1, 1000]$, $n_\omega = 3000$, $P = 101$, ω_p={5, 10, 20, 30, 40, 50, ... , 500, 510, 520,...,1000} and RMA number of modes is equal to 500.

The computation time required for this analysis is written in tab. 2 showing that the KGP and DGP (run in parallel) are approximately 18 times faster than the DM and 6 times faster than the DM runs in parallel. Also the KGP and DGP are 2 time faster than the RMA when they are run in parallel.

$Method$	$c.t.[s]$	$c.t.parallel[s]$
DM	274.732	91.884
KGP	42.267	15.522
DGP	39.731	15.507
RMA	38.236	31.439

Table 2: Computation time using DM, mP, KGP, DGP and RMA.

The amplitude of the frequency response for dof UY of the node 153 , marked with a circle over the upper beam in fig. 11, is depicted in figs. 13, 14, 15 and 16.

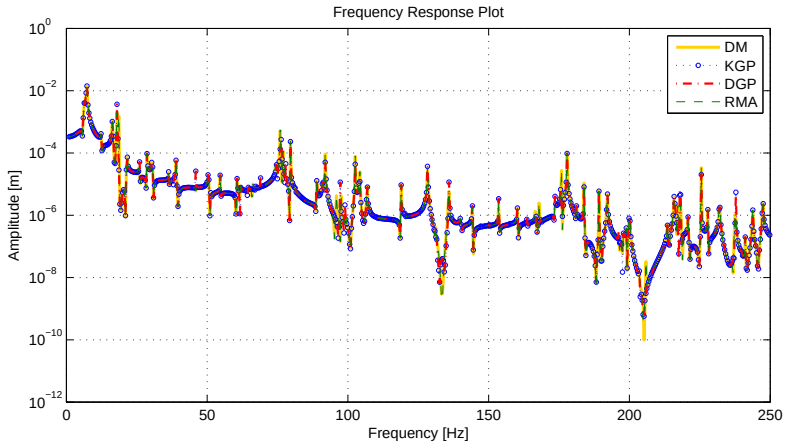


Figure 13: Amplitude of the response using DM, KGP, DGP and RMA: $\Delta_\omega = [1, 250]$.

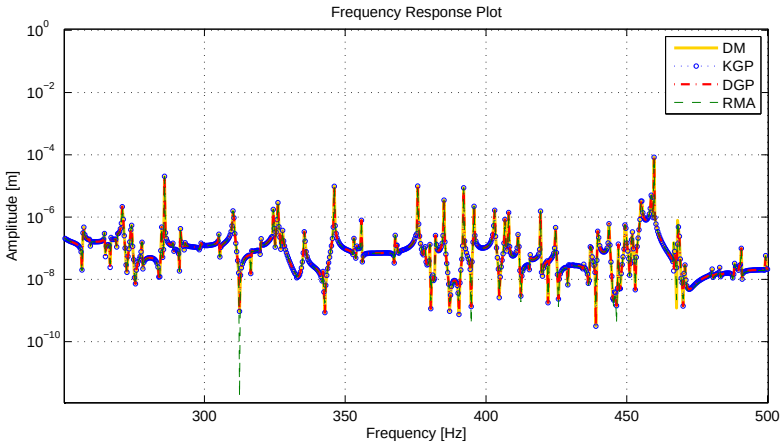


Figure 14: Amplitude of the response using DM, KGP, DGP and RMA: $\Delta_\omega = [250, 500]$.

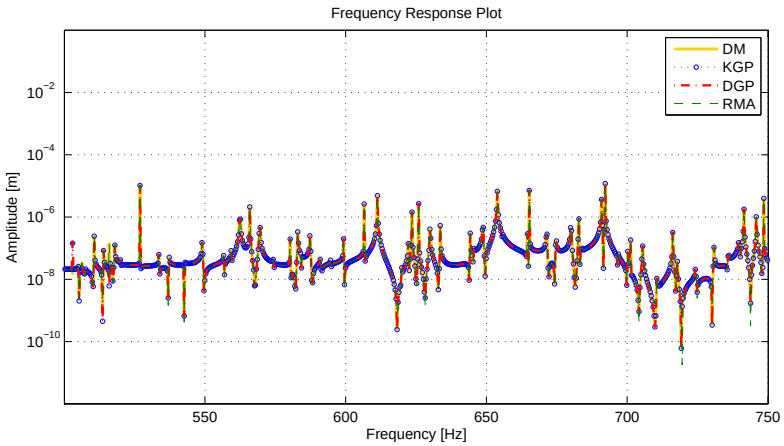


Figure 15: Amplitude of the response using DM, KGP, DGP and RMA: $\Delta_\omega = [500, 750]$.

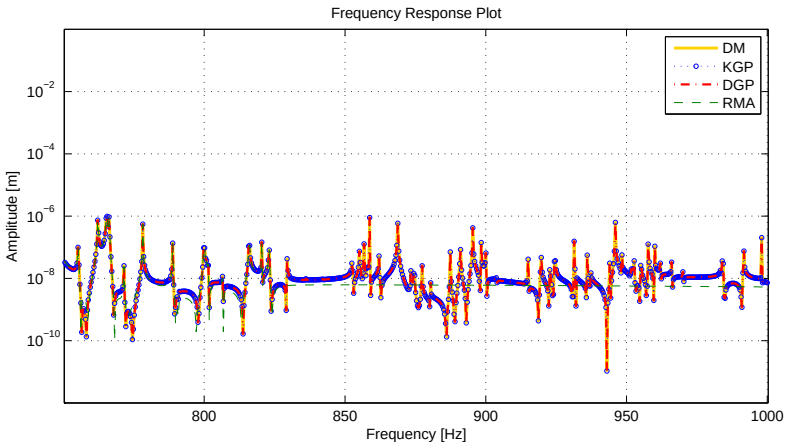


Figure 16: Amplitude of the response using DM, KGP, DGP and RMA: $\Delta_\omega = [750, 1000]$.

The solutions given by the KGP and DGP overlap the exact solution almost everywhere in

the Δ_ω and the RMA using 500 modes is able to capture the exact solution until the 865Hz, see Fig. 16, using almost the same amount of time.

7 CONCLUSIONS

As shown in this work several approaches are available to have a reduced-order model for a second order dynamical system of the form given by eq. (2). Some of them have been extensively used in the structural dynamics fields when $\Delta\omega$ is a narrow band and the damping matrix is proportional to the stiffness and mass matrix. Those are: the GR, IRS and DR. Another classical technique like RMA is suitable when the eigen-frequencies of the modal matrix are the main eigen-frequencies in $\Delta\omega$ which does not limit the interval $\Delta\omega$. In a similar way, the Interpolatory MOR are applicable in a large $\Delta\omega$ with all the interpolatory points contained in that interval.

For non proportional-damped systems the CMA and the Interpolatory MOR are the best choices, but the CMA increases the dimension of the matrices. The mP is applicable in a narrow $\Delta\omega$ or when few interpolation points are needed. The KGP and DGP can be applied to any kind of system defined by eq. (2) over a narrow or large $\Delta\omega$. Thus they are very promising methods to obtain a reduced-order model with cheaper computational cost and good approximation to the exact solution for proportional and non-proportional damped systems.

REFERENCES

[1] Zeeman, E. C. (1993). Controversy in science: on the ideas of Daniel Bernoulli and Ren Thom. *Nieuw Arch Wisk*, 11, 257.

[2] Guyan, R.J. (1965) Reduction of stiffness and mass matrices. *AIAA journal* 3.2, 380-380. .

[3] Irons, B. (1965). Structural eigenvalue problems-elimination of unwanted variables. *AIAA journal*, 3(5), 961-962.

[4] Leung, A. Y. T. (1978). An accurate method of dynamic condensation in structural analysis. *International Journal for Numerical Methods in Engineering*, 12(11), 1705-1715.

[5] Friswell, M. I., Garvey, S. D., & Penny, J. E. T. (1995). Model reduction using dynamic and iterated IRS techniques. *Journal of sound and vibration*, 186(2), 311-323.

[6] Bampton, M. C., & Craig, Jr, R. R. (1968). Coupling of substructures for dynamic analyses. *Aiaa Journal*, 6(7), 1313-1319.

[7] Su, T. J., & Craig, R. R. (1991). Model reduction and control of flexible structures using Krylov vectors. *Journal of guidance, control, and dynamics*, 14(2), 260-267.

[8] Young, J. T., & Haile, W. B. (2000). Primer on the Craig-Bampton method. *Finite Element Modeling Continuous Improvement*, NASA.

[9] Bai, Z. (2002). Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Applied Numerical Mathematics*, 43(1), 9-44.

[10] Freund, R. W. (2003). Model reduction methods based on Krylov subspaces. *Acta Numerica* , 12, 267-319.

[11] Beattie, C., & Gugercin, S. (2009). Interpolatory projection methods for structure-preserving model reduction. *Systems & Control Letters*, 58(3), 225-232.

[12] Salimbahrami, S. B. (2005). *Structure preserving order reduction of large scale second order models* (Doctoral dissertation, Universität München).

[13] Avery, P., Farhat, C., & Reese, G. (2007). Fast frequency sweep computations using a multipoint Padbased reconstruction method and an efficient iterative solver. *International Journal for Numerical Methods in Engineering*, 69(13), 2848-2875.

[14] Hetmaniuk, U., Tezaur, R., & Farhat, C. (2012). Review and assessment of interpolatory model order reduction methods for frequency response structural dynamics and acoustics problems. *International Journal for Numerical Methods in Engineering*, 90(13), 1636-1662.

[15] Craig, R. R., & Kurdila, A. J. (2006). *Fundamentals of structural dynamics*. John Wiley & Sons.

[16] Meirovitch, L. (2010). *Fundamentals of vibrations*. Waveland Press.

[17] O'Callahan, J. C. (1989, February). A procedure for an improved reduced system (IRS) model. *In Proceedings of the 7th international modal analysis conference* (Vol. 1, pp. 17-21). Union College Press, Schenectady, NY.

[18] Gordis, J. H. (1992). An analysis of the improved reduced system (IRS) model reduction procedure. In *10th International Modal Analysis Conference* (Vol. 1, pp. 471-479).

[19] Blair, M. A., Camino, T. S., & Dickens, J. M. (1991). An iterative approach to a reduced mass matrix. In *9th Conference International Modal Analysis Conference (IMAC)* (Vol. 1, pp. 621-626).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

TOWARDS A NON-INTRUSIVE PROPER GENERALIZED DECOMPOSITION SCHEME FOR MODEL ORDER REDUCTION

X. Zou[1,2], P. Díez[2], M. Conti[1], and F. Auricchio[1]

[1]Università degli Studi di Pavia
Via Ferrata 3, 27100 Pavia, Italy
e-mail: {xi.zou, michele.conti, auricchio}@unipv.it

[2] Universitat Politècnica de Catalunya
Jordi Girona 1-3 E-08034 Barcelona, Spain
e-mail: pedro.diez@upc.edu

Keywords: Proper Generalized Decomposition, Non-intrusive, Model Order Reduction.

Abstract. *Proper Generalized Decomposition (PGD) has become a popular technique of model order reduction to solve the so-called curse of dimensionality in multidimensional partial differential equations, especially for real-time simulations. Therefore, it is very interesting to implement the PGD algorithm non-intrusively for a better integration with off-the-shelf finite element (FE) codes. This paper provides a non-intrusive scheme for PGD computation based on commercial softwares to minimize the implementation effort. Specifically, also for the sake of simplicity, we chose linear elasticity models with two independent material parameters as the extra coordinates for the preliminary implementations. The strategy is using proper separated representation of the functional of material properties to split the problem into sequentially independent linear equations, and solve the equations of mechanical problem, which usually have a large scale, with an efficient external solver. The proposed non-intrusive PGD scheme can be easily realized using arbitrary codes for iteration control cooperating with commercial FE softwares. Numerical examples were realized with MATLAB (The MathWorks, Inc., USA) as the main procedure for iterations, and Abaqus (Dassault Systèmes, France) as the external solver. In particular, we have applied the algorithm on a biomedical problem. The proposed scheme is extensible and also considered to be a promising bridge between academic research and industrial application.*

1 INTRODUCTION

Modern simulation-based engineering sciences relies on physical models depending on many parameters described by partial differential equations (PDEs) to be solved in a multidimensional parametric space, where the depending parameters are viewed as extra coordinates. For this reason, the dimensionality of the problem is increased and the computational cost is raised exponentially, leading to the so-called *curse of dimensionality*.

Proper generalized decomposition (PGD) method [1, 2, 3, 4] is a recently developed model order reduction technique particularly powerful on tackling the curse of dimensionality. Readers may refer to [5] for a recent review on PGD method. As a model order reduction method, it can be divided into off-line phase and on-line phase. In this paper, we will focus on the off-line phase of PGD method, which produces a computational vademecum [6] consisting of functions of the parameters. Based on the vademecum, the value of the functions can be calculated as fast as real-time in the on-line phase, so that to enable real-time and multi-query simulation.

Numerical methods for solving PDEs, especially the finite element method [7], are intensely implemented in both academia and industry in the form of computer aided engineering (CAE), covering a wide range of fields. There are numerous finite element (FE) codes highlighting different features. Academical institutes tend to develop less complexed but more flexible codes for research purposes. Industrial companies invest resources on commercial CAE softwares and make rules for the application routines, because the provided reliable results with robust FE solvers, and user-friendly graphical interfaces for the pre-processing and post-processing can improve productivity. To bridge the gap in between, we propose a scheme for the application of PGD method in a non-intrusive manner. The idea is using a general-purposed code to perform input/output (I/O) and to control the whole program, in which the key subroutines are executed by calling the external solver. The non-intrusive implementation is guaranteed for PGD method, thanks to the feasibility of separated representation for functions of parameters. Recent work by Courard *et al.* [8] has implied the non-intrusive PGD idea that starts from taking geometric parameters as extra coordinates, and makes SAMCEF (Siemens, Germany) work with an in-house MATLAB code for the generation of virtual charts to optimize engineering design process.

In this work we concentrate on physical models of linear elasticity, whose governing equation is vector-valued. The problems considered include a very simple but typical one-dimensional model, and a practical three-dimensional model as an application in biomechanics. As a preliminary step towards a complete non-intrusive PGD scheme, we take the material properties as the extra coordinates. Further work is in progress for solving problems with load or boundary conditions or geometric parameters as the extra coordinates based on the proposed scheme.

2 PROBLEM STATEMENT

2.1 Governing equation in linear elasticity with two materials

Under the infinitesimal deformation assumption, the relationship between strain tensor $\boldsymbol{\varepsilon}$ and displacement $\mathbf{u}$ reads

$$\boldsymbol{\varepsilon} = \nabla_s \mathbf{u} = \frac{1}{2}\left[\nabla \mathbf{u} + (\nabla \mathbf{u})^T\right]. \tag{1}$$

The stress tensor $\boldsymbol{\sigma}$ and the strain tensor are related with elasticity tensor $\mathbb{C}$ by Hook's law

$$\boldsymbol{\sigma} = \mathbb{C} : \boldsymbol{\varepsilon}. \tag{2}$$

Consider an elastic domain $\Omega \subset \mathbb{R}^d, (d = 1, 2, 3)$, the function of interest is displacement $\mathbf{u}$, of the following boundary value problem (BVP) of the govering equations. The strong form

of the BVP is stated as follows: find $\mathbf{u}$ satisfying the equilibrium equation and the boundary conditions

$$\nabla \cdot \boldsymbol{\sigma} + \mathbf{b} = 0 \quad \text{in } \Omega, \tag{3}$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \Gamma_D, \tag{4}$$

$$\boldsymbol{\sigma} \cdot \mathbf{n} = \mathbf{t}_N \quad \text{on } \Gamma_N, \tag{5}$$

where $\mathbf{b}$ represents the known external body forces. The Dirichlet boundary $\mathbf{u}_D$ and Neumann boundary traction $\mathbf{t}_N$ are also known data sets.

To simulate an elastic body composed of two, for example, parts with different material properties, the domain is partitioned into two subdomains Ω_1 and Ω_2. For the sake of simplicity, we consider only the homogeneous, isotropic materials with the same Poisson's ratio, so that the elasticity tensor $\mathbb{C}$ depends only on Young's Modulus E within each subdomain. So that the Young's Modulus E can be written as

$$E(\mathbf{x}) = \begin{cases} E_1 & \text{for } \mathbf{x} \in \Omega_1, \\ E_2 & \text{for } \mathbf{x} \in \Omega_2. \end{cases} \tag{6}$$

Geometrically, we suppose that $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2$ and $\partial\Omega_1 \cap \partial\Omega_2$ is the interface between the two open subdomains. In principle, there is no difficulty to increase the number of subdomains with the same strategy.

The standard weak form of the problem reads as follows: find $\mathbf{u} \in V$ such that

$$a(\mathbf{u}, \mathbf{v}) = l(\mathbf{v}), \quad \forall \mathbf{v} \in V_0, \tag{7}$$

where $V := \{\mathbf{u} \in H^1(\Omega) : \mathbf{u} = \mathbf{u}_D \text{ on } \Gamma_D\}$, and its corresponding test function space is $V_0 := \{\mathbf{u} \in H^1(\Omega) : \mathbf{u} = 0 \text{ on } \Gamma_D\}$. The bilinear and linear forms $a(\cdot,\cdot)$ and $l(\cdot)$ are given by

$$a(\mathbf{u}, \mathbf{v}) := \int_\Omega \nabla_s \mathbf{u} : \mathbb{C} : \nabla_s \mathbf{v} \, d\Omega \quad \text{and} \quad l(\mathbf{v}) := \int_\Omega \mathbf{b} \cdot \mathbf{v} \, d\Omega + \int_{\Gamma_N} \mathbf{t}_N \cdot \mathbf{v} \, d\Gamma. \tag{8}$$

2.2 PGD separated representation with material properties as extra coordinates

The general procedure of the parametrization problem is discussed in [9]. Here we introduce the PGD weak form simply by assuming the material properties as extra coordinates. The displacement is then generalized as $\mathbf{u} = \mathbf{u}(\mathbf{x}, E_1, E_2)$, in which $\mathbf{x} \in \Omega$, $E_1 \in I_{E_1}$ and $E_2 \in I_{E_2}$, and thus the weak form reads: find $\mathbf{u} \in V \otimes L^2(I_{E_1}) \otimes L^2(I_{E_2})$, such that

$$\mathcal{A}(\mathbf{u}, \mathbf{v}) = \mathcal{L}(\mathbf{v}), \quad \forall \mathbf{v} \in V_0 \otimes L^2(I_{E_1}) \otimes L^2(I_{E_2}), \tag{9}$$

where the bilinear and linear forms are defined by

$$\mathcal{A}(\mathbf{u}, \mathbf{v}) := \int_{I_{E_1}} \int_{I_{E_2}} a(\mathbf{u}, \mathbf{v}) \, dE_2 dE_1 \quad \text{and} \quad \mathcal{L}(\mathbf{v}) := \int_{I_{E_1}} \int_{I_{E_2}} l(\mathbf{v}) \, dE_2 dE_1. \tag{10}$$

It can be seen that the dimensionality of the problem is increased by the introduced extra coordinates, which may cause the so-called *curse of dimensionality*. To overcome this with PGD methodology, the separated representation is used to approximate $\mathbf{u}(\mathbf{x}, E_1, E_2)$ as

$$\mathbf{u}(\mathbf{x}, E_1, E_2) \approx \mathbf{u}^n_{\text{PGD}}(\mathbf{x}, E_1, E_2) = \sum_{m=1}^{n} \boldsymbol{\chi}^m(\mathbf{x}) \varepsilon_1^m(E_1) \varepsilon_2^m(E_2). \tag{11}$$

This representation implies that we can approximate the solution by using only n terms of modes, which are products of the separated functions $\boldsymbol{\chi}(\mathbf{x})$, $\varepsilon_1(E_1)$ and $\varepsilon_2(E_2)$. Hereafter, the explicit dependence of these functions is frequently omitted in the notations since readers can easily identify from the names and the subscripts. The normalization of the products is useful by defining the magnitude of the modes

$$M^m := \|\boldsymbol{\chi}^m\| \cdot \|\varepsilon_1^m\| \cdot \|\varepsilon_2^m\|. \tag{12}$$

Derivation of the PGD algorithm requires assuming that the bilinear and linear forms $a(\mathbf{u}, \mathbf{v})$ and $l(\mathbf{v})$ are also separable. That is to say, the functions involved should be separable. For our purposes, as generally seen in real structural problems, we assume that the load is independent of the material properties; while the elasticity tensor $\mathbb{C}$ is dependent on the Young's modulus of the subdomains in a generalized form of Equation (6)

$$\mathbb{C}(\mathbf{x}, E_1, E_2) = \sum_{k=1}^{n_{\mathbb{C}}} \mathbb{C}_k(\mathbf{x}) \mu_1^k(E_1) \mu_2^k(E_2), \tag{13}$$

where $\mathbb{C}_k(\mathbf{x})$, $\mu_1^k(E_1)$ and $\mu_2^k(E_2)$ describe the material properties in each of the $n_{\mathbb{C}}$ subdomains. In the simplest two-material case, without losing generality, the Heaviside-like function $H_k(\mathbf{x}) : \Omega \to \{0, 1\}$ is used as

$$\mathbb{C}(\mathbf{x}, E_1, E_2) = \mathbb{C}_1 H_1(\mathbf{x}) E_1 + \mathbb{C}_2 H_2(\mathbf{x}) E_2, \tag{14}$$

in which $\mathbb{C}_i$ is a constant tensor and

$$H_i(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \Omega_i, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

The function related to Young's modulus reads

$$\mu_i^j(E_i) = \begin{cases} E_i, & i = j, \\ 1, & i \neq j. \end{cases} \tag{16}$$

In practical FE analysis, following the standard approach, $\mathbb{C}$ is represented in a discretized form by the stiffness matrix $\mathbf{K}$, which is computed by FE softwares as long as the user inputs given values for E_1 and E_2 and specifies the mapping of material properties to the subdomains (or element sets) in the pre-processing.

2.3 Non-intrusive PGD mode search with alternated directions

In general, the number of modes n which needed to approximate $\mathbf{u}_{\text{PGD}}^n$ is not known a priori. The greedy approach is used to calculate the modes on-the-fly with proper stopping criteria. Note that

$$\mathbf{u}_{\text{PGD}}^n(\mathbf{x}, E_1, E_2) = \mathbf{u}_{\text{PGD}}^{n-1}(\mathbf{x}, E_1, E_2) + \boldsymbol{\chi}^n \varepsilon_1^n \varepsilon_2^n, \tag{17}$$

we can compute the modes term by term. From now on the superscript n for the functions is omitted since we are more focused on the terms. Due to the bilinearity, we have turned the problem into: find $\boldsymbol{\chi} \in V$, $\varepsilon_1 \in L^2(I_{E_1})$ and $\varepsilon_2 \in L^2(I_{E_2})$ such that

$$\mathcal{A}(\boldsymbol{\chi}\varepsilon_1\varepsilon_2, \mathbf{v}) = \mathcal{L}(\mathbf{v}) - \mathcal{A}(\mathbf{u}_{\text{PGD}}^{n-1}, \mathbf{v}), \quad \forall \mathbf{v} \in V_0 \otimes L^2(I_{E_1}) \otimes L^2(I_{E_2}). \tag{18}$$

To apply FE approximation properly, the fixed-point iteration method is used by treating the terms of test function $\mathbf{v}$ separately from normal variation $\mathbf{v} = \mathbf{u}^* = \boldsymbol{\chi}^*\varepsilon_1\varepsilon_2 + \boldsymbol{\chi}\varepsilon_1^*\varepsilon_2 + \boldsymbol{\chi}\varepsilon_1\varepsilon_2^*$. With this idea, the alternating directions approach is adopted to split (18) into three sequential linear problems:

1. Assuming ε_1 and ε_2 known, update $\boldsymbol{\chi}$ such that

$$\mathcal{A}(\boldsymbol{\chi}\varepsilon_1\varepsilon_2, \boldsymbol{\chi}^*\varepsilon_1\varepsilon_2) = \mathcal{L}(\boldsymbol{\chi}^*\varepsilon_1\varepsilon_2) - \mathcal{A}(\mathbf{u}_{\text{PGD}}^{n-1}, \boldsymbol{\chi}^*\varepsilon_1\varepsilon_2), \quad \forall \boldsymbol{\chi}^* \in V_0. \tag{19}$$

2. Assuming $\boldsymbol{\chi}$ and ε_2 known, update ε_1 such that

$$\mathcal{A}(\boldsymbol{\chi}\varepsilon_1\varepsilon_2, \boldsymbol{\chi}\varepsilon_1^*\varepsilon_2) = \mathcal{L}(\boldsymbol{\chi}\varepsilon_1^*\varepsilon_2) - \mathcal{A}(\mathbf{u}_{\text{PGD}}^{n-1}, \boldsymbol{\chi}\varepsilon_1^*\varepsilon_2), \quad \forall \varepsilon_1^* \in L^2(I_{E_1}). \tag{20}$$

3. Assuming $\boldsymbol{\chi}$ and ε_1 known, update ε_2 such that

$$\mathcal{A}(\boldsymbol{\chi}\varepsilon_1\varepsilon_2, \boldsymbol{\chi}\varepsilon_1\varepsilon_2^*) = \mathcal{L}(\boldsymbol{\chi}\varepsilon_1\varepsilon_2^*) - \mathcal{A}(\mathbf{u}_{\text{PGD}}^{n-1}, \boldsymbol{\chi}\varepsilon_1\varepsilon_2^*), \quad \forall \varepsilon_2^* \in L^2(I_{E_2}). \tag{21}$$

We would like to address that (19) is a mechanical problem, since $\boldsymbol{\chi}$ is a vector-valued function, that usually takes much more cost to compute, while (20) and (21) are one-dimensional problems, since ε_1 and ε_2 are scalar functions, which can be solved with little computational cost. The key idea of the non-intrusive scheme we propose is making use of off-the-shelf FE softwares to solve the mechanical problem numerically with the expectation of efficiency improvement for both coding and computing.

2.4 FE approximation and resulting linear algebra formulation

Choosing proper subspaces $V^h \subset V$, $V_0^h \subset V_0$, $V_1^h \subset L^2(I_{E_1})$ and $V_2^h \subset L^2(I_{E_2})$, we perform the Galerkin approximation by choosing proper discrete approximations following or imitating the standard FE formulations. That is

$$\boldsymbol{\chi} \approx \mathbf{N}^T\mathbf{U}, \quad \varepsilon_1 \approx \mathbf{N}_1^T\boldsymbol{\varepsilon}_1, \quad \varepsilon_2 \approx \mathbf{N}_2^T\boldsymbol{\varepsilon}_2, \tag{22}$$

and thus

$$\mathbf{u}_{\text{PGD}}^n \approx \sum_{m=1}^{n} \mathbf{N}^T\mathbf{U}^m \cdot \mathbf{N}_1^T\boldsymbol{\varepsilon}_1^m \cdot \mathbf{N}_2^T\boldsymbol{\varepsilon}_2^m, \tag{23}$$

where $\mathbf{N}$ is the matrix of shape functions of the standard FE approach for the mechanical problem, $\mathbf{N}_1$ and $\mathbf{N}_2$ are vectors of shape functions for one-dimensional problems. Note that the shape functions are fixed, only the degree of freedom (DOF) vectors $\mathbf{U}$, $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_1$ will be updated throughout the computation. Empirically, to initialize the computation, we assume $\mathbf{U}$ as a zero vector while $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are vectors in which each component equals to one. The dimensions of the DOF vectors can be chosen according to the desired precision.

In standard FE theory, the global stiffness matrix $\mathbf{K}$ is an assembly of elemental stiffness matrices, so it has a natural separable formulation. According to (13), the global stiffness matrix $\mathbf{K}$ can be rewritten as

$$\mathbf{K} = E_1\mathbf{K}_1 + E_2\mathbf{K}_2, \tag{24}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are stiffness-like matrices. It is convenient that we still refer $\mathbf{K}_1$ and $\mathbf{K}_2$ to stiffness matrices without any ambiguity.

With the help of external FE software, we can obtain $\mathbf{K}_1$ and $\mathbf{K}_2$ in the non-intrusive manner. In particular, using Equation (24), $\mathbf{K}_1$ can be obtained by letting $E_1 = 1$ and $E_2 = 0$, while $\mathbf{K}_2$ be obtained by $E_1 = 0$ and $E_2 = 1$.

By putting (22) to (24), along with (8), (10), (11) and (13) into (19) to (21), we can obtain the algebraic formulation as

1. Assuming $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ known, update $\mathbf{U}$ by solving

$$(M_1 C_2 \mathbf{K}_1 + M_2 C_1 \mathbf{K}_2)\mathbf{U} = Q_1 Q_2 \mathbf{F} - \sum_{k=1}^{n-1} (M_1^k C_2^k \mathbf{K}_1 + M_2^k C_1^k \mathbf{K}_2)\mathbf{U}^k. \tag{25}$$

   Letting $\tilde{\mathbf{K}} = M_1 C_2 \mathbf{K}_1 + M_2 C_1 \mathbf{K}_2$ and $\tilde{\mathbf{F}} = Q_1 Q_2 \mathbf{F} - \sum_{k=1}^{n-1} (M_1^k C_2^k \mathbf{K}_1 + M_2^k C_1^k \mathbf{K}_2)\mathbf{U}^k$, Equation (25) is simplified to $\tilde{\mathbf{K}}\mathbf{U} = \tilde{\mathbf{F}}$, which has a standard FE form, thus can be seen as a fictitious mechanical problem and be sent to the external FE software for a non-intrusive solution, taking advantage of its robust solvers.

2. Assuming $\mathbf{U}$ and $\boldsymbol{\varepsilon}_2$ known, update $\boldsymbol{\varepsilon}_1$ by solving

$$(K_1 C_2 \mathbf{M}_1 + M_2 K_2 \mathbf{C}_1)\boldsymbol{\varepsilon}_1 = F Q_2 \mathbf{Q}_1 - \sum_{k=1}^{n-1} (K_1^k C_2^k \mathbf{M}_1 + K_2^k M_2^k \mathbf{C}_1)\boldsymbol{\varepsilon}_1^k. \tag{26}$$

3. Assuming $\mathbf{U}$ and $\boldsymbol{\varepsilon}_1$ known, update $\boldsymbol{\varepsilon}_2$ by solving

$$(M_1 K_1 \mathbf{C}_2 + K_2 C_1 \mathbf{M}_2)\boldsymbol{\varepsilon}_2 = F Q_1 \mathbf{Q}_2 - \sum_{k=1}^{n-1} (K_1^k M_1^k \mathbf{C}_2 + K_2^k C_1^k \mathbf{M}_2)\boldsymbol{\varepsilon}_2^k. \tag{27}$$

The coefficients and matrices are defined as follows

$$\mathbf{K}_i := \int_{\Omega_i} \nabla_s \mathbf{N} : \mathbb{C}_i : \nabla_s \mathbf{N}^T \, d\Omega, \quad K_i := \mathbf{U}^T \mathbf{K}_i \mathbf{U}, \quad K_i^k := \mathbf{U}^T \mathbf{K}_i \mathbf{U}^k, \quad i = 1, 2. \tag{28}$$

$$\mathbf{M}_i := \int_{I_{E_i}} E_i \mathbf{N}_i \mathbf{N}_i^T \, dE_i, \quad M_i := \boldsymbol{\varepsilon}_i^T \mathbf{M}_i \boldsymbol{\varepsilon}_i, \quad M_i^k := \boldsymbol{\varepsilon}_i^T \mathbf{M}_i \boldsymbol{\varepsilon}_i^k, \quad i = 1, 2. \tag{29}$$

$$\mathbf{C}_i := \int_{I_{E_i}} \mathbf{N}_i \mathbf{N}_i^T \, dE_i, \quad C_i := \boldsymbol{\varepsilon}_i^T \mathbf{C}_i \boldsymbol{\varepsilon}_i, \quad C_i^k := \boldsymbol{\varepsilon}_i^T \mathbf{C}_i \boldsymbol{\varepsilon}_i^k, \quad i = 1, 2. \tag{30}$$

$$\mathbf{Q}_i := \int_{I_{E_i}} \mathbf{N}_i^T \, dE_i, \quad Q_i := \mathbf{Q}_i^T \boldsymbol{\varepsilon}_i, \quad i = 1, 2. \tag{31}$$

$$\mathbf{F} := \int_{\Omega} \mathbf{N}^T \mathbf{b} \, d\Omega + \int_{\Gamma_N} \mathbf{N}^T \mathbf{t}_N \, d\Gamma, \quad F := \mathbf{U}^T \mathbf{F}. \tag{32}$$

In the definitions above, the load vector $\mathbf{F}$ is exactly the same as that in standard FE theory, since we have the assumption that it is independent of the material properties. Therefore, $\mathbf{F}$ can be obtained directly from the external FE software. As previously stated, $\mathbf{K}_i$ can be obtained non-intrusively by giving proper values of E_1 and E_2. By definition, $\mathbf{M}_i$ and $\mathbf{C}_i$ are quasi-diagonal sparse matrices, and $\mathbf{Q}_i$ is a vector, all of them can be calculated in conventional ways.

2.5 Non-intrusive PGD algorithm

We take advantage of MATLAB as the main code, and Abaqus as the external FE solver. See the flowchart of non-intrusive PGD algorithm in Figure 1.

The pre-processing of the mechanical model can be performed in Abaqus/CAE, which will produce an input file containing data of the node-element topology, the load and boundary conditions, and the material properties with associated element sets. In our non-intrusive scheme, stiffness matrix $\mathbf{K}_i$ and load vector $\mathbf{F}$ are generated within Abaqus. In addition, the Python interface provided by Abaqus is also used for data I/O.

The matrices $\mathbf{M}_i$, $\mathbf{C}_i$ and vector $\mathbf{Q}_i$ are all calculated and stored in MATLAB once for all. While the scalar parameters K_i, M_i, C_i, Q_i are updated and stored with mode index k in each iteration.

The idea that using the external FE solver to resolve the mechanical problem (25) can also be generalized to the parametric problems (26) and (27). However, the parametric problems are one-dimensional, which can be solved very quickly within the main code, so that this generalization is usually unnecessary.

Two error limits, η_1 and η_2, are introduced within the stopping criteria of the iteration loops: η_1 controls the iteration of alternative direction, η_2 controls the search for PGD modes. In the following examples, we have empirically chosen $\eta_1 = 10^{-5}$ and $\eta_2 = 1$.

In standard FE analysis in mechanics, one is usually more interested in the global nodal result $\mathbf{U}$ other than the field result $\mathbf{u}(\mathbf{x})$. To be compatible, we can rewrite (23) as

$$\mathbf{u}^n_{\text{PGD}}(\mathbf{x}, E_1, E_2) \approx \mathbf{N}^T(\mathbf{x})\mathbf{U}_{\text{PGD}}(E_1, E_2), \tag{33}$$

by introducing

$$\mathbf{U}_{\text{PGD}} := \sum_{m=1}^{n} \mathbf{U}^m \cdot \mathbf{N}_1^T \boldsymbol{\varepsilon}_1^m \cdot \mathbf{N}_2^T \boldsymbol{\varepsilon}_2^m. \tag{34}$$

So that we can use $\mathbf{U}_{\text{PGD}}$ to make comparisons with standard FE result $\mathbf{U}_{\text{FE}}$.

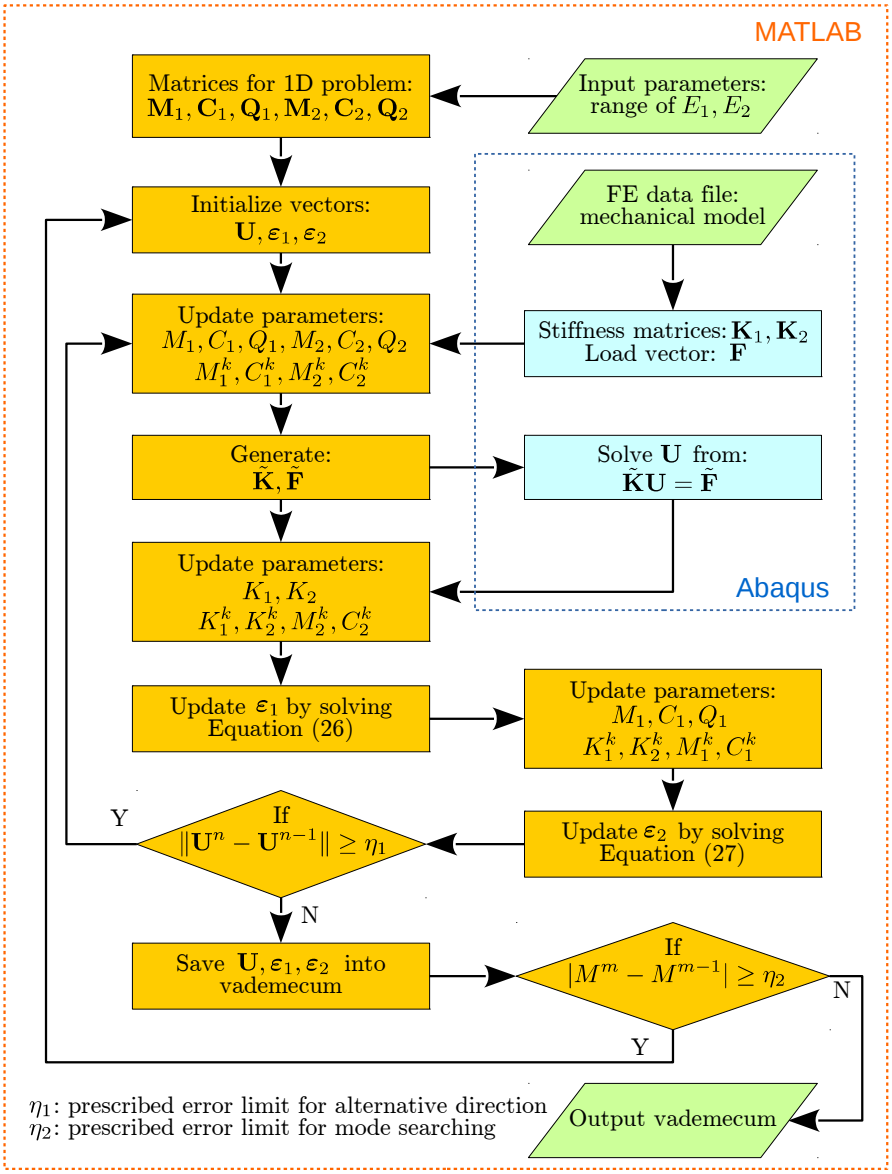


Figure 1: Flowchart of non-intrusive PGD algorithm exemplified by MATLAB as the main program and Abaqus as the external solver.

3 NUMERICAL EXAMPLES

3.1 Example I: a one-dimensional model problem

Consider a one-dimensional mechanical model problem consisting of two vertical bars, shown in Figure 2, with unit cross-section area $A_1 = A_2 = 1$ mm^2 and with different lengths $L_1 = 10$ mm and $L_2 = 20$ mm and Young's moduli $E_1 \in [10, 100]$ MPa and $E_2 \in [20, 200]$ MPa. The lower end (Node 3) is fixed ($u_3 = 0$) and the upper end (Node 1) loaded with a concentrated force $F = -10$ N. From standard FE theory, we know that the load vector is $\mathbf{F} = F[1, 0, 0]^T$ and the stiffness matrix is naturally separable

$$\mathbf{K} = E_1\mathbf{K}_1 + E_2\mathbf{K}_2 = \frac{E_1}{L_1}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{E_2}{L_2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}. \tag{35}$$

Assuming no buckling would happen, we have the analytical solution of a stiffness approach

$$\mathbf{U}_{\text{ex}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \mathbf{K}^{-1}\mathbf{F} = F\begin{bmatrix} L_1/E_1 + L_2/E_2 \\ L_2/E_2 \\ 0 \end{bmatrix}. \tag{36}$$

Note that taking E_1 and E_2 as extra coordinates, Equation (36) has implied that the solutions for u_1 and u_2 are nonlinear. We select 100 DOFs to discretize the extra coordinates for functions ε_1 and ε_2.

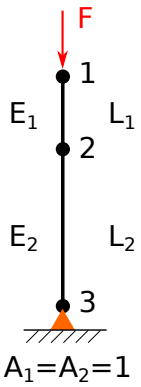


Figure 2: Sketch of the one-dimensional mechanical problem: two vertical bars.

Solving the problem with the proposed non-intrusive PGD approach, the resulting magnitudes decrease with mode sequence, which is plotted in Figure 3. From Figure 4 we can find that the relative error, which is defined as $(u_{\text{PGD}} - u_{\text{ex}})/u_{\text{ex}}$, tends to be larger when E_1 and E_2 reach their lower bounds due to the nonlinearity.

Another approach is to take the slenderness of the materials, $S_1 = 1/E_1$ and $S_2 = 1/E_2$, as extra coordinates. So that we are able to rewrite (35) and (36) as

$$\mathbf{K} = \frac{1}{S_1}\mathbf{K}_1 + \frac{1}{S_2}\mathbf{K}_2 = \frac{1}{L_1S_1}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{L_2S_2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \tag{37}$$

$$\mathbf{U}_{\text{ex}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = F\begin{bmatrix} L_1S_1 + L_2S_2 \\ L_2S_2 \\ 0 \end{bmatrix}. \tag{38}$$

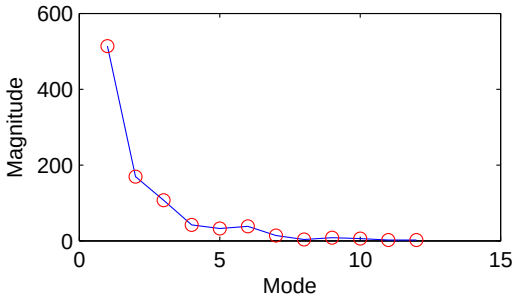


Figure 3: PGD mode-magnitudes relationship of Example I with stiffness approach.

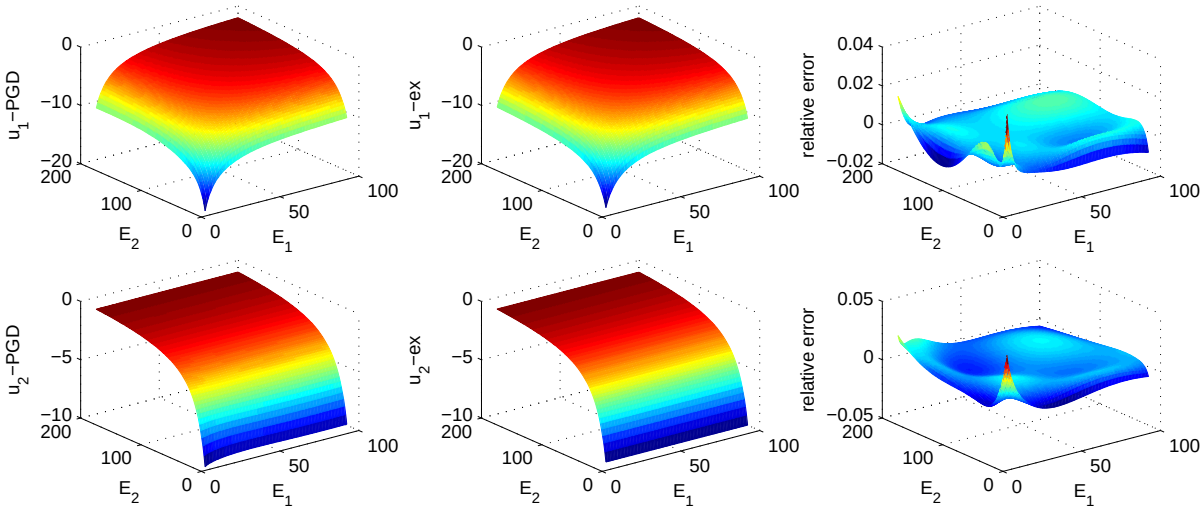


Figure 4: Verification of nodal displacements (u_1 and u_2) for Example I with stiffness approach.

The slenderness approach has converted u_1 and u_2 into linear functions of S_1 and S_2, which is demonstrated in Equation (38) and illustrated in Figure 6. Thanks to the linearity, the modes required to achieve prescribed accuracy is decreased from twelve (Figure 3) to nine (Figure 5).

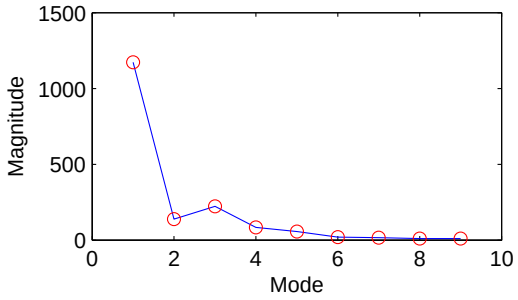


Figure 5: PGD mode-magnitudes relationship of Example I with slenderness approach.

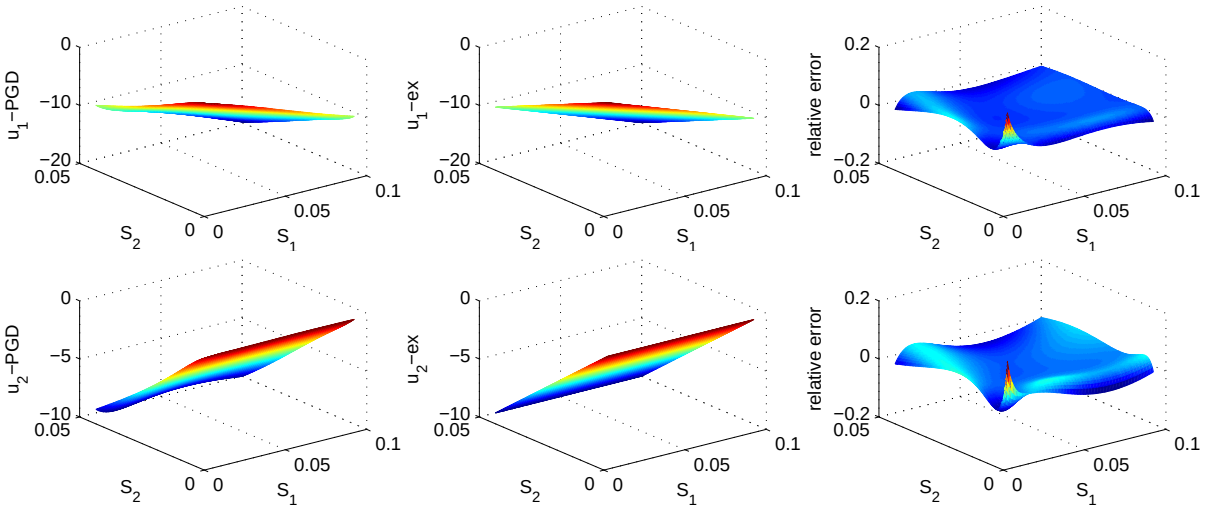


Figure 6: Verification of nodal displacements (u_1 and u_2) for Example I with slenderness approach.

3.2 Example II: a three-dimensional model of human proximal femur

As an example for biomechanical applications of the non-intrusive PGD scheme, a FE model of human proximal femur is built with tetrahedral elements in Abaqus/CAE. The femur bone is fixed at the distal end and loaded vertically at the femur head with $F = -740$ N (uniformly distributed to 74 nodes), see the sketch in Figure 7. The Young's modulus of the proximal epiphysis is $E_1 \in [8000, 10000]$ MPa and that of the cortical bone is $E_2 \in [9000, 15000]$ MPa.

Solving the problem with the proposed non-intrusive PGD scheme, the resulting mode-magnitude relationship demonstrates very well the linearity of the three-dimensional model: from Mode 2 on, the magnitudes are close to zero, which means a certain accuracy could be achieved with only two modes. The mode-magnitude plot is shown in Figure 8 and the displacement at the femur head is plotted in components in Figure 9. Since we have no analytical solution for this three-dimensional problem, there is no straightforward way to plot the error surface like Figure 4 and Figure 6. To verify the proposed non-intrusive PGD approach, we select a fixed (E_1, E_2) and compare displacements between standard FE result $\mathbf{U}_{\text{FE}}$ and non-intrusive PGD result $\mathbf{U}_{\text{PGD}}$ by computing nodal error $\mathbf{U}_{\text{PGD}} - \mathbf{U}_{\text{FE}}$. All seven modes are used to compute $\mathbf{U}_{\text{PGD}}$. For the purpose of illustration, we compute the magnitude of the nodal error at each node, and the result is shown in Figure 10.

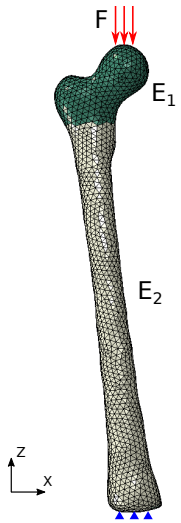


Figure 7: Sketch of the human proximal femur model.

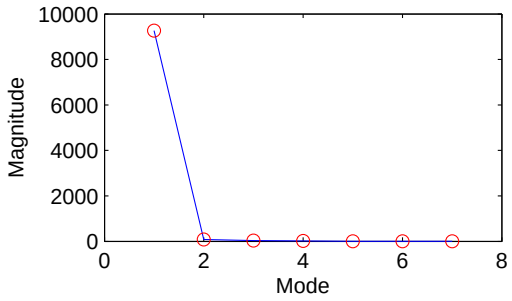


Figure 8: PGD mode-magnitude relationship for human proximal femur model.

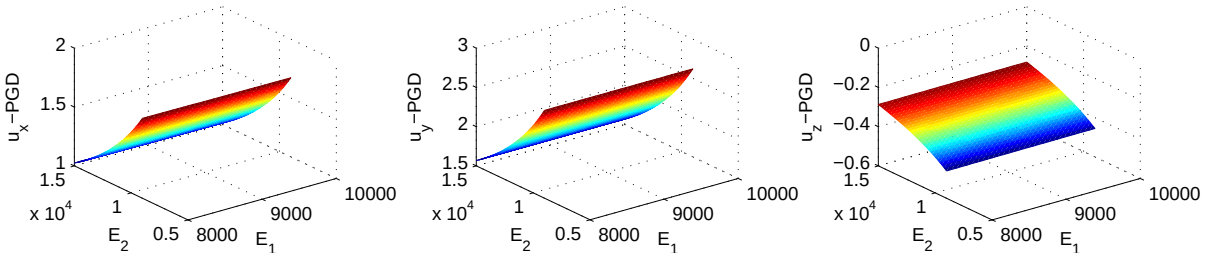


Figure 9: Non-intrusive PGD results of displacement components (u_x, u_y, u_z) at a node on the femur head. The dependence on E_1 is much less apparent than that on E_2.

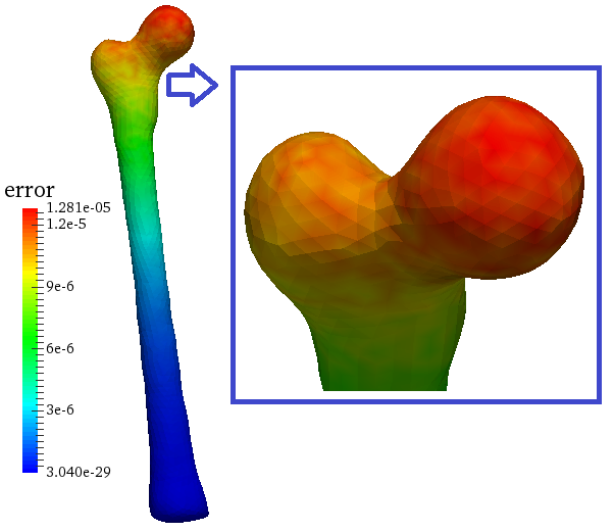


Figure 10: Magnitude of nodal error of non-intrusive PGD result with reference to standard FE result.

4 CONCLUSIONS

This work addresses a non-intrusive approach for solving multidimensional PDEs with PGD methods taking advantage of commercial FE codes with high-maturity. The material properties are considered as extra coordinates in the multidimensional governing equations of linear elasticity. One of the most significant highlights is in computation of large-scaled multidimensional problems, the effort of developing and certifying codes for solving physical problems is saved by the non-intrusive approach. In particular, the stiffness matrices and load vector are also computed by the external FE software during the pre-processing.

The proposed non-intrusive PGD scheme is highly extensible and has promisingly wide applications. For example, the presented linear elasticity problems can be easily extended to nonlinear elasticity problems [10]. In addition, geometric parameters, load and boundary conditions are also able to be taken into account as extra coordinates to be solved within the scheme.

5 ACKNOWLEDGEMENT

We acknowledge the support from the European Education, Audiovisual and Culture Executive Agency (EACEA) under the Erasmus Mundus Joint Doctorate *Simulation in Engineering and Entrepreneurship Development* (SEED), FPA 2013-0043.

Dr. M. Conti acknowledges European Reasearch Council through the ERC Starting Grant ISOBIO: Isogeometric Methods for Biomechanics (No. 259229).

REFERENCES

[1] A. Ammar, B. Mokdad, F. Chinesta, R. Keunings, A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex fluids. *Journal of Non-Newtonian Fluid Mechanics*, **139**, 153–176, 2006.

[2] A. Ammar, B. Mokdad, F. Chinesta, R. Keunings, A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modelling of complex fluids. Part II: Transient simulation using space-time separated representations. *Journal of Non-Newtonian Fluid Mechanics*, **144**, 98–121, 2007.

[3] A. Ammar, The proper generalized decomposition: A powerful tool for model reduction. *International Journal of Material Forming*, **3**, 89–102, 2010.

[4] F. Chinesta, A. Ammar, E. Cueto, Recent Advances and New Challenges in the Use of the Proper Generalized Decomposition for Solving Multidimensional Models. *Archives of Computational Methods in Engineering*, **17**, 327–350, 2010.

[5] F. Chinesta, P. Ladevèze, and E. Cueto, A Short Review on Model Order Reduction Based on Proper Generalized Decomposition. *Archives of Computational Methods in Engineering*, **18**, 395–404, 2011.

[6] F. Chinesta, A. Leygue, F. Bordeu, J.V. Aguado, E. Cueto, D. Gonzalez, I. Alfaro, A. Ammar, A. Huerta, PGD-Based Computational Vademecum for Efficient Design, Optimization and Control. *Archives of Computational Methods in Engineering*, **20**, 31–59, 2013.

[7] O.C. Zienkiewicz, R.C. Taylor, J.Z. Zhu, *The finite element method, Vol. I, 7th Edition*. Butterworth-Heinemann, 2013.

[8] A. Courard, D. Néron, P. Ladevèze, L. Ballere, Integration of PGD-virtual charts into an engineering design process. *Computational Mechanics*, 2015.

[9] S. Zlotnik, P. Díez, D. Modesto, A. Huerta, Proper generalized decomposition of a geometrically parametrized heat problem with geophysical applications. *International Journal for Numerical Methods in Engineering*, **103**, 737–758, 2015.

[10] S. Niroomandi, I. Alfaro, D. Gonzlez, E. Cueto, F. Chinesta, Model order reduction in hyperelasticity: a proper generalized decomposition approach. *International Journal for Numerical Methods in Engineering*, **96**, 129–149, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A PGD-BASED TIME SPACE DECOMPOSITION FOR THE UNSTEADY NAVIER-STOKES EQUATIONS APPLIED TO INCOMPRESSIBLE FLOWS

M. Visonneau[1]

[1]Centrale Nantes, LHEEA UMR6598, Nantes, France
e-mail: Michel.Visonneau@ec-nantes.fr

Keywords: PGD, Navier-Stokes, Finite-volume, Unsteady, Unstructured, ISIS-CFD

Abstract. *This article addresses the theoretical and numerical formulation of an algorithm based on the Proper Generalized Decomposition (PGD) applied to the solution of unsteady incompressible viscous flows. By using a separated functional description for the space and time variables, one can formulate an algorithm which may replace the traditional incremental approach and, consequently, may reduce drastically the computational time needed for the simulation of complex unsteady flows. The choice of spatial and temporal modes for velocity and pressure is discussed and several academic applications are provided, illustrating the benefits and challenges associated with this new computational approach.*

1 INTRODUCTION

The Proper Generalized Decomposition (PGD) proposes a new theoretical framework to build the continuous partial differential equations which govern the modes defining the modal decomposition of the physical quantities of interest. Separated representations were first introduced by P. Ladevèze ([1], [2]). This remarkable idea was widely extended by F. Chinesta and his co-workers ([3], [4]). The separated representation combined with the Reduced Basis concept built with a greedy algorithm is therefore a very attractive formulation which can be used to iteratively build by successive enrichments an *a priori* separated modal description of the physical fields involved in the mathematical model of interest. In this article, one describes the use and adaptation of the PGD paradigm to build a solution of the unsteady Navier-Stokes Equations for incompressible flows. Since one does not want to reduce the framework of applications to simplified Cartesian-like or large aspect ratio geometries, the separated modal decomposition concerns only the time and spatial dimensions where the space is considered as a non-separated three-dimensional entity. From that point of view, the objective of this work is to build a constructive continuous model providing an *a priori* decomposition which is very similar to the well-known *a posteriori* time-space SVD reconstruction. No further hypothesis except the existence of a separated time-space decomposition has to be made, which means that the proposed methodology is valid for unsteady flows around geometries of arbitrary complexity. The general computational framework is the one used for the code ISIS-CFD ([5]), i.e. a generalized unstructured finite volume discretization in space and a second order accurate discretisation in time.

The existence of a separated time-space solution for the unsteady incompressible Navier-Stokes equations can not be theoretically established but one feels that the parabolicity in time and the mixed parabolic/elliptic character in space for the unsteady incompressible Navier-Stokes equations provide a mathematical background which is relatively well suited to this separated modal decomposition. In a traditional incremental approach, the unsteady solution is obtained by solving n time-steps quasi-steady three-dimensional problems (where n is the number of time steps), which becomes prohibitively expensive when the number of time steps or the global duration of the simulation are large. Replacing this traditional incremental approach by a modal strategy may result in huge gains in terms of computational efficiency if the number of modes does not grow with the number of time steps. As a matter of fact, in a separated modal representation, to compute n time steps, one needs to compute $Q * N$ three-dimensional fields, where Q is the number of enrichments (or modes) and N is the number of (fixed-point) iterations needed to determine each modal group. It is clear that if $n >> N * Q$, the PGD paradigm will become extremely attractive. Moreover, since the continuous equations used to determine the temporal modes are simple ODEs, one can use a time step as small as needed without any significant penalty in terms of computational time. By combining the stored spatial and temporal modes, one can restore the full unsteady flow evolution at any time and any point with strongly reduced storage requirements, which makes extremely easy the *a-posteriori* detailed analysis of the unsteady flow. Finally, a very interesting by-product of this formulation is a reduced order flow model which could be used in situations where a high-fidelity solution is not necessarily needed like in active flow control, for instance.

In this article, one will describe the choices of decomposition for the variables involved in the unsteady Navier-Stokes decomposition, the proposed general continuous formu-

lations, the treatment of the unsteady boundary conditions and several illustrations on simple unsteady laminar flows.

2 PGD FORMULATION OF THE UNSTEADY NAVIER-STOKES EQUATIONS FOR INCOMPRESSIBLE FLOWS

2.1 The partial differential equations

To build the continuous PGD formulation, one starts from the following formulation of the unsteady incompressible Navier-Stokes equations:

$$\begin{array}{lcl} \dfrac{\partial u}{\partial t} + \hat{u} + \dfrac{\partial p}{\partial x} & = & f^u(\overline{x},t) \\ \dfrac{\partial v}{\partial t} + \hat{v} + \dfrac{\partial p}{\partial y} & = & f^v(\overline{x},t) \\ \hat{u} & = & Div(u\overrightarrow{U^{cv}}) - \dfrac{\partial}{\partial x}(\nu\dfrac{\partial u}{\partial x}) - \dfrac{\partial}{\partial y}(\nu\dfrac{\partial u}{\partial y}) \\ \hat{v} & = & Div(v\overrightarrow{U^{cv}}) - \dfrac{\partial}{\partial x}(\nu\dfrac{\partial v}{\partial x}) - \dfrac{\partial}{\partial y}(\nu\dfrac{\partial v}{\partial y}) \\ Div(\overrightarrow{U}) & = & 0 \end{array} \tag{1}$$

where $\overrightarrow{f} = (f^u, f^v)$ represents momentum source term, $\overrightarrow{U^{cv}}(u,v)$ is the convective velocity and $\overrightarrow{\hat{U}} = (\hat{u}, \hat{v})$ is an intermediate pseudo-acceleration field.

2.2 The time space PGD decomposition

Let us suppose that the velocity, pressure, pseudo-acceleration and velocity flux field can be expressed according to the following separated modal expansions :

$$\begin{array}{lcl} u(\overline{x},t) & \approx & \sum_{k=1}^{i} X_k^u(\overline{x}) T_k^u(t) \\ v(\overline{x},t) & \approx & \sum_{k=1}^{i} X_k^v(\overline{x}) T_k^u(t) \\ \hat{u}(\overline{x},t) & \approx & \sum_{k=1}^{i} \hat{X}_k^u(\overline{x}) \hat{T}_k^u(t) \\ \hat{v}(\overline{x},t) & \approx & \sum_{k=1}^{i} \hat{X}_k^v(\overline{x}) \hat{T}_k^u(t) \\ p(\overline{x},t) & \approx & \sum_{k=1}^{i} X_k^p(\overline{x}) T_k^p(t) \end{array} \tag{2}$$

The only hypothesis made here is that the temporal modes are identical for each component of the vectorial fields. Otherwise, one supposes that the other dependent variables introduced in this separated decomposition have their own temporal modes which will have to be determined by independent differential equations.
When injected in the formulation (1), one gets the following non-linear coupled equations:

$$
\begin{aligned}
&\sum_{k=1}^{i} X_k^u \frac{dT_k^u}{dt} + \sum_{k=1}^{i} \hat{X}_k^u \hat{T}_k^u + \sum_{k=1}^{i} \frac{\partial X_k^p}{\partial x} T_k^p = f^u(\overline{x},t) \\
&\sum_{k=1}^{i} X_k^v \frac{dT_k^u}{dt} + \sum_{k=1}^{i} \hat{X}_k^v \hat{T}_k^u + \sum_{k=1}^{i} \frac{\partial X_k^p}{\partial y} T_k^p = f^v(\overline{x},t) \\
&\sum_{k=1}^{i} \hat{X}_k^u \hat{T}_k^u - \sum_{k=1}^{i} \sum_{l=1}^{i} Div(X_k^u \overrightarrow{X_l}) T_k^u T_l^u + \sum_{k=1}^{i} T_k^u [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y})] = 0 \\
&\sum_{k=1}^{i} \hat{X}_k^v \hat{T}_k^u - \sum_{k=1}^{i} \sum_{l=1}^{i} Div(X_k^v \overrightarrow{X_l}) T_k^u T_l^u + \sum_{k=1}^{i} T_k^u [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y})] = 0 \\
&\sum_{k=1}^{i} Div(\overrightarrow{X_k}) T_k^u = 0
\end{aligned}
\tag{3}
$$

One can write now the weak formulation in the space-time domain of these equations:

$$
\begin{aligned}
&\int_0^{T_{max}} \int_\Omega (X_i^u T_i^{u*} + X_i^{u*} T_i^u) \left[\sum_{k=1}^{i} X_k^u \frac{dT_k^u}{dt} + \sum_{k=1}^{i} \hat{X}_k^u \hat{T}_k^u \right. \\
&\left. + \sum_{k=1}^{i} \frac{\partial X_k^p}{\partial x} T_k^p \right] d\overline{x} dt = \int_0^{T_{max}} \int_\Omega (X_i^u T_i^{u*} + X_i^{u*} T_i^u) f^u(\overline{x},t) d\overline{x} dt \\
&\int_0^{T_{max}} \int_\Omega (X_i^v T^{u*_i} + X_i^{v*} T_i^u) \left[\sum_{k=1}^{i} X_k^v \frac{dT_k^u}{dt} + \sum_{k=1}^{i} \hat{X}_k^v \hat{T}_k^u \right. \\
&\left. + \sum_{k=1}^{i} \frac{\partial X_k^p}{\partial y} T_k^p \right] d\overline{x} dt = \int_0^{T_{max}} \int_\Omega (X_i^v T_i^{u*} + X_i^{v*} T_i^u) f^v(\overline{x},t) d\overline{x} dt \\
&\int_0^{T_{max}} \int_\Omega (X_i^u T^{u*_i} + X_i^{u*} T_i^u) \left[\sum_{k=1}^{i} \hat{X}_k^u \hat{T}_k^u - \sum_{k=1}^{i} \sum_{l=1}^{i} Div(X_k^u \overrightarrow{X_l}) T_k^u T_l^u \right. \\
&\left. + \sum_{k=1}^{i} T_k^u [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y})] \right] d\overline{x} dt = 0 \\
&\int_0^{T_{max}} \int_\Omega (X_i^v T^{u*_i} + X_i^{u*} T_i^u) \left[\sum_{k=1}^{i} \hat{X}_k^v \hat{T}_k^u - \sum_{k=1}^{i} \sum_{l=1}^{i} Div(X_k^v \overrightarrow{X_l}) T_k^u T_l^u \right. \\
&\left. + \sum_{k=1}^{i} T_k^u [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y})] \right] d\overline{x} dt = 0 \\
&\int_0^{T_{max}} \int_\Omega (X_i^p T_i^{p*} + X_i^{p*} T_i^p) \left[\sum_{k=1}^{i} Div(\overrightarrow{X_k}) T_k^u \right] = 0
\end{aligned}
\tag{4}
$$

where $X_i^{u*}, X_i^{v*}, T_i^{u*}$, (resp. X_i^{p*}, T_i^{p*}) are velocity (resp. pressure) test functions, $[0, T_{max}]$ is the temporal interval and Ω is the spatial domain of integration.

2.3 The incompressibility condition

The incompressibility condition is a steady partial differential equation which expresses the local conservation of mass:

$$
Div(\overrightarrow{U}) = 0 \tag{5}
$$

Using a PGD decomposition of the velocity, one gets the following equation:

$$
\sum_{k=1}^{i} Div(\overrightarrow{X}_k) T_k^u = 0 \tag{6}
$$

Projected on the time and space dimensions using a Galerkin projection, one gets two equations for the spatial and temporal modes, respectively:

$$\sum_{k=1}^{i}(\int_0^{T_{max}} T_i^p T_k^u dt) Div(\overrightarrow{X}_k) = 0 \tag{7}$$

and

$$\sum_{k=1}^{i}(\int_\Omega X_i^p Div(\overrightarrow{X}_k) dV) T_k^u = 0 \tag{8}$$

The following PGD coefficients are introduced:

$$\begin{array}{rcl} \xi_{ik} & = & \int_0^{T_{max}} T_i^p T_k^u dt \\ e_k^p & = & \int_\Omega X_i^p Div(\overrightarrow{X}_k) dV \end{array} \tag{9}$$

which leads to:

$$\begin{array}{rcl} \sum_{k=1}^{i} \xi_{ik} Div(\overrightarrow{X}_k) & = & 0 \\ \sum_{k=1}^{i} e_k^p T_k^u & = & 0 \end{array} \tag{10}$$

Actually, if the first mode is built such that it satisfies the steady Navier-Stokes equations as it will be explained later on, $Div(\overrightarrow{X_1}) = 0$.
This means that the second mode satisfies the same conditions, and by recurrence, that every spatial mode is incompressible for any k. Therefore, the PGD solution will be decomposed on the basis of spatial modes which satisfy individually the incompressibility condition. This means also that the temporal equation is trivially satisfied since $e_k^p = 0$ and can not be used.

3 DETERMINING THE SPATIAL MODES

3.1 The PGD momentum equations

To build the spatial PDE's and temporal ODE's needed to determine the spatial and temporal modes, we use the standard Galerkin projection of the equations (3) alternatively on the spatial and temporal directions and gets a set of coupled non-linear equations. The spatial PGD equations are defined by:

$$\begin{array}{rcl} \sum_{k=1}^{i} \alpha_{ik} X_k^u + \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^u + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial x} & = & \delta_i^u(\overline{x}) \\ \sum_{k=1}^{i} \alpha_{ik} X_k^v + \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^v + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial y} & = & \delta_i^v(\overline{x}) \\ \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^u - \sum_{k=1}^{i} \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^u \overrightarrow{X_l}) + \sum_{k=1}^{i} \beta_{ik} [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y})] & = & 0 \\ \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^v - \sum_{k=1}^{i} \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^v \overrightarrow{X_l}) + \sum_{k=1}^{i} \beta_{ik} [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y})] & = & 0 \\ Div(\overrightarrow{X}^i) & = & 0 \end{array} \tag{11}$$

with the following coefficients:

$$\begin{array}{lcl} \alpha_{ik} & = & \int_0^{T_{max}} T_i^u \frac{dT_k^u}{dt} dt \\ \hat{\beta}_{ik} & = & \int_0^{T_{max}} T_i^u \hat{T}_k^u dt \\ \beta_{ik} & = & \int_0^{T_{max}} T_i^u T_k^u dt \\ \gamma_{ikl} & = & \int_0^{T_{max}} T_i^u T_k^u T_l^u dt \\ \zeta_{ik} & = & \int_0^{T_{max}} T_i^u T_k^p dt \\ \delta_i^u(\overline{x}) & = & \int_0^{T_{max}} T_i^u f^u(\overline{x},t) dt \\ \delta_i^v(\overline{x}) & = & \int_0^{T_{max}} T_i^u f^v(\overline{x},t) dt \end{array} \tag{12}$$

3.2 Transformation of the incompressibility condition into a Laplace pressure equation - Formulation 1

Let us now see how it is possible to transform the incompressibility condition into an operational partial differential equation enabling the determination of the pressure spatial modes.
As indicated before, each spatial velocity mode satisfies:

$$Div(\overrightarrow{X}_i) = 0 \tag{13}$$

The modal momentum equations read:

$$\begin{array}{lcl} \sum_{k=1}^{i} \alpha_{ik} X_k^u + \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^u + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial x} & = & \delta_i^u(\overline{x}) \\ \sum_{k=1}^{i} \alpha_{ik} X_k^v + \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^v + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial y} & = & \delta_i^v(\overline{x}) \end{array} \tag{14}$$

This means that :

$$\begin{array}{lcl} \alpha_{ii} X_i^u + \hat{\beta}_{ii} \hat{X}_i^u + \zeta_{ii} \frac{\partial X_i^p}{\partial x} & = & \delta_i^u(\overline{x}) \\ & - & \sum_{k=1}^{i-1} \alpha_{ik} X_k^u - \sum_{k=1}^{i-1} \hat{\beta}_{ik} \hat{X}_k^u - \sum_{k=1}^{i-1} \zeta_{ik} \frac{\partial X_k^p}{\partial x} \\ \alpha_{ii} X_i^v + \hat{\beta}_{ii} \hat{X}_i^v + \zeta_{ii} \frac{\partial X_i^p}{\partial y} & = & \delta_i^v(\overline{x}) \\ & - & \sum_{k=1}^{i-1} \alpha_{ik} X_k^v - \sum_{k=1}^{i-1} \hat{\beta}_{ik} \hat{X}_k^v - \sum_{k=1}^{i-1} \zeta_{ik} \frac{\partial X_k^p}{\partial y} \end{array} \tag{15}$$

Using the incompressibility of every spatial modal velcity yields the so-called "Laplace pressure equation" which couples together the spatial pressure and pseudo-acceleration modes:

$$-\hat{\beta}_{ii} Div(\overrightarrow{\hat{X}}_i) - \zeta_{ii} \Delta X_i^p = -\delta_i^{p*} \tag{16}$$

with:

$$\delta_i^{p*} = \delta_i^p - \sum_{k=1}^{i-1} \hat{\beta}_{ik} Div(\overrightarrow{\hat{X}}_k) - \sum_{k=1}^{i-1} \zeta_{ik} \Delta X_k^p \tag{17}$$

It also provides the relation which should be used to update the modal velocity flux at the control volume interface ∂V_c:

$$\begin{array}{lcl} \alpha_{ii} \overrightarrow{X}_i \cdot \overrightarrow{n} & = & \overrightarrow{\delta_i(\overline{x})} \cdot \overrightarrow{n} - \hat{\beta}_{ii} \overrightarrow{\hat{X}}_i \cdot \overrightarrow{n} - \zeta_{ii} \overrightarrow{\nabla X_i^p} \cdot \overrightarrow{n} \\ & - & \sum_{k=1}^{i-1} \alpha_{ik} \overrightarrow{X}_k \cdot \overrightarrow{n} - \sum_{k=1}^{i-1} \hat{\beta}_{ik} \overrightarrow{\hat{X}}_k \cdot \overrightarrow{n} - \sum_{k=1}^{i-1} \zeta_{ik} \overrightarrow{\nabla X_k^p} \cdot \overrightarrow{n} \end{array} \tag{18}$$

where every term is reconstructed at the interface and not interpolated from cell-centered values in agreement with the Rhie and Chow procedure to avoid any checkerboard oscillations.

3.3 Transformation of the incompressibility condition into a Poisson pressure equation - Formulation 2

If one wishes to build a pressure equation which resembles the pressure equation used in the classic incremental formulation of the unsteady Navier Stokes equations, we should proceed in a slightly different way which is explained now. The momentum equations are first rebuilt by replacing the pseudo-acceleration field by its expression, which leads to :

$$\begin{array}{rl} & \sum_{k=1}^{i} \alpha_{ik} X_k^u + \sum_{k=1}^{i} \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^u \overrightarrow{X_l}) \\ - & \sum_{k=1}^{i} \beta_{ik} [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y})] + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial x} = \delta_i^u(\overline{x}) \\ & \sum_{k=1}^{i} \alpha_{ik} X_k^v + \sum_{k=1}^{i} \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^v \overrightarrow{X_l}) \\ - & \sum_{k=1}^{i} \beta_{ik} [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y})] + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial y} = \delta_i^v(\overline{x}) \end{array} \tag{19}$$

If we introduce the following source terms δ_i^{u*} and δ_i^{v*}:

$$\begin{array}{rcl} \delta_i^{u*} & = & \delta_i^u - \sum_{k=1}^{i-1} \{\alpha_{ik} X_k^v + \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^u \overrightarrow{X_l}) - \beta_{ik} \frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) - \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y})] \\ & + & \zeta_{ik} \frac{\partial X_k^p}{\partial x}\} \\ \delta_i^{v*} & = & \delta_i^v - \sum_{k=1}^{i-1} \{\alpha_{ik} X_k^v + \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^u \overrightarrow{X_l}) - \beta_{ik} \frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) - \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y})] \\ & + & \zeta_{ik} \frac{\partial X_k^p}{\partial y}\} \end{array} \tag{20}$$

the PGD momentum equations for the current i mode read:

$$\begin{array}{l} \alpha_{ii} X_i^u + \sum_{l=1}^{i} \gamma_{iil} Div(X_i^u \overrightarrow{X_l}) - \beta_{ii} [\frac{\partial}{\partial x}(\nu \frac{\partial X_i^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_i^u}{\partial y})] + \zeta_{ii} \frac{\partial X_i^p}{\partial x} = \delta_i^{u*} \\ \alpha_{ii} X_i^v + \sum_{l=1}^{i} \gamma_{iil} Div(X_i^v \overrightarrow{X_l}) - \beta_{ii} [\frac{\partial}{\partial x}(\nu \frac{\partial X_i^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_i^v}{\partial y})] + \zeta_{ii} \frac{\partial X_i^p}{\partial y} = \delta_i^{v*} \end{array} \tag{21}$$

Once discretised with a finite volume discretization, the momentum equations read for each control volume V_c:

$$\begin{array}{rcl} Vol_c \frac{\alpha_{ii}}{\beta_{ii}} X_c^u + \sum C_{nb} X_{nb}^u + C_d X_c^u + Vol_c \frac{\zeta_{ii}}{\beta_{ii}} Discr\left[\frac{\partial}{\partial x} X^p\right] + Src_U & = & 0 \\ Vol_c \frac{\alpha_{ii}}{\beta_{ii}} X_c^v + \sum C_{nb} X_{nb}^v + C_d X_c^v + Vol_c \frac{\zeta_{ii}}{\beta_{ii}} Discr\left[\frac{\partial}{\partial y} X^p\right] + Src_V & = & 0 \end{array} \tag{22}$$

where:

$$\begin{array}{rcl} Src_U & = & -Vol_c \frac{\delta_i^{u*}}{\beta_{ii}} \\ Src_V & = & -Vol_c \frac{\delta_i^{v*}}{\beta_{ii}} \end{array} \tag{23}$$

The indices i of the mode i are no more mentioned for the sake of simplicity. The operator $Discr$ stands for "Discretization of". Vol_c is the volume of the cell of integration, C_{nb} and C_d are the discretisation coefficients of the convective-diffusive common operator present in both momentum equations.
In order to build a pressure equation, a new discrete pseudo-velocity field $(\widetilde{X}_c^u, \widetilde{X}_c^v)$ is introduced:

$$\begin{array}{rcl} \widetilde{X}_c^u - \frac{1}{Vol_c} \sum C_{nb} X_{nb}^u & = & \frac{SrcU}{Vol_c} = -\frac{\delta_i^{u*}}{\beta_{ii}} \\ \widetilde{X}_c^v - \frac{1}{Vol_c} \sum C_{nb} X_{nb}^v & = & \frac{SrcV}{Vol_c} = -\frac{\delta_i^{v*}}{\beta_{ii}} \end{array} \tag{24}$$

which leads to:

$$\begin{array}{rcl} X_c^u & = & -C_p \left(\widetilde{X}_c^u + Vol_c \frac{\zeta_{ii}}{\beta_{ii}} Discr\left[\frac{\partial}{\partial x} X^p\right]\right) \\ X_c^v & = & -C_p \left(\widetilde{X}_c^v + Vol_c \frac{\zeta_{ii}}{\beta_{ii}} Discr\left[\frac{\partial}{\partial y} X^p\right]\right) \end{array} \tag{25}$$

with $C_{Diag} = C_d + Vol_c \frac{\alpha_{ii}}{\beta_{ii}}$ and $C_p = \frac{Vol_c}{C_{Diag}}$.
By using the solenoidality condition, one gets and alternate pressure equation which can be written with the following semi-discretized formulation:

$$-Div(C_p \overrightarrow{\tilde{X}}) - Div(C_p \frac{\zeta_{ii}}{\beta_{ii}} \overrightarrow{\nabla X^p}_i) \quad = \quad 0 \tag{26}$$

3.4 Comments on the spatial formulation

In this section, we have built the spatial PGD equations which have to be solved to determine the spatial velocity and pressure modes. In order to have spatial velocity modes satisfying individually the incompressibility condition, we have proposed two possibilities to transform the solenoidality condition into a pressure equation. The first one is built from the continuous formulation of the momentum equations in which an intermediate pseudo-acceleration vectorial field has been introduced. Taking the divergence of these equations leads to a partial differential equation combining a Laplace pressure equation and the divergence of the pseudo-acceleration field. The second option which corresponds to the usual procedure for the finite volume approach for incompressible flows consists in discretizing first the momentum equations and reconstructing the fluxes of the velocity components and pressure gradient, while all the contributions of the neighboring data and source terms are gathered into the so-called pseudo-velocity fluxes. Proceeding in this way transforms the solenoidality condition into a Poisson pressure equation and a divergence of the pseudo velocity vectorial field. In that case, the discretization coefficients of the momentum equations enter the formulation of the spatial pressure equation through the central coefficient C_p and the definition of the pseudo-velocity fields.

3.5 Final spatial formulation

To summarize, one gets the final coupled non-linear formulation for the spatial PGD modes:

$$\begin{array}{lcl}
\sum_{k=1}^{i} \alpha_{ik} X_k^u + \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^u + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial x} & = & \delta_i^u(\overline{x}) \\
\sum_{k=1}^{i} \alpha_{ik} X_k^v + \sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^v + \sum_{k=1}^{i} \zeta_{ik} \frac{\partial X_k^p}{\partial y} & = & \delta_i^v(\overline{x}) \\
\sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^u - \sum_{k=1}^{i} \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^u \overrightarrow{X_l}) + \sum_{k=1}^{i} \beta_{ik} [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y})] & = & 0 \\
\sum_{k=1}^{i} \hat{\beta}_{ik} \hat{X}_k^v - \sum_{k=1}^{i} \sum_{l=1}^{i} \gamma_{ikl} Div(X_k^v \overrightarrow{X_l}) + \sum_{k=1}^{i} \beta_{ik} [\frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y})] & = & 0 \\
-\sum_{k=1}^{i} \hat{\beta}_{ik} Div(\overrightarrow{\hat{X}}_k) - \sum_{k=1}^{i} \zeta_{ik} \Delta X_k^p & = & -\delta_i^p(\overline{x}) \\
\textit{or} & & \\
-\beta_{ii} Div(C_p \overrightarrow{\tilde{X}}_i) - \zeta_{ii} Div(C_p \nabla X_i^p) & = & 0
\end{array} \tag{27}$$

with:

$$\begin{array}{lcl}
\delta_i^u(\overline{x}) & = & \int_0^{T_{max}} T_i^u f^u(\overline{x},t)dt \\
\delta_i^v(\overline{x}) & = & \int_0^{T_{max}} T_i^u f^v(\overline{x},t)dt \\
\delta_i^p(\overline{x}) & = & \int_0^{T_{max}} T_i^u Div(\overrightarrow{f}(\overline{x},t))dt \\
\alpha_{ik} & = & \int_0^{T_{max}} T_i^u \frac{dT_k^u}{dt}dt \\
\hat{\beta}_{ik} & = & \int_0^{T_{max}} T_i^u \hat{T}_k^u dt \\
\beta_{ik} & = & \int_0^{T_{max}} T_i^u T_k^u dt \\
\gamma_{ikl} & = & \int_0^{T_{max}} T_i^u T_k^u T_l^u dt \\
\zeta_{ik} & = & \int_0^{T_{max}} T_i^u T_k^p dt
\end{array} \tag{28}$$

4 DETERMINING THE TEMPORAL MODES

4.1 Comments on the pressure temporal modes

It has been noticed previously that it is impossible to build an equation to determine the temporal modes of the pressure T_i^p if one starts from the original incompressibility condition $Div(\overrightarrow{U}) = 0$. Moreover, if one starts from the classic semi-discretized Poisson equation for the pressure (formulation 2), it is also impossible to build a consistent temporal differential equation since this semi-discretized equation comprises terms (C_p and $\overrightarrow{\widetilde{X}_i}$) which come from the discretisation of the momentum equations. It appears therefore that the only valid starting point to determine the pressure temporal modes is the differential Laplace pressure equation introduced in 3.2.

4.2 Temporal PGD ordinary differential equations

Based on the set of equations (3), the temporal PGD equations consist in a set of five coupled non-linear algebraic differential equations:

$$\begin{array}{lcl}
\sum_{k=1}^i a_{ik}^u \frac{dT_k^u}{dt} + \sum_{k=1}^i \hat{b}_{ik}^u \hat{T}_k^u + \sum_{k=1}^i e_{ik}^u T_k^p & = & d_i^u(t) \\
\sum_{k=1}^i a_{ik}^v \frac{dT_k^u}{dt} + \sum_{k=1}^i \hat{b}_{ik}^v \hat{T}_k^u + \sum_{k=1}^i e_{ik}^v T_k^p & = & d_i^v(t) \\
\sum_{k=1}^i \hat{b}_{ik}^u \hat{T}_k^u - \sum_{k=1}^i \sum_{l=1}^i c_{ikl}^u T_k^u T_l^u - \sum_{k=1}^i b_{ik}^u T_k^u & = & 0 \\
\sum_{k=1}^i \hat{b}_{ik}^v \hat{T}_k^u - \sum_{k=1}^i \sum_{l=1}^i c_{ikl}^v T_k^u T_l^u - \sum_{k=1}^i b_{ik}^v T_k^u & = & 0 \\
\sum_{k=1}^i a_{ik}^p T_k^p + \sum_{k=1}^i \hat{b}_{ik}^p \hat{T}_k^u & = & d_i^p(t)
\end{array} \tag{29}$$

where the following coefficients have been introduced:

$$
\begin{array}{rcl}
a_{ik}^u & = & \int_\Omega X_i^u X_k^u d\overline{x} \\
a_{ik}^v & = & \int_\Omega X_i^v X_k^v d\overline{x} \\
\hat{b}_{ik}^u & = & \int_\Omega X_i^u \hat{X}_k^u d\overline{x} \\
\hat{b}_{ik}^v & = & \int_\Omega X_i^v \hat{X}_k^v d\overline{x} \\
c_{ikl}^u & = & \int_\Omega X_i^u Div(X_k^u \overrightarrow{X_l}) d\overline{x} \\
c_{ikl}^v & = & \int_\Omega X_i^v Div(X_k^v \overrightarrow{X_l}) d\overline{x} \\
b_{ik}^u & = & \int_\Omega X_i^u \left[-\frac{\partial}{\partial x}(\nu \frac{\partial X_k^u}{\partial x}) - \frac{\partial}{\partial y}(\nu \frac{\partial X_k^u}{\partial y}) \right] d\overline{x} \\
b_{ik}^v & = & \int_\Omega X_i^v \left[-\frac{\partial}{\partial x}(\nu \frac{\partial X_k^v}{\partial x}) - \frac{\partial}{\partial y}(\nu \frac{\partial X_k^v}{\partial y}) \right] d\overline{x} \\
e_{ik}^u & = & \int_\Omega X_i^u \frac{\partial X_k^p}{\partial x} d\overline{x} \\
e_{ik}^v & = & \int_\Omega X_i^v \frac{\partial X_k^p}{\partial y} d\overline{x} \\
d_i^u & = & \int_\Omega X_i^u f^u(\overline{x},t) d\overline{x} \\
d_i^v & = & \int_\Omega X_i^v f^v(\overline{x},t) d\overline{x} \\
a_{ik}^p & = & \int_\Omega X_i^p (-Div \left[\overrightarrow{\nabla X_k^p} \right]) d\overline{x} \\
\hat{b}_{ik}^p & = & \int_\Omega X_i^p (-Div \left[\overrightarrow{\hat{X}_k} \right]) d\overline{x} \\
d_i^p & = & \int_\Omega X_i^p (-Div[\overrightarrow{f}]) d\overline{x}
\end{array}
\tag{30}
$$

We have now five ordinary differential equations to determine three unknowns T_i^u, $\hat{T}_i^u$ and T_i^p. We will use the sum of the first and second equations and the sum of the third and fourth equations to provide a system coupling together T_i^u, $\hat{T}_i^u$ and T_i^p:

$$
\begin{array}{lcl}
\sum_{k=1}^i (a_{ik}^u + a_{ik}^v) \frac{dT_k^u}{dt} + \sum_{k=1}^i (\hat{b}_{ik}^u + \hat{b}_{ik}^v) \hat{T}_k^u + \sum_{k=1}^i (e_{ik}^u + e_{ik}^v) T_k^p & = & d_i^u(t) + d_i^v(t) \\
\sum_{k=1}^i (\hat{b}_{ik}^u + \hat{b}_{ik}^v) \hat{T}_k^u - \sum_{k=1}^i \sum_{l=1}^i (\hat{c}_{ikl}^u + \hat{c}_{ikl}^v) T_k^u T_l^u - \sum_{k=1}^i (\hat{f}_{ik}^u + \hat{f}_{ik}^v) T_k^u & = & 0 \\
\sum_{k=1}^i a_{ik}^p T_k^p + \sum_{k=1}^i \hat{b}_{ik}^p \hat{T}_k^u & = & d_i^p(t)
\end{array}
\tag{31}
$$

and these coupled non-linear differential equations will be iteratively solved with an appropriate linearisation strategy. It is worthwhile to underline that the last algebraic equation provides a non-linear (actually quadratic) relation between the pressure and the velocity temporal modes which is the temporal counterpart of the Bernoulli equation.
At this stage, it should be noticed that it is not necessary easy to provide accurate evaluations of these temporal coefficients with a formal 2nd order accurate finite volume solver. This means that specific discretisation schemes have to be built to provide an accurate enough discretisation of these operators on a fully unstrctured grid. For the sake of conciseness, we will not elaborate more on this topic which is however felt crucial for the modal convergence of the PGD decomposition.

5 SPACE-TIME CONSISTENCY RELATIONS

It might be useful to notice that there are additional equations linking together the spatial and temporal PGD coefficients. These relations are obtained, for instance, by integrating over the whole flow domain the spatial modal equations. This yields the following

relations:

$$\begin{array}{lcl}
\sum_{l=1}^{k} \alpha_{kl} a_{il}^{u} + \sum_{l}^{k} \hat{\beta}_{kl} b_{il}^{u} + \sum_{l}^{k} \zeta_{kl} e_{il}^{u} & = & \int_{\Omega} X_i^u \delta_k^u d\overline{x} \\
\sum_{l=1}^{k} \alpha_{kl} a_{il}^{v} + \sum_{l}^{k} \hat{\beta}_{kl} b_{il}^{v} + \sum_{l}^{k} \zeta_{kl} e_{il}^{v} & = & \int_{\Omega} X_i^v \delta_k^v d\overline{x} \\
\sum_{l=1}^{k} \hat{\beta}_{kl} \hat{b}_{ik}^{u} - \sum_{l=1}^{k} \sum_{m=1}^{k} \gamma_{klm} c_{ikm}^{u} - \sum_{l=1}^{k} \beta_{kl} b_{ik}^{u} & = & 0 \\
\sum_{l=1}^{k} \hat{\beta}_{kl} \hat{b}_{ik}^{v} - \sum_{l=1}^{k} \sum_{m=1}^{k} \gamma_{klm} c_{ikm}^{v} - \sum_{l=1}^{k} \beta_{kl} b_{ik}^{v} & = & 0 \\
\sum_{l=1}^{k} \zeta_{kl} a_{ik}^{p} T_k^p + \sum_{l=1}^{k} \hat{\beta}_{kl} \hat{b}_{ik}^{p} & = & \int_{\Omega} X_i^p \delta_k^p d\overline{x}
\end{array} \tag{32}$$

which are valid for any indices i and k. These relations establish a kind of global time-space consistency of the PGD coefficients and can be used to ensure structural relations which might not be verified at a discrete level, due to the influence of the spatial discretisation errors on the determination of the temporal PGD coefficients.

6 COMMENTS ON THE GENERAL COMPUTATIONAL STRATEGY

This PGD algorithm is therefore comprised of four embedded loops:

1. an enrichment loop which adds new temporal and spatial modes to enrich the modal decomposition until modal convergence,

2. a fixed point loop which ensures the mutual convergence of the couple of spatial and temporal modes which are solved sequentially. It should be noted here that we have chosen to converge on the non-linearity at each fixed point iteration, either for the spatial or temporal modes. Although such a procedure is likely to be very expensive, we preferred to use it to ensure a safe overall convergence.

3. a non-linear loop which accounts for the non-linearities present in the spatial PDE's and the temporal ODE's,

4. a linear coupling loop which takes into account the linear coupling existing between the linearized spatial momentum and pressure equations on one hand, and on the other hand, the linear coupling between the linearized temporal equations for the temporal velocity and pressure modes.

In this version of the **ISIS-CFD** code, the linear coupling between the spatial equations is treated with a fully coupled formulation for which the fully-coupled saddle-point linear system corresponding to the linearised momentum and pressure equations is directly solved by *ad-hoc* preconditioned Krylov solvers in the spirit of previous developments performed in our team ([6]).

7 TAKING INTO ACCOUNT THE BOUNDARY CONDITIONS AND RELATED SIMPLIFICATIONS

7.1 Satisfying unsteady boundary conditions

The PGD formulation is well posed when the boundary conditions are homogeneous. When the unsteadiness of the flow can be reached naturally, thanks to the development of internal instabilities in the flow like for the unsteady behind a cylinder, this peculiarity of the PGD methodology does not pose any problem. However, when the unsteadiness of the flow is imposed by moving boundaries, one should build a first set of modes which

satisfy the unsteady boundary conditions. This means that the velocity boundary condition should be written in a time space separated form in order to comply with the present choice. For simple academic flows like the unsteady driven cavity, or the flow between two concentric cyinders, it is easy to design a procedure to take into account the unsteadiness of the boundary conditions.
Let us take for instance the case of a driven square cavity where the horizontal component of the velocity is imposed at the upper wall by:

$$U_{wall} = f(t) \tag{33}$$

We select the steady flow solution associated with $U_{wall} = 1$ to provide the initial spatial modes and one multiplies this steady solution by the temporal function involved at the upper wall of the driven cavity to build a peculiar flow solution satisfying the unsteady boundary conditions but, of course, not the unsteady Navier-Stokes equations. The steady flow solution $X_1^u(\overline{x}), X_1^v(\overline{x}), X_1^p(\overline{x})$ satisfies the steady Navier-Stokes equations and we can use it as first spatial modes, the first temporal velocity modes being provided by $T_1^u(t) = f(t)$ while the first pressure temporal mode has still to be specified. The first modes are therefore designed in such a way that the unsteady boundary conditions are satisfied, which means that the next modes just need to satisfy homogeneous Dirichlet (velocity) or Neumann (pressure) boundary conditions.

$$\begin{array}{rcl} u(\overline{x},t) & = & u^{(2)}(\overline{x},t) + X_1^u(\overline{x})T_1^u(t) \\ v(\overline{x},t) & = & v^{(2)}(\overline{x},t) + X_1^v(\overline{x})T_1^u(t) \\ p(\overline{x},t) & = & p^{(2)}(\overline{x},t) + X_1^p(\overline{x})T_1^p(t) \end{array} \tag{34}$$

However, we also need to prescribe the temporal and spatial first modes for the pressure and the pseudo-acceleration in the framework of the proposed formulation. For the pseudo-acceleration, we can simply use its definition:

$$\begin{array}{rcl} \hat{b}_{11}^u \hat{T}_{11}^u & = & c_{111}^u (T_1^u)^2 + b_{11}^u T_1^u \\ \hat{b}_{11}^v \hat{T}_{11}^u & = & c_{111}^v (T_1^u)^2 + b_{11}^v T_1^u \end{array} \tag{35}$$

Moreover, the fact that the first spatial mode is solution of the steady Navier-Stokes equations provides the following relations:

$$\begin{array}{rcl} \hat{b}_{11}^u & = & c_{111}^u + b_{11}^u \\ \hat{b}_{11}^v & = & c_{111}^v + b_{11}^v \end{array} \tag{36}$$

which leads us to the formula used to initialize the first pseudo-acceleration temporal mode:

$$\hat{T}_1^u = \frac{(c_{111}^u + c_{111}^v)(T_1^u)^2 + (b_{11}^u + b_{11}^v) T_1^u}{c_{111}^u + c_{111}^v + b_{11}^u + b_{11}^v} \tag{37}$$

The optimal initialisation of the first pressure temporal mode is still an open problem. To determine the initial temporal pressure mode, we might suppose that the flow solution based on the first PGD mode satisfies the incompressibility condition:

$$a_{11}^p T_1^p + \hat{b}_{11}^p \hat{T}_1^u = 0 \tag{38}$$

If one supposes that the first spatial pressure mode is solution of the steady Navier-Stokes equations:

$$a_{11}^p + \hat{b}_{11}^p = 0 \tag{39}$$

this leads to:

$$T_1^p = \hat{T}_1^u \tag{40}$$

But, there is no justification of the above-mentioned hypothesis and the determination of the first temporal mode can be considered as an open question.
When the unsteady boundary conditions are more complex, one can build a time-space Singular Value Decomposition of the unsteady boundary conditions, which leads to the following expression:

$$\begin{array}{rcl} u^{(bnd)} & = & \sum_{k=1}^{N} X_k^{u(bnd)}(\overline{x}) T_k^u(t) \\ v^{(bnd)} & = & \sum_{k=1}^{N} X_k^{v(bnd)}(\overline{x}) T_k^u(t) \\ p^{(bnd)} & = & \sum_{k=1}^{N} X_k^{p(bnd)}(\overline{x}) T_k^p(t) \end{array} \tag{41}$$

We can then solve N steady Navier-Stokes equations with the boundary conditions provided by $X_k^{u(bnd)}$ and $X_k^{v(bnd)}$ to initialize the spatial modes in the flow domain. Once these modes have been determined, the remaining PGD modes will satisfy homogeneous velocity conditions. This procedure can be tedious in case of very complex unsteady boundary conditions but this is a known limitation of the PGD paradigm.

7.2 Simplifications related with the choice of the initial temporal and spatial modes

As indicated before, when the boundary conditions are not homogeneous, it is necessary to select the k first temporal and spatial modes such that the unsteady boundary conditions are satisfied by their combination. It is interesting to notice that if, these first k spatial modes satisfy also the steady incompressible Navier-Stokes equations and if $\hat{T}_k^u = T_k^p$ for all the k modes used to ensure the boundary conditions, then for every next $i > k$ mode, the following relations hold:

$$\begin{array}{l} -Div[\overrightarrow{\nabla X_i^p}] - Div[\overrightarrow{\hat{X}_i}] = 0 \\ \hat{T}_i^u = T_i^p \end{array} \tag{42}$$

which greatly simplifies the PGD formulation and reduces the number of independent PGD coefficients.

8 SIMPLIFICATIONS FOR THE STOKES EQUATIONS

For the Stokes equations, the non-linear convective terms are absent. If one chooses the first spatial mode such that it satisfies the steady Stokes equations, we have the following relations:

$$\begin{array}{rcl} \hat{X}_1^u & = & -\frac{\partial X_1^p}{\partial x} \\ \hat{X}_1^v & = & -\frac{\partial X_1^p}{\partial y} \\ \hat{X}_1^u & = & \frac{\partial}{\partial x}(\nu \frac{\partial X_1^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_1^u}{\partial y}) \\ \hat{X}_1^v & = & \frac{\partial}{\partial x}(\nu \frac{\partial X_1^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_1^v}{\partial y}) \end{array} \tag{43}$$

and then, for every i:

$$\begin{array}{rcl} \hat{b}_{i1}^u & = & b_{i1}^u \\ \hat{b}_{i1}^v & = & b_{i1}^v \\ \hat{b}_{i1}^u & = & -e_{i1}^u \\ \hat{b}_{i1}^v & = & -e_{i1}^v \\ a_{i1}^p & = & -\hat{b}_{i1}^p \end{array} \tag{44}$$

Since $\hat{T}_1^u = T_1^u$, if one chooses $T_1^p = T_1^u$, this leads to the following coupled equations for the second PGD spatial and temporal modes:

$$\begin{array}{lcl}
\alpha_{22} X_2^u + \hat{\beta}_{22} \hat{X}_2^u + \zeta_{22} \frac{\partial X_2^p}{\partial x} & = & -\alpha_{21} X_1^u \\
\alpha_{22} X_2^v + \hat{\beta}_{22} \hat{X}_2^v + \zeta_{22} \frac{\partial X_2^p}{\partial y} & = & -\alpha_{21} X_1^v \\
\hat{\beta}_{22} \hat{X}_2^u + \beta_{22} [\frac{\partial}{\partial x}(\nu \frac{\partial X_2^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_2^u}{\partial y})] & = & 0 \\
\hat{\beta}_{22} \hat{X}_2^v + \beta_{22} [\frac{\partial}{\partial x}(\nu \frac{\partial X_2^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_2^v}{\partial y})] & = & 0 \\
-\zeta_{22} Div[\overrightarrow{\nabla X_2^p}] - \hat{\beta}_{22} Div[\hat{\vec{X}}_2] & = & 0 \\
(a_{22}^u + a_{22}^v) \frac{dT_2^u}{dt} + (\hat{b}_{22}^u + \hat{b}_{22}^v) \hat{T}_2^u + (e_{22}^u + e_{22}^v) T_2^p & = & -(a_{21}^u + a_{21}^v) \frac{dT_1^u}{dt} \\
(\hat{b}_{22}^u + \hat{b}_{22}^v) \hat{T}_2^u - (b_{22}^u + b_{22}^v) T_2^u & = & 0 \\
a_{22}^p T_2^p + \hat{b}_{22}^p \hat{T}_2^u & = & 0
\end{array} \tag{45}$$

The solutions of this system are :

$$\begin{array}{lcl}
T_2^p & = & T_2^u \\
\hat{T}_2^u & = & T_2^u \\
(a_{22}^u + a_{22}^v) \frac{dT_2^u}{dt} + (\hat{a}_{22}^u + e_{22}^u + \hat{a}_{22}^v + e_{22}^v) T_2^u & = & -(a_{21}^u + a_{21}^v) \frac{dT_1^u}{dt} \\
-Div[\overrightarrow{\nabla X_2^p}] - Div[\hat{\vec{X}}_2] & = & 0 \\
\hat{X}_2^u + [\frac{\partial}{\partial x}(\nu \frac{\partial X_2^u}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_2^u}{\partial y})] & = & 0 \\
\hat{X}_2^v + [\frac{\partial}{\partial x}(\nu \frac{\partial X_2^v}{\partial x}) + \frac{\partial}{\partial y}(\nu \frac{\partial X_2^v}{\partial y})] & = & 0 \\
\alpha_{22} \hat{X}_2^u + \hat{\beta}_{22} (\hat{X}_2^u + \frac{\partial X_2^p}{\partial x}) & = & -\alpha_{21} \hat{X}_1^u \\
\alpha_{22} \hat{X}_2^v + \hat{\beta}_{22} (\hat{X}_2^v + \frac{\partial X_2^p}{\partial y}) & = & -\alpha_{21} \hat{X}_1^v
\end{array} \tag{46}$$

It is trivial to demonstrate by recurrence for the next temporal modes that $\hat{T}_i^u = T_i^u$ and $\hat{T}_i^p = T_i^u$, which proves that this formulation naturally provides a decomposition for which every temporal pressure mode is equal to the velocity temporal mode, thanks to the linearity of the Stokes equations.

9 APPLICATIONS

9.1 Analytic solutions

9.1.1 Stokes equations

A. Dumont ([7]) used in his PhD thesis the following 2D analytic solutions for the Stokes equations:

$$\begin{array}{lcl}
\frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} - \frac{\partial}{\partial x}(\nu \frac{\partial u}{\partial x}) - \frac{\partial}{\partial y}(\nu \frac{\partial u}{\partial y}) + f^u(\overline{x}, t) & = & 0 \\
\frac{\partial v}{\partial t} + \frac{\partial p}{\partial y} - \frac{\partial}{\partial x}(\nu \frac{\partial v}{\partial x}) - \frac{\partial}{\partial y}(\nu \frac{\partial v}{\partial y}) + f^v(\overline{x}, t) & = & 0 \\
-Div\left[\overrightarrow{\nabla p}\right] - Div\left[\hat{\vec{U}}\right] + f^p(\overline{x}, t) & = & 0
\end{array} \tag{47}$$

with:

$$\begin{array}{rcl} u(x,y,t) & = & cos(\omega x)sin(\omega y)e^{-t} \\ v(x,y,t) & = & -sin(\omega x)cos(\omega y)e^{-t} \\ p(x,y,t) & = & (1-2\omega^2\nu)sin(\omega x)sin(\omega y)e^{-t} \\ f^u(x,y,t) & = & -(1-\omega)(2\omega^2\nu-1)cos(\omega x)sin(\omega y)e^{-t} \\ f^v(x,y,t) & = & -(1+\omega)(1-2\omega^2\nu)sin(\omega x)cos(\omega y)e^{-t} \\ f^p(x,y,t) & = & -(\dfrac{\partial f^u}{\partial x}+\dfrac{\partial f^v}{\partial y}) \\ & = & -2\omega^2(2\omega^2\nu-1)sin(\omega x)sin(\omega y)e^{-t} \end{array} \tag{48}$$

with $\omega = 2\pi$. This manufactured solution is imposed in a squared cavity $[0.;1.] \times [0.;1.]$ with the following boundary conditions and the PGD decomposition converges in one mode analytically and numerically.

9.1.2 Navier-Stokes equations

The unsteady Taylor-Green vortex is an analytic solution of the unsteady Navier-Stokes equations. It is given by:

$$\begin{array}{rcl} u(x,y,t) & = & sin(\omega x)cos(\omega y)e^{-2\omega^2\nu t} \\ v(x,y,t) & = & -cos(\omega x)sin(\omega y)e^{-2\omega^2\nu t} \\ p(x,y,t) & = & \frac{1}{4}(cos(2\omega x)+cos(2\omega y))e^{-4\omega^2\nu t} \end{array} \tag{49}$$

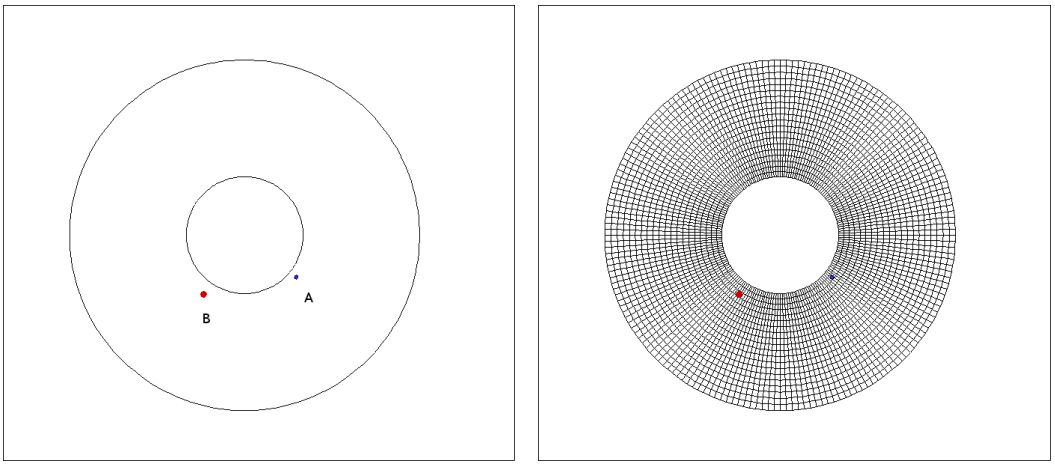
with $\omega = 2\pi$. In that case, the spatial and temporal modes are provided by:

$$\begin{array}{rcl} X_1^u & = & sin(\omega x)cos(\omega y) \\ X_1^v & = & -cos(\omega x)sin(\omega y) \\ X_1^p & = & \frac{1}{4}(cos(2\omega x)+cos(2\omega y)) \\ T_1^u & = & e^{-2\omega^2\nu t} \\ T_1^p & = & e^{-4\omega^2\nu t} \end{array} \tag{50}$$

As previously, the PGD decomposition converges in one mode analytically and numerically. This is the first example where the temporal mode for the pressure is the square of the temporal velocity mode, justifying the choice of independent pressure and velocity temporal modes.

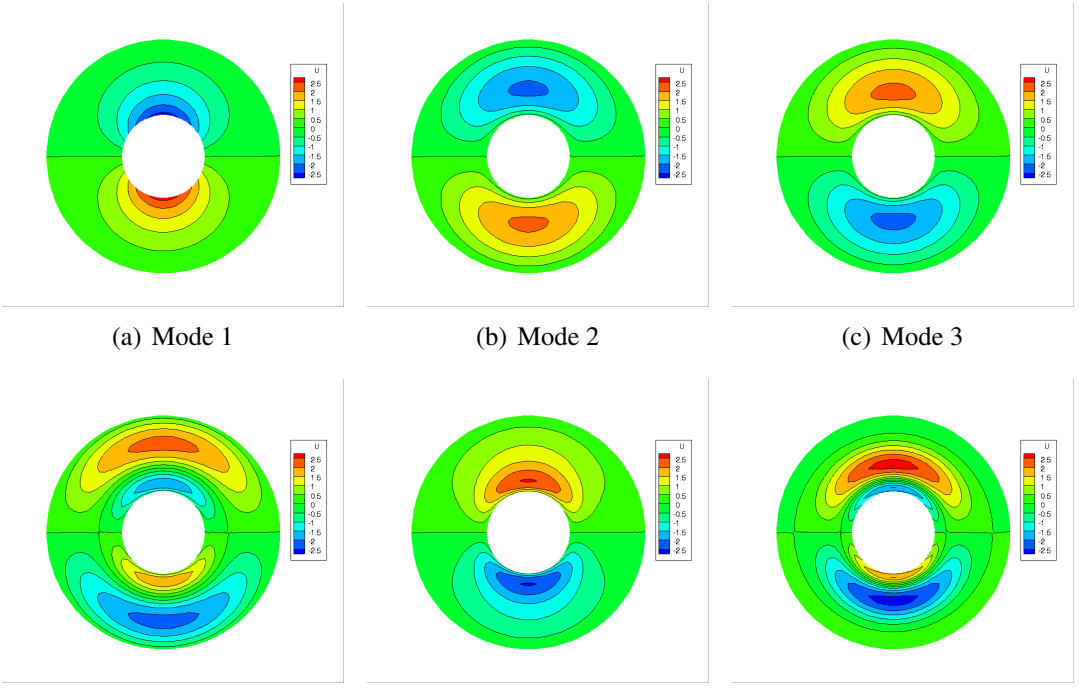
9.2 UNSTEADY FLOW BETWEEN TWO CONCENTRIC CYLINDERS

The overall procedure is validated on an academic flow configuration, the computation of the unsteady viscous flow between two concentric cylinders (see figure1(a)). The flow domain is limited by two cylinders and (R_{in}=1., R_{out}=3.0) and the inner cylinder is rotating with a rotation velocity defined by $\Omega(t) = \Omega_0 sin(2\pi t)$. The Reynolds number of this flow is $Re = 20$. A structured grid composed of 180 (resp. 21) points distributed circumferentially (resp. radially) is shown in figure 1(b) where two specific blue (A) and red (B) points are introduced. The point A is located close to the inner rotating wall while the point B is positioned further away in a region where the convective effects are less prominent.
Figures 2, 3, 4 show the six first normalized spatial modes for the two components of the velocity and the pressure.



(a) Computational domain (b) Grid

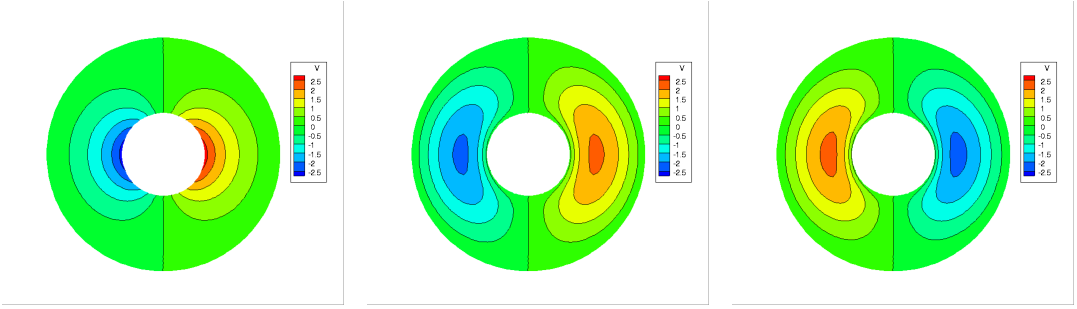
Figure 1: Unsteady viscous flow between two concentric cylinders - Computational domain



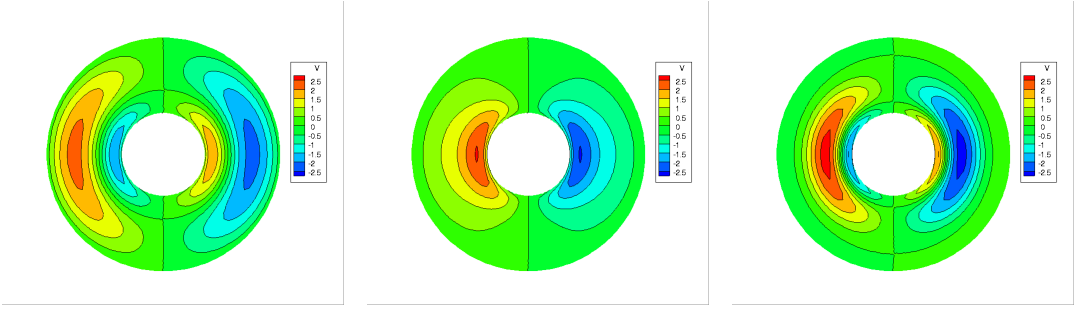
(a) Mode 1 (b) Mode 2 (c) Mode 3

(d) Mode 4 (e) Mode 5 (f) Mode 6

Figure 2: The six first normalized spatial PGD modes for U

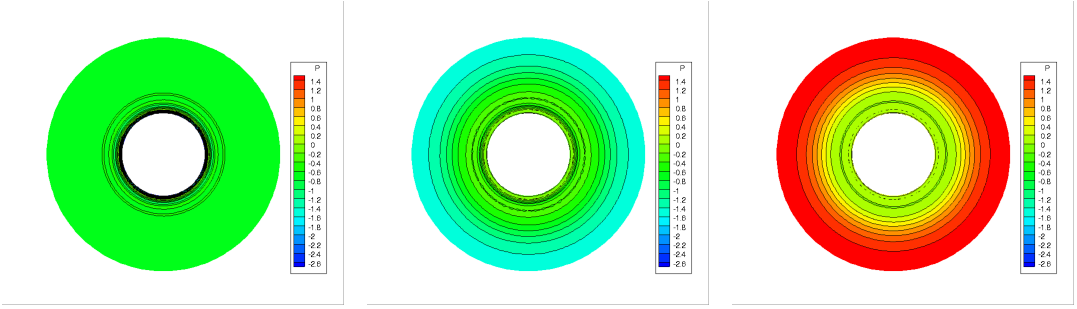


(a) Mode 1 (b) Mode 2 (c) Mode 3

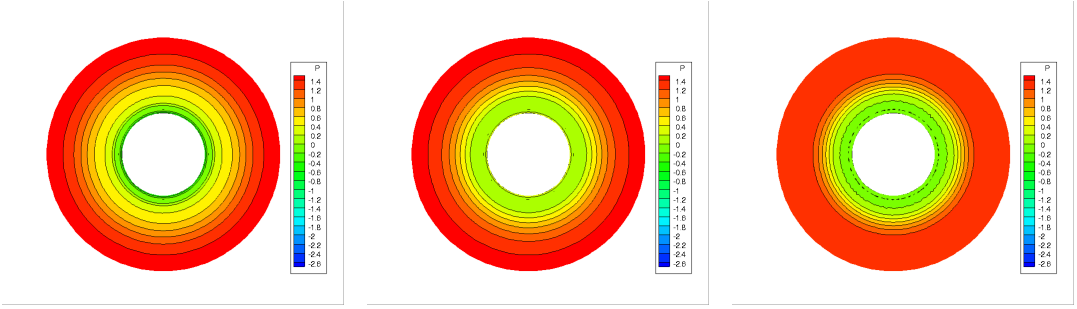


(d) Mode 4 (e) Mode 5 (f) Mode 6

Figure 3: The six first normalized spatial PGD modes for V

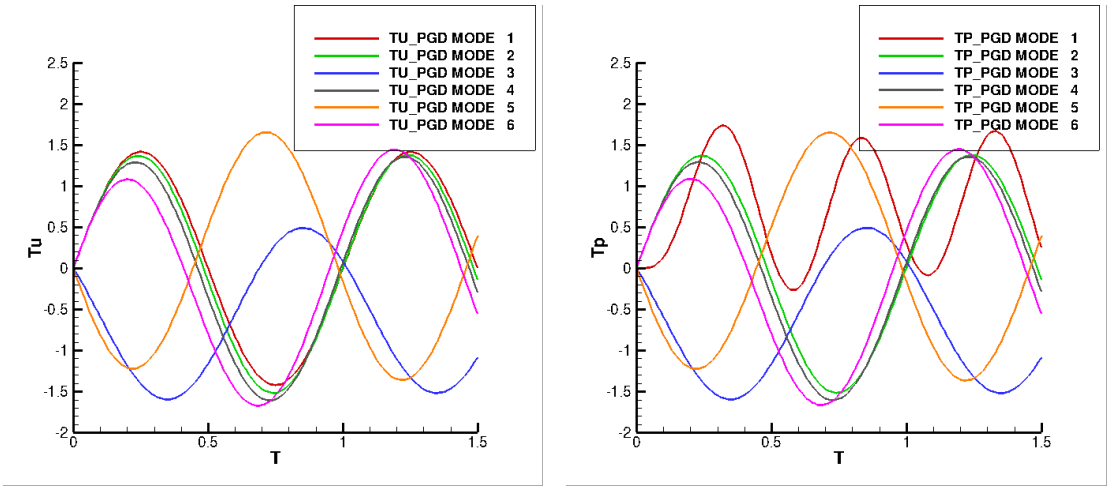


(a) Mode 1 (b) Mode 2 (c) Mode 3



(d) Mode 4 (e) Mode 5 (f) Mode 6

Figure 4: The six first normalized spatial PGD modes for P



(a) Temporal velocity modes

(b) Temporal pressure modes

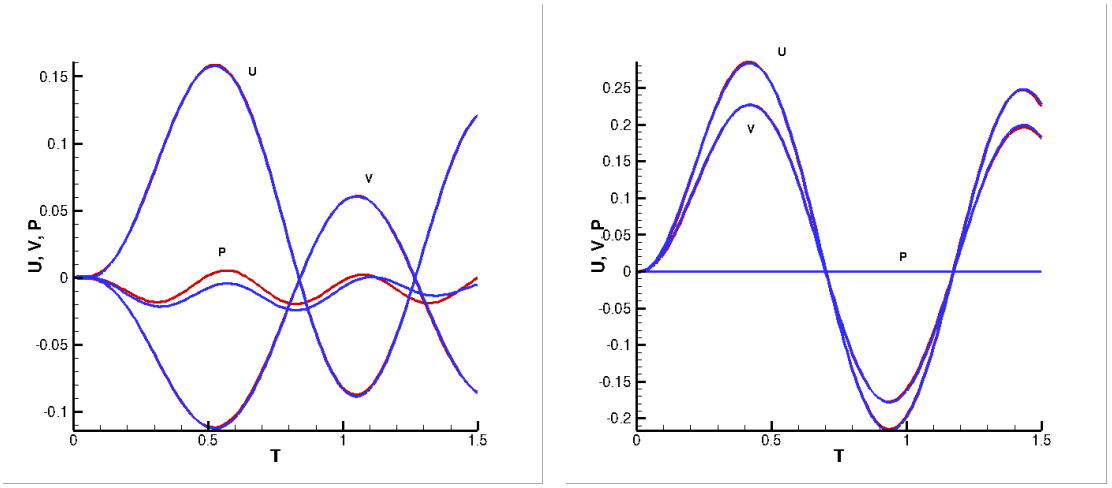
Figure 5: The six first normalized temporal PGD modes for the velocity and pressure

The six first normalized temporal modes for the velocity and pressure are also shown in the figures 5.

The figures 6 show a comparison of the temporal evolution of the velocity components and pressure at the two points mentioned before. 30 modes are used to build this PGD decomposition. We can notice a very good agreement on the velocity components and a less perfect agreement on the pressure, especially at points (here point A) which are located in a region where the convective effects are significant (and therefore $T_i^u \neq T_i^p$). It should be underlined here that the discretization errors committed in the discrete evaluation of the temporal PGD coefficients preclude any perfect agreement between the incremental and the modal approaches for a given time and space discretization. To draw conclusions on the solution convergence, it would be necessary to perform a systematic grid and time convergence study, which has not yet been carried out in this preliminary study.

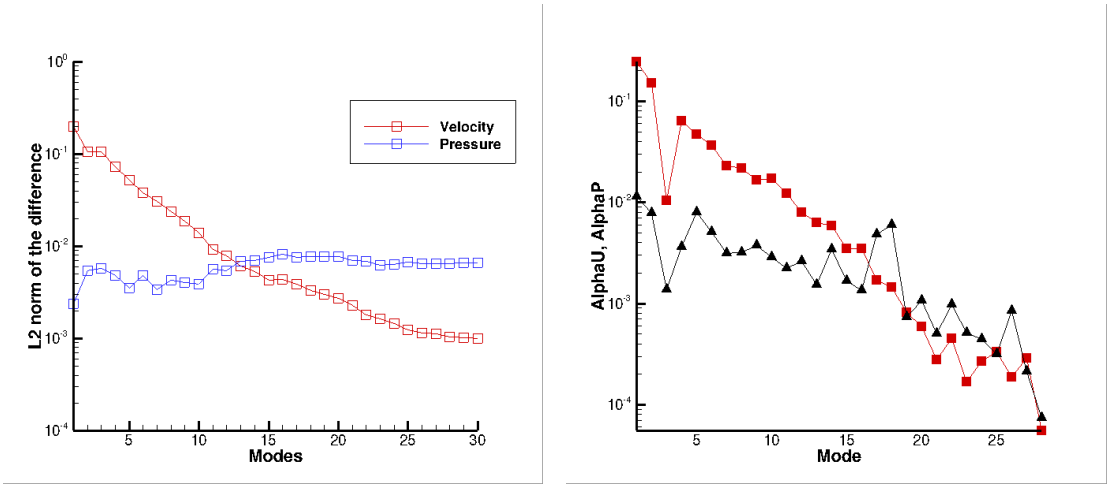
The figure 7(a) shows the L2 norm in space and time of the difference between the solutions computed with the incremental and modal approaches. It is clear that the difference on the velocity fields is reduced as long as the modal content is enriched before reaching a minimum value of 10^{-3} which is representative of the different discretization errors characterizing each methodology. However, such a convergence is not observed on the pressure field which is still affected by a systematic (but small) defect. It is surprising to observe that the convergence on the velocity is not affected by the lack of convergence on the pressure.

The figure 7(b) shows the evolution of the norms of the velocity (in red) and pressure (in black) PGD modes. The modal convergence is smoother on the velocity than on the pressure. We can notice that it is difficult to reach more than five orders of magnitude for the Navier-Stokes equations, although a zero-machine convergence is reached without any problem for the Stokes equations (not shown here for the sake of conciseness). It is however interesting to assess the quality of the PGD decomposition with a reduced number of modes. Figure 8 show the temporal signal at the same points A and B for only 15 modes. One can notice that the agreement on the velocity and pressure signals is already quite satisfactory.



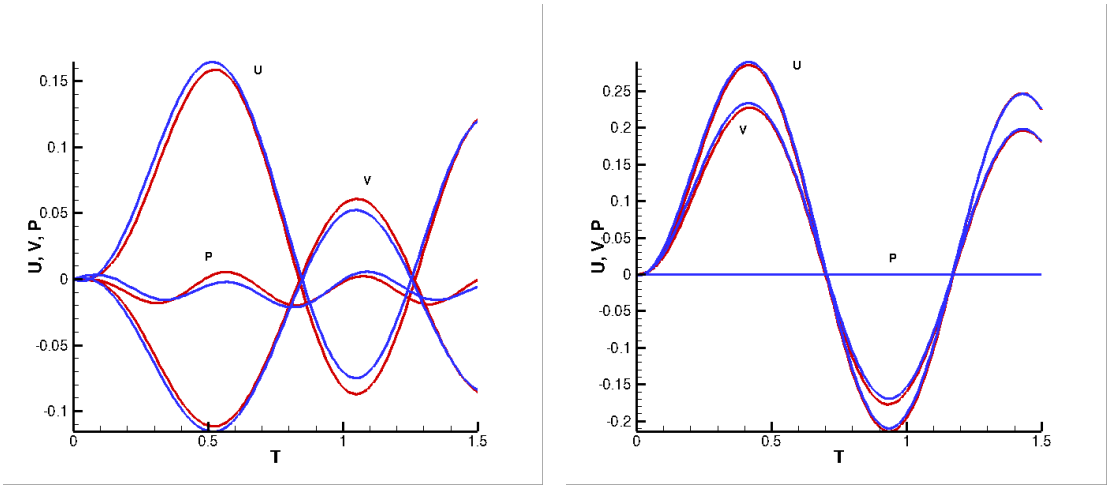
(a) Point A (b) Point B

Figure 6: Comparison between the incremental (red) and modal (blue) temporal signals for the two components of the velocity and the pressure (30 modes)



(a) L2 Norm of the difference (b) Modal convergence

Figure 7: Modal convergence of the velocity and pressure



(a) Point A (b) Point B

Figure 8: Comparison between the incremental (red) and modal (blue) temporal signals for the two components of the velocity and the pressure (15 modes)

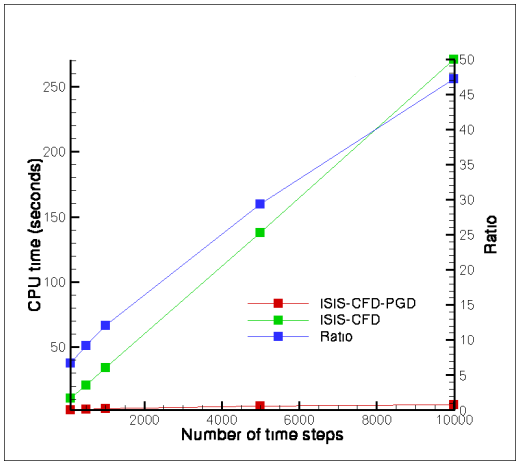


Figure 9: Evolution of the computational time to solve the unsteady Navier-Sokes equations - Comparison between incremental (ISIS-CFD) and modal (ISIS-CFD-PGD) approaches

Finally, the figure 9 shows a comparison of the evolution of computational time with the number of time steps, both for a classical incremental approach (ISIS-CFD) and for the present modal decomposition (ISIS-CFD-PGD). This comparison is made for a very modest grid and already illustrates the impressive speed-up which can be obtained with the modal paradigm.

10 CONCLUSIONS AND PERSPECTIVES

This article has briefly described a first formulation of a time space separated PGD decomposition applied to the unsteady Navier-Stokes equations. While the implementation of the PGD paradigm in the momentum equations is relatively straightforward for the velocity spatial and temporal modes, the determination of the temporal pressure modal decomposition poses a severe problem. This difficulty has to be related with the fact that the conservation of mass is a steady constraint linking velocity gradients. It is therefore

necessary to transform this solenoidality condition into a pressure equation. Moreover, contrary to what is done in traditional finite volume segregated computational approaches, the pressure equation which is used here is not a semi-discretized Poisson equation but a Laplace pressure equation obtained by taking the divergence of the momentum equations. The first results are encouraging although the robustness of the overall procedure should be improved. The lack of machine-zero modal convergence for such a simple flow configuration might be an issue for higher Reynolds numbers which should be understood and solved. The test case used to illustrate this first formulation is very academic but the formulation is written in such a way that it can be applied to arbitrary geometries on fully unstructured grids. It is worthwhile to notice that the PGD decomposition has been implemented into an industrial fully unstructured finite volume solver **ISIS-CFD** worldwide distributed by NUMECA Int. under the name FINE^{TM}/Marine. This means that it is feasible to introduce a PGD formulation into industrial softwares without dramatic modifications of the code as long as a time-space separation is envisaged.

REFERENCES

[1] P. Ladeveze, The large time increment method for the analyze of structures with nonlinear constitutive relation described by internal variables, *Comptes Rendus Académie des Sciences Paris*, **309**, 1095-1099, 1989.

[2] P. Ladeveze, J.-C. Passieux and D. Néron, The latin multiscale computational method and the proper generalized decomposition, *Computer Methods in Applied Mechanics and Engineering*, **199**, 1287-1296, 2010.

[3] F. Chinesta, P. Ladeveze and E. Cueto, A short review in model order reduction based on Proper Generalized Decomposition, *Archives of Cmoputational Methods in Engneering*, **18**, 395-404, 2011.

[4] F. Chinesta, R. Keunings and A. Leygue, The Proper Generalized Decomposition for adavanced numerical simulations. A primer, *Springerbriefs*, Springer, 2014.

[5] P. Queutey, M. Visonneau, An Interface Capturing Method for Free-Surface Hydrodynamic Flows. *Computers & Fluids*, **36**, 1481-1510, 2007.

[6] X. Vasseur, G.B. Deng, J.Piquet and M. Visonneau, A new fully coupled method for computing turbulent flows. *Computers & Fluids*, **30(4)**, 445-472, 2001.

[7] A. Dumon, Réduction dimensionnelle de type PGD pour la résolution des écoulements incompressibles, *PhD thesis*, Université de La Rochelle, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Crete Island, Greece, 5–10 June 2016

RATIONAL MODELING OF MULTIVARIATE MULTI-FIDELITY DATA

Elizabeth Rita Samuel[1],Dirk Deschrijver[1],Luc Knockaert[1],Tom Dhaene[1],Annie Cuyt[2]

[1]Ghent University - iMinds
Gaston Crommenlaan 8, 9050 Ghent, Belgium
e-mail: {dirk.deschrijver,luc.knockaert,tom.dhaene}@intec.ugent.be

[2] University of Antwerp
Middelheimlaan 1, 2020 Antwerp, Belgium
e-mail: {annie.cuyt}@ua.ac.be

Keywords: Rational functions, multi-fidelity data, modeling, barycentric interpolation.

Abstract. *Accurate multi-fidelity modeling is of high importance in the present day engineering design process. It allows to model computationally expensive simulations at a reduced cost by combining simulations with variable fidelity levels. In this paper, a novel algorithm is proposed to build multivariate models from variable fidelity simulations using rational functions. The modeling is based on high-fidelity data and low-fidelity data that is sampled over a parameter space of interest. The former is assumed to be computationally expensive and sparse, whereas the latter is cheaper to obtain but comes at a lower accuracy. It is shown that accurate rational models can be built at a reduced cost by combining these types of data. The effectiveness of the algorithm is applied to several examples and confirmed by numerical results.*

1 INTRODUCTION

Nowadays, there is a growing need for accurate behavioral models in many engineering applications and research disciplines. These models mimic the input-output behavior of a system, which is usually characterized by multiple parameters (also called variables) that describe the physical properties of the system over a certain parameter range of interest. The availability of such models has proven very useful for tasks like optimization, design space exploration, virtual prototyping and sensitivity analysis. In order to build models, a representative set of data samples must be collected by performing simulations. Unfortunately, as systems are becoming increasingly complex, the computational cost associated with simulations gets prohibitively high and limits the complexity of systems that can be handled.

The modeling of data with different levels of fidelity can be very useful in areas where the computational effort of an analysis is substantial, such as in computational fluid dynamics (CFD) and finite element (FE) analysis. The term "fidelity" refers to the rigorousness of the simulations and the level of detail by which an underlying system is represented. For example, depending on the mesh refinement of a system, physics-based simulations can be performed to collect data having various accuracy levels. On one hand, a dense mesh discretization leads to time-consuming simulations used to generate very accurate data, henceforth called high-fidelity (HF) data. On the other hand, a coarse mesh discretization significantly speeds up the simulations and generates data that captures the overall trend, although it is often less precise. This type of data is called low-fidelity (LF) data. It was shown in [1, 2, 3] that accurate models can be built at a reduced cost by combining data with different fidelity levels, hereby combining the best of both worlds.

A well known algorithm for multi-fidelity modeling is the auto-regressive co-kriging algorithm, which is a geostatistical technique that originates from mining and geology [4, 5]. It models the HF data exactly and accounts for spatial correlation in the LF data. In the case of linear time-invariant systems that are described by a set of differential equations, however, it is preferable to model the data with rational functions because this model type corresponds better to the underlying nature of the system [6]. The modeling of data using rational functions has been studied intensively in literature, and several algorithms are available. The reader is referred to [7, 8, 9, 10] for a survey and analysis of univariate methods. More details on algorithms for multivariate rational modeling are provided in [11, 12, 13, 14]. In these works, it is assumed that the data originates from a single source and that all available data samples have a comparable level of fidelity. This is one of the key aspects that will be addressed in this paper.

This paper proposes a novel algorithm that builds multivariate models from multi-fidelity data using rational functions. To this end, it makes use of the barycentric interpolation algorithm which has excellent numerical properties [15, 16]. First, a multivariate rational model that exactly interpolates the HF data is built using barycentric interpolation. Second, the remaining model coefficients are calculated such that the (calibrated) LF data is approximated in a least-squares (LS) sense. The algorithm is applied to model the quality factor and inductance of a spiral inductor. Numerical results confirms the effectiveness of the approach.

2 ALGORITHM OUTLINE

The goal of the algorithm is to identify a multivariate rational function $\mathcal{R}$ that approximates the behavioral response of an underlying system f. For notational convenience, it will be assumed that this system depends on 2 parameters (α and β), although the extension to multivari-

ate responses is analogous and conceptually straightforward.

$$\mathcal{R}(\alpha,\beta)=\frac{N(\alpha,\beta)}{D(\alpha,\beta)}=\frac{\sum_{i=0}^{I}\sum_{j=0}^{J}c_{ij}\alpha^i\beta^j}{\sum_{i=0}^{I}\sum_{j=0}^{J}\tilde{c}_{ij}\alpha^i\beta^j}. \tag{1}$$

To ensure that $\mathcal{R}$ matches f as closely as possible, some data is needed to estimate the unknown model coefficients c and $\tilde{c}$. To this end, the system f is simulated using different levels of fidelity, resulting in two disjoint sets of data

$$\text{HF} = \{(\alpha_k^H,\beta_k^H,f^H(\alpha_k^H,\beta_k^H))\}_{k=1}^{K_H} \tag{2}$$
$$\text{LF} = \{(\alpha_i^L,\beta_i^L,f^L(\alpha_i^L,\beta_i^L))\}_{i=1}^{K_L} \tag{3}$$

The HF data is sampled on a sparse grid of $K_H = M \times N$ points that are equidistantly spaced in the (α,β) parameter space. The LF data comprises K_L data samples that can be scattered throughout the parameter space. Taking into account the computational cost of simulations, it is assumed that the HF data is sparse when compared to the LF data ($K_H << K_L$). To ensure that both types of data are compatible, the ratio between HF and LF data is computed as a calibration factor to scale the LF data accordingly [17, 18, 19]. Then, a two-step modeling approach is proposed:

1. Compute a rational model that exactly interpolates all the HF data by using a barycentric interpolation formula.

2. Calculate the coefficients of the model in such a way that the (calibrated) LF data is approximated in a LS sense.

It is assumed that the HF data is too sparse in order to generate a global model with sufficient accuracy. However, by exploiting the information that is contained within the LF data, it is possible to control the behavior of model $\mathcal{R}$ inbetween the HF data samples. By combining both types of data, a model $\mathcal{R}$ with good overall accuracy is obtained.

3 RATIONAL MODELING ALGORITHM

In order to facilitate a mixture of interpolation and LS approximation, the multivariate rational function $\mathcal{R}$ in (1) is represented in a different functional form [20]. To fix the barycentric representation of the rational function [15, 16], the input values of the HF data are used as the interpolation points such that $\mathcal{R}(\alpha_k^H,\beta_k^H)=f^H(\alpha_k^H,\beta_k^H)$ for $k=1,...,K_H$ and for any choice of non-zero barycentric weights ($\boldsymbol{w}=[w_k]\in\mathbb{R}^{K_H\times 1}$ with $w_k\neq 0$).

$$\mathcal{R}(\alpha,\beta,\boldsymbol{w})=\frac{N(\alpha,\beta,\boldsymbol{w})}{D(\alpha,\beta,\boldsymbol{w})}=\frac{\sum_{k=1}^{K_H}\frac{w_k}{(\alpha-\alpha_k^H)(\beta-\beta_k^H)}f^H(\alpha_k^H,\beta_k^H)}{w_0+\sum_{k=1}^{K_H}\frac{w_k}{(\alpha-\alpha_k^H)(\beta-\beta_k^H)}}. \tag{4}$$

The choice of $\boldsymbol{w}$ in (4) allows one to enforce additional properties to the model. In this work, these additional degrees of freedom are exploited in such a way that the interpolation model also approximates the LF data in a LS sense. This corresponds to finding optimal value for

the weight $\boldsymbol{w}$ such that sum of the squared residuals is minimized. Note that a residual is the difference between the actual function value f and the model response.

$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{K_L} \left| \frac{N(\alpha_i^L, \beta_i^L, \boldsymbol{w})}{D(\alpha_i^L, \beta_i^L, \boldsymbol{w})} - f^L(\alpha_i^L, \beta_i^L) \right|^2 \tag{5}$$

Unfortunately, problem (5) is non-linear and therefore it becomes difficult to solve in a fast and accurate way. Instead, Levi proposes to solve a linear approximation of the problem [21].

$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{K_L} |N(\alpha_i^L, \beta_i^L, \boldsymbol{w}) - f^L(\alpha_i^L, \beta_i^L) D(\alpha_i^L, \beta_i^L, \boldsymbol{w})|^2 \tag{6}$$

This way, an estimate for $\boldsymbol{w}$ can be found by solving a linear set of LS equations (7) of the form $\boldsymbol{Ax} = \boldsymbol{b}$ using the matlab operator '\' (mldivide). Note that this problem is overdetermined because $K_L >> K_H$. One can the weight $w_0 = 1$ in (4) without loss of generality.

$$\boldsymbol{A} = [a_{ik}] \in \mathbb{R}^{K_L \times K_H} \quad \text{and} \quad \boldsymbol{x} = \boldsymbol{w} \in \mathbb{R}^{K_H \times 1} \quad \text{and} \quad \boldsymbol{b} = [b_i] \in \mathbb{R}^{K_L \times 1}$$

$$\text{where} \quad a_{ik} = \frac{f^H(\alpha_k^H, \beta_k^H) - f^L(\alpha_i^L, \beta_i^L)}{(\alpha_i^L - \alpha_k^H)(\beta_i^L - \beta_k^H)} \quad \text{and} \quad b_i = f^L(\alpha_i^L, \beta_i^L). \tag{7}$$

It is known that Levi's formulation is biased and not equivalent to (6), because a weighting factor $1/|D(\alpha_i^L, \beta_i^L, \boldsymbol{w})|^2$ is omitted in each term of the summation. To improve the accuracy of the model, the estimated weights w_k are refined iteratively by applying the Sanathanan-Koerner (SK) iteration [9] with explicit weighting as illustrated in [22].

$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{K_L} \left| \frac{N^t(\alpha_i^L, \beta_i^L, \boldsymbol{w}) - f^L(\alpha_i^L, \beta_i^L) D^t(\alpha_i^L, \beta_i^L, \boldsymbol{w})}{D^{t-1}(\alpha_i^L, \beta_i^L, \boldsymbol{w})} \right|^2. \tag{8}$$

This iteration multiplies each row in the LS matrix with an explicit weighting function $1/|D^{t-1}(\alpha, \beta, \boldsymbol{w})|^2$, which is based on Levi's initial (iteration step 0) or the previous (iteration step $t-1$) estimate of the denominator. This way, updated values of the weights can be found in successive steps $t = 0, ..., T$. The final model is then $\mathcal{R}^T(\alpha, \beta, \boldsymbol{w}) = N^T(\alpha, \beta, \boldsymbol{w})/D^T(\alpha, \beta, \boldsymbol{w})$. Although the solution of (8) is not guaranteed to converge to the solution of (5) either, the SK iteration is known to provide good results for sufficiently high signal-to-noise ratios [22].

4 MODEL COEFFICIENT RELAXATION

In (7) it was chosen to set $w_0 = 1$, because numerator and denominator can be divided by the same constant value without loss of generality. In some cases, this can lead to a poor convergence of the SK iteration. This problem was investigated in [23] and it was proposed to make w_0 a variable in the LS problem. This gives rise to a modified set of equations $\begin{bmatrix} \boldsymbol{A} & -\boldsymbol{b} \end{bmatrix} \boldsymbol{x} = 0$. To avoid the trivial null solution, one additional equation is added to the LS problem.

$$\sum_{i=1}^{K_L} \left(\sum_{k=1}^{K_H} \frac{w_k}{(\alpha_i^L - \alpha_k^H)(\beta_i^L - \beta_k^H)} + w_0 \right) = K_L. \tag{9}$$

Equation (9) is scaled in relation to the size of $f^L(\alpha^L, \beta^L)$ in the LS problem by a scaling factor

$$scale = \frac{||\boldsymbol{f}^L(\alpha^L, \beta^L)||_2}{K_L} \tag{10}$$

Note that this additional equation does not impose any further constraints on the LS problem other than preventing the trivial null solution. The reader is referred to [23] for a more details.

5 NUMERICAL RESULTS

In this example, the rational modeling of a three turn spiral inductor is considered as shown in Figure 1. The dielectric has a thickness of 300 μm, a relative dielectric constant $\epsilon_r = 9.6$ and a loss tangent tan δ = 0.0002. The conductivity of the metallic layers is equal to $\sigma = 5.8 \times 10^7$ S/m. The spacing between conductors is equal to 10 μm.

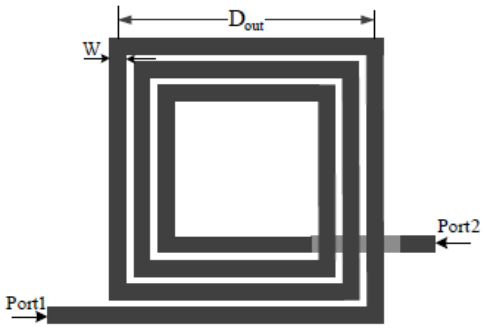


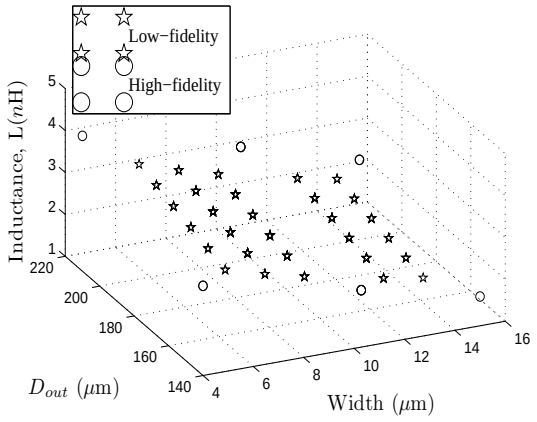
Figure 1: Layout of the spiral inductor (Top View).

The width of the conductors W and the outer length D_{out} are considered as the 2 input parameters (α and β) and their corresponding ranges are shown in Table 1. The rational modeling technique is used to model the inductance L and quality factor Q of the spiral inductor at frequency $freq = 2.4$ GHz. The L and Q are calculated from the simulated admittance parameters (Y-parameters) as in (11). The Y-parameters describe the electrical behavior of linear electrical networks where the terminal currents can be expressed in terms of the terminal voltages [24].

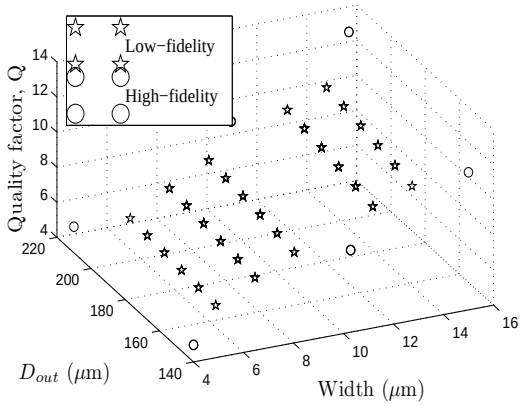
Table 1: Range of the design parameters.

Parameter	Width (W)	Outer length (D_{out})
Lower bound	4 μm	140 μm
Upper bound	15 μm	210 μm

$$\begin{aligned} L(W, D_{out}) &= \frac{1}{2\pi freq} \Im m \frac{1}{Y_{12}(freq, W, D_{out})} \\ Q(W, D_{out}) &= \frac{\Im m(Y_{11}(freq, W, D_{out}))}{\Re e(Y_{11}(freq, W, D_{out}))}. \end{aligned} \tag{11}$$



(a) High−fidelity and low−fidelity samples



(b) High−fidelity and low−fidelity samples

Figure 2: HF and LF data samples considered for the modeling using rational function.

The Y-parameters are obtained from the full-wave electromagnetic simulator Agilent Advanced Design System1 (ADS) Momentum [1], whereas the LF data are obtained using circuit schematic ADS simulations. The circuit schematic simulations are based on analytical formulas for the electrical behavior of the spiral inductor, whereas the electromagnetic simulations are based on the solution of Maxwell's equations, which is computationally more expensive. A set of LF and a set of HF data samples are collected on a grid of 5×6 (W, D_{out}) and 3×2 (W, D_{out}) respectively for the frequency 2.4 GHz to build the rational model. The computation time to get LF and HF admittance parameters at one sample point in the 2D (W, D_{out}) parameter space is roughly equal to 0.0301s and 0.2541s, respectively. Note that these timing results were performed on an Intel$^{(R)}$ Core$^{(TM)}$ 2 Duo P8700 2.53 GHz machine with 2 GB RAM and has been implemented in Matlab R2012b on the Windows 7 platform. For checking the accuracy of the model, relative error is used, which is defined as follows, where f represents either L or Q.

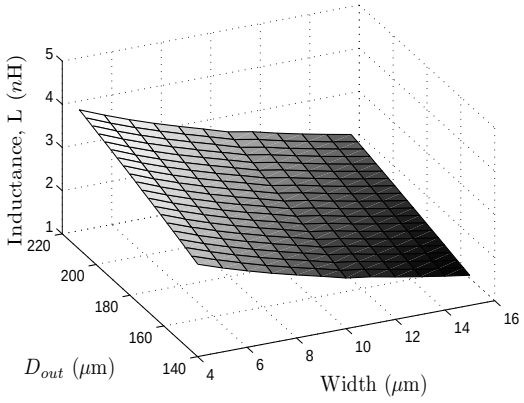
$$Rel.Err = \frac{|f(W, D_{out}) - \mathcal{R}(W, D_{out})|}{|f(W, D_{out})|}. \tag{12}$$

In both cases, 5×6 and 3×2 samples of LF and HF are considered as shown in Figure 2. Then using the proposed technique a rational model is obtained for the L and Q. Figure 3 (a) and (b) plots the L and Q respectively on a denser grid of 12×20 and is validated with HF data on a similar grid. The relative error (12) for L and Q is plotted in Figure 4 (a) and (b) respectively and can be seen that the model accurately captures the true underlying system. The evolution of the normalized root-mean-square error (NRMSE) is compared in terms of iteration count in Figure 5. The NRMSE for K_v validation points is defined as,

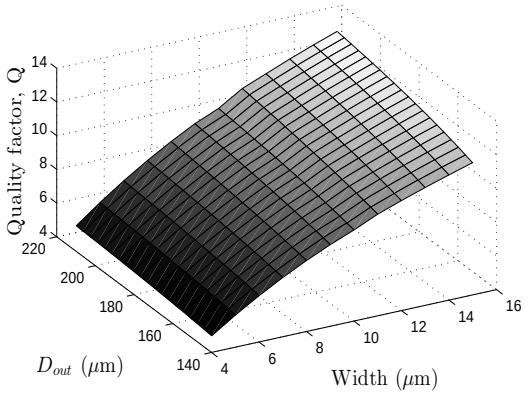
$$NRMSE = \frac{\sqrt{\frac{\Sigma_{k=1}^{K_v}(f(W_k, D_{outk}) - \mathcal{R}(W_k, D_{outk}))^2}{K_v}}}{max(f(W, D_{out})) - min(f(W, D_{out}))}. \tag{13}$$

To illustrate the correlation of the multi-fidelity model with the samples considered for modeling, the Spearman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) is studied for both cases. In Table 2, the multi-fidelity model is

[1] www.eesof.com, Agilent Technologies EEsof EDA, Santa Rosa, CA.



(a) Multi−fidelity model for Inductance



(b) Parametric rational model for Quality factor, Q.

Figure 3: (a) and (b) plots the model obtained for inductance (L) and quality factor (Q) using the rational function.

compared with the HF and LF training samples that were used for the modeling as shown in Figure 2. It can be seen that the samples and the model have a strong linear relationship as the SROCC and PLCC is 0.99. The model is also compared with the HF data used for validation and is seen that the SROCC and PLCC is equal to 0.99, which again infers that the model response is strongly correlated with the true L and Q values, see Figure 6. Table 2 also gives the values of the NRMSE (13) to assess the multi-fidelity model accuracy for both L and Q.

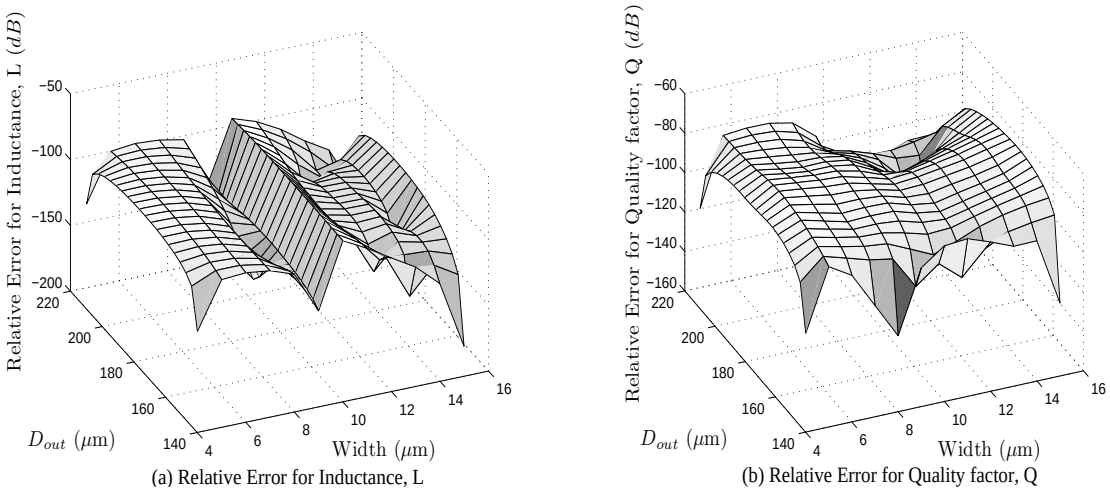
Table 2: CORRELATION OF MULTI-FIDELITY MODEL WITH THE ACTUAL SAMPLES.

Data set	For Inductance (L)			For Quality Factor (Q)		
	NRMSE	SROCC	PLCC	NRMSE	SROCC	PLCC
High-fidelity training	0	1	1	0	1	1
Low-fidelity training	0.092	0.99	0.99	0.08	0.99	0.99
Validation (high-fidelity)	0.012	0.99	0.99	0.003	0.99	0.99

Thus, the proposed technique generates a rational model from the LF and HF data which is more accurate than a model built using either the sparse HF samples or dense LF samples only. Also the model is built at a reduced cost when compared to a model that is built from a larger amount of HF samples only (see Table 3), leading to a speedup of 21 times.

6 CONCLUSIONS

A novel approach for multivariate modeling of systems using a combination of variable fidelity simulations has been proposed. The algorithm first makes use of multidimensional barycentric interpolation for computing a rational model with the HF data, and then the coefficients of this model are iteratively calculated using the (calibrated) LF data in least-square sense. It is shown that rational function models can be built at a reduced cost by combining different types of data. The effectiveness of the technique is illustrated using a numerical result.



(a) Relative Error for Inductance, L

(b) Relative Error for Quality factor, Q

Figure 4: (a) and (b) plots the relative error of the model with the original model which is the high-fidelity data.

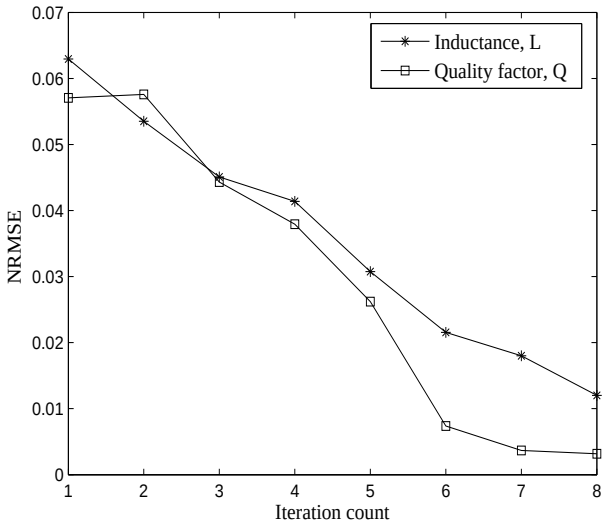


Figure 5: Evolution of NRMSE in terms of iteration count t.

ACKNOWLEDGMENTS

This research has been funded by the Fund for Scientific Research in Flanders (FWO-Vlaanderen) and the Interuniversity Attraction Poles Programme BESTCOM initiated by the Belgian Science Policy Office.

REFERENCES

[1] P. Craig, M. Goldstein, A. Seheult, and J. Smith, "Constructing partial prior specifications for models of complex physical systems.," *Journal of the Royal Statistical Society Series D (The Statistician)*, vol. 47, no. 1, pp. 37–53, 1998.

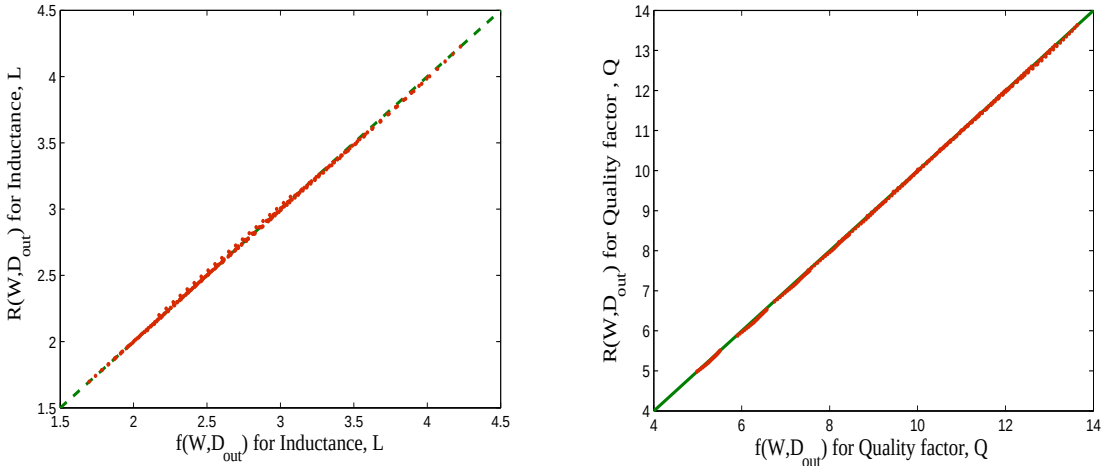


Figure 6: Plots the observed value to the predicted value for the inductance (L) and quality factor (Q).

Table 3: COMPUTATION TIMES.

Data set	samples	CPU time (s)
Calculation of LF training samples	5×6	0.903
Calculation of HF training samples	3×2	1.525
Rational function modeling		0.267
Evaluation of multi-fidelity model over a dense grid	12×20	0.209
	Total	2.904
	Speedup	21
Calculation of HF data over a dense grid	12×20	60.984

[2] M. Bakr, J. Bandler, K. Madsen, and J. Søndergaard, "An Introduction to the Space Mapping Technique," *Optimization and Engineering*, vol. 2, no. 4, pp. 369–384, 2001.

[3] D. Huang, T. Allen, W. Notz, and R. Miller, "Sequential kriging optimization using multiple-fidelity evaluations," *Structural and Multidisciplinary Optimization*, vol. 32, no. 5, pp. 369–382, 2006.

[4] M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, pp. 1–13, Mar. 2000.

[5] I. Couckuyt, F. Declercq, T. Dhaene, H. Rogier, and L. Knockaert, "Surrogate-based infill optimization applied to electromagnetic problems," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 20, no. 5, pp. 492–501, 2010.

[6] J. Willems and J. Polderman, *Introduction to Mathematical Systems Theory: A Behavioral Approach*. NewYork: Springer-Verlag, 1998.

[7] E. Safe and R. S. Varga, *Padé and Rational Approximation: Theory and Applications*. NewYork: Academic Press Inc., 1977.

[8] Y.-l. F. Chiang, "A modified remes algorithm," *SIAM J. Sci. Stat. Comput.*, vol. 9, pp. 1058–1072, Nov. 1988.

[9] C. K. Sanathanan and J. Koerner, "Transfer function synthesis as a ratio of two complex polynomials," *IEEE Transactions on Automatic Control*, vol. 8, no. 1, pp. 56–58, 1963.

[10] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *IEEE Transaction on Power Delivery*, vol. 14, pp. 1052–1061, Jul. 1999.

[11] R. Lehmensiek and P. Meyer, "Creating accurate multivariate rational interpolation models of microwave circuits by using efficient adaptive sampling to minimize the number of computational electromagnetic analyses," *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, pp. 1419–1430, Aug 2001.

[12] A. Lamecki, P. Kozakowski, and M. Mrozowski, "Efficient implementation of the cauchy method for automated cad-model construction," *IEEE Microwave and Wireless Components Letters*, vol. 13, pp. 268–270, July 2003.

[13] D. Deschrijver, T. Dhaene, and D. De Zutter, "Robust parametric macromodeling using multivariate orthonormal vector fitting," *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, no. 7, pp. 1661–1667, 2008.

[14] A. Lamecki, L. Balewski, and M. Mrozowski, "Adaptive cad-model construction schemes," *IEEE Transactions on Magnetics*, vol. 45, pp. 1538–1541, March 2009.

[15] C. Schneider and W. Werner, "Some new aspects of rational interpolation," *Mathematics of Computation*, vol. 47, no. 175, pp. pp. 285–299, 1986.

[16] L. Knockaert, "A simple and accurate algorithm for barycentric rational interpolation," *IEEE Signal Processing Letters*, vol. 15, pp. 154–157, 2008.

[17] V. V. Toropov, F. V. Keulen, V. L. Markine, and L. F. Alvarez, "Multipoint approximation based on response surface fitting: a summary of recent developments," *First ASMKO UK/ ISSMO.*, pp. 371–380, 1999.

[18] R. T. Haftka, "Combining global and local approximations.," *AIAA Journal*, vol. 29, no. 9, pp. 1523–1525, 1991.

[19] K. J. Chang, R. T. Haftka, G. L. Giles, and P. J. Kao, "Sensitivity-based scaling for approximating structural response.," *Journal of Aircraft*, vol. 30, no. 2, pp. 283–287, 1993.

[20] D. Deschrijver and T. Dhaene, "Stability and passivity enforcement of parametric macromodels in time and frequency domain," *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, pp. 2435–2441, Nov 2008.

[21] E. C. Levy, "Complex-curve fitting," *IRE Transactions on Automatic Control*, vol. AC-4, pp. 37–43, May 1959.

[22] D. Deschrijver, B. Gustavsen, and T. Dhaene, "Advancements in iterative methods for rational approximation in the frequency domain," *IEEE Transactions on Power Delivery*, vol. 22, no. 3, pp. 1633–1642, 2007.

[23] B. Gustavsen, "Improving the pole relocating properties of vector fitting," *Power Delivery, IEEE Transactions on*, vol. 21, no. 3, pp. 1587–1592, 2006.

[24] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach*. NewYork: Dover Publications, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ON THE USE OF S-PARAMETER TRANSFORMATIONS TO IMPROVE SURROGATE MODEL BEHAVIOUR OF MULTIPORT NETWORKS

Petrie Meyer

Dept. Electrical and Electronic Engineering, Stellenbosch University
Private Bag X1, Matieland, 7602, South Africa
e-mail: pmeyer@sun.ac.za

Keywords: Surrogate Models, Multiport, S-Paramters.

Abstract. *This paper will show how the utilization of the Generalized Multi-Mode Scattering Parameter Transformation for multiport S-parameters can result in the reduction of the number of S-parameters needed to model a multiport network, and how correct choices can result in S-parameters which exhibit improved numerical behaviour in a given parameter space. An example of a two-port, six-line transmission line with the conductors spaced as pairs is used to show the difference between sets of common-differential mode excitations, and sets of single-ended excitations will be shown.*

1 INTRODUCTION

Multiport networks pose particular problems to surrogate modelling algorithms, as an N-port network is typically described by an NxN Scattering matrix, or N^2 frequency and parameter dependent variables. In general, the parameter set can be reduced using symmetry, reciprocity, and if applicable, the passive lossless condition, but the number of separate models required always grow quite rapidly with N.

To exploit symmetry to obtain a reduced S-parameter set, the network has to be excited by correct combinations of port excitations which enforce the symmetry conditions. In addition, different sets of excitations often result in sets of S-parameters which exhibit varying behaviour as a function of the parameter space, and a specific excitation can result in very well-behaved S-parameters, while a different excitation can result in (for instance) resonant spikes.

To find these more optimum sets of excitations is quite difficult, as the full response of the network has to be simulated for each full set of excitations. While commercial EM-solvers can all calculate combined responses from any full set of excitations, they can do this only for uncoupled ports. For coupled ports, e.g. in a multi-conductor transmission line, each different set of excitations requires a new full 3D EM analysis, which quickly becomes prohibitively expensive.

Recently, the author developed a transformation technique which allows for the transformation of multiport S-parameters between different sets of excitations, using only one set of EM-analyses [1, 2]. The technique works for coupled and uncoupled ports, and is very simple to implement. It has been implemented on multiport antennas and transmission lines. This paper will show how the utilization of the transformation technique on multiport S-parameters can result in the reduction of the number of S-parameters needed to model a multiport network, and how correct choices can result in S-parameters which exhibit improved numerical behaviour in a given parameter space. An example of a six-line transmission line with the conductors spaced as pairs will be used to show difference between sets of common-differential mode excitations, and sets of single-ended excitations.

2 S-PARAMETER TRANSFORMATIONS

The basis of equivalent port representations is the creation of new sets of port voltages and currents which consist of linear combinations of the port voltages and currents of another port description. A generalized procedure to do this was presented in [1], together with a number of examples showing the implementation for classical circuit and antenna problems. Here, a brief mathematical overview is given, and the reader is referred to [1] for details.

Consider an N-port network excited in two different ways to form two equivalent N-port networks A and B, as shown in Fig. 1. Each network is described by an S-matrix $\mathbf{S}^{\mathrm{A,B}}$, with port voltages $\left(V_{\mathrm{n}}^{\mathrm{A,B}}\right)$, port currents $\left(I_{\mathrm{n}}^{\mathrm{A,B}}\right)$, and port impedances $\left(Z_{\mathrm{n}}^{\mathrm{A,B}}\right)$ indicated for both networks. In addition, the incident $\left(a_{\mathrm{n}}^{\mathrm{A,B}}\right)$ and reflected $\left(b_{\mathrm{n}}^{\mathrm{A,B}}\right)$ waves are also shown at each port. For the case where the port voltages and port currents of network B are linear combinations of the voltages and currents of network A, the technique establishes a transformation from $\mathbf{S}^{\mathrm{A}}$ to $\mathbf{S}^{\mathrm{B}}$ for an arbitrary set of port impedances (e.g. where network A consists of N single-ended ports, each terminated by the same reference impedance, and network B of sets of differential ports and possible single-ended ports, each terminated by a different impedance).

For the completely general case, the port voltages and currents of network B can be expressed

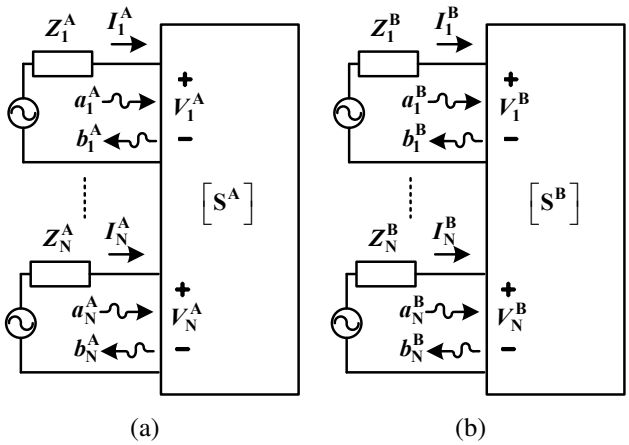


Figure 1: N-port networks with ports using (a) mode set A and (b) mode set B

as linear combinations of any or all of the port voltages and currents of network A. That is,

$$\begin{aligned} V_{\mathrm{n}}^{\mathrm{B}} &= k_{\mathrm{n1}}^{\mathrm{v}} V_1^{\mathrm{A}} + k_{\mathrm{n2}}^{\mathrm{v}} V_2^{\mathrm{A}} + \cdots + k_{\mathrm{nN}}^{\mathrm{v}} V_{\mathrm{N}}^{\mathrm{A}} \\ I_{\mathrm{n}}^{\mathrm{B}} &= k_{\mathrm{n1}}^{\mathrm{i}} I_1^{\mathrm{A}} + k_{\mathrm{n2}}^{\mathrm{i}} I_2^{\mathrm{A}} + \cdots + k_{\mathrm{nN}}^{\mathrm{i}} I_{\mathrm{N}}^{\mathrm{A}} \\ &\text{for} \quad \mathrm{n} = 1, ..., N \end{aligned} \tag{1}$$

or

$$\mathbf{V}^{\mathrm{B}} = \mathbf{K}^{\mathrm{v}} \mathbf{V}^{\mathrm{A}} \tag{2a}$$

$$\mathbf{I}^{\mathrm{B}} = \mathbf{K}^{\mathrm{i}} \mathbf{I}^{\mathrm{A}} \tag{2b}$$

where $\mathbf{V}^{\mathrm{A,B}}$ and $\mathbf{I}^{\mathrm{A,B}}$ denote $[N \times 1]$ matrices containing the port voltages and currents of networks A and B, and $\mathbf{K}^{\mathrm{v,i}}$ are $[N \times N]$ matrices as in (3) and (4).

$$\mathbf{K}^{\mathrm{v}} = \begin{bmatrix} k_{11}^{\mathrm{v}} & k_{12}^{\mathrm{v}} & \cdots & k_{1\mathrm{N}}^{\mathrm{v}} \\ \vdots & \vdots & \ddots & \vdots \\ k_{\mathrm{N1}}^{\mathrm{v}} & k_{\mathrm{N2}}^{\mathrm{v}} & \cdots & k_{\mathrm{NN}}^{\mathrm{v}} \end{bmatrix} \tag{3}$$

$$\mathbf{K}^{\mathrm{i}} = \begin{bmatrix} k_{11}^{\mathrm{i}} & k_{12}^{\mathrm{i}} & \cdots & k_{1\mathrm{N}}^{\mathrm{i}} \\ \vdots & \vdots & \ddots & \vdots \\ k_{\mathrm{N1}}^{\mathrm{i}} & k_{\mathrm{N2}}^{\mathrm{i}} & \cdots & k_{\mathrm{NN}}^{\mathrm{i}} \end{bmatrix} \tag{4}$$

From [1], $\mathbf{K}^{\mathrm{v}}$ and $\mathbf{K}^{\mathrm{i}}$ must be related by (5), with $\mathcal{I}_{\mathrm{N}}$ denoting an $[N \times N]$ identity matrix, for conservation of power under the transformation to be guaranteed.

$$\mathbf{K}^{\mathrm{v}\dagger} \mathbf{K}^{\mathrm{i}} = \mathcal{I}_{\mathrm{N}} \tag{5}$$

The incident and reflected waves at each port of both networks A and B can be related to the port voltage and current at that port using the classical power-wave definitions for real port impedances as

$$\mathbf{V}^{\mathrm{A,B}} = \left(\mathbf{Z}^{\mathrm{A,B}}\right)^{\frac{1}{2}} \left(\mathbf{a}^{\mathrm{A,B}} + \mathbf{b}^{\mathrm{A,B}}\right) \tag{6a}$$

$$\mathbf{I}^{\mathrm{A,B}} = \left(\mathbf{Z}^{\mathrm{A,B}}\right)^{-\frac{1}{2}} \left(\mathbf{a}^{\mathrm{A,B}} - \mathbf{b}^{\mathrm{A,B}}\right) \tag{6b}$$

where $\{\mathbf{a}^{\mathrm{A,B}}, \mathbf{b}^{\mathrm{A,B}}\}$ denote $[N \times 1]$ matrices containing the complex magnitudes of the incident and reflected waves respectively, and $\mathbf{Z}^{\mathrm{A,B}}$ are diagonal matrices containing the characteristic port impedances as shown in (7).

$$\mathbf{Z}^{\mathrm{A,B}} = \begin{bmatrix} Z_1^{\mathrm{A,B}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z_{\mathrm{N}}^{\mathrm{A,B}} \end{bmatrix} \tag{7}$$

Defining the matrices

$$\mathbf{M}_{\mathrm{S}} = \frac{1}{2}\left(\mathbf{Z}^{\mathrm{B}}\right)^{-\frac{1}{2}} \mathbf{K}^{\mathrm{v}} \left(\mathbf{Z}^{\mathrm{A}}\right)^{\frac{1}{2}} + \frac{1}{2}\left(\mathbf{Z}^{\mathrm{B}}\right)^{\frac{1}{2}} \mathbf{K}^{\mathrm{i}} \left(\mathbf{Z}^{\mathrm{A}}\right)^{-\frac{1}{2}} \tag{8a}$$

$$\mathbf{M}_{\mathrm{C}} = \frac{1}{2}\left(\mathbf{Z}^{\mathrm{B}}\right)^{-\frac{1}{2}} \mathbf{K}^{\mathrm{v}} \left(\mathbf{Z}^{\mathrm{A}}\right)^{\frac{1}{2}} - \frac{1}{2}\left(\mathbf{Z}^{\mathrm{B}}\right)^{\frac{1}{2}} \mathbf{K}^{\mathrm{i}} \left(\mathbf{Z}^{\mathrm{A}}\right)^{-\frac{1}{2}} \tag{8b}$$

the incident and reflected waves of network B can be expressed in terms of those of network A by (9a) and (9b).

$$\mathbf{a}^{\mathrm{B}} = \mathbf{M}_{\mathrm{S}}\mathbf{a}^{\mathrm{A}} + \mathbf{M}_{\mathrm{C}}\mathbf{b}^{\mathrm{A}} \tag{9a}$$

$$\mathbf{b}^{\mathrm{B}} = \mathbf{M}_{\mathrm{C}}\mathbf{a}^{\mathrm{A}} + \mathbf{M}_{\mathrm{S}}\mathbf{b}^{\mathrm{A}} \tag{9b}$$

Finally, with $\mathbf{b}^{\mathrm{A}} = \mathbf{S}^{\mathrm{A}}\mathbf{a}^{\mathrm{A}}$ and $\mathbf{b}^{\mathrm{B}} = \mathbf{S}^{\mathrm{B}}\mathbf{a}^{\mathrm{B}}$, it follows from (9a) and (9b) that $\mathbf{S}^{\mathrm{B}}$ can be expressed in terms of $\mathbf{S}^{\mathrm{A}}$ by (10).

$$\mathbf{S}^{\mathrm{B}} = \left(\mathbf{M}_{\mathrm{C}} + \mathbf{M}_{\mathrm{S}}\mathbf{S}^{\mathrm{A}}\right)\left(\mathbf{M}_{\mathrm{S}} + \mathbf{M}_{\mathrm{C}}\mathbf{S}^{\mathrm{A}}\right)^{-1} \tag{10}$$

A very important special case occurs when the port impedances of network B are related to the port impedances of network A by (11).

$$\mathbf{Z}^{\mathrm{B}} = \mathbf{K}^{\mathrm{v}}\,\mathbf{Z}^{\mathrm{A}}\,\left(\mathbf{K}^{\mathrm{i}}\right)^{-1} \tag{11}$$

In this case, $\mathbf{M}_{\mathrm{C}} = \mathbf{0}$, and the transformation in (10) simply becomes that shown in (12).

$$\mathbf{S}^{\mathrm{B}} = \mathbf{M}_{\mathrm{S}}\mathbf{S}^{\mathrm{A}}\left(\mathbf{M}_{\mathrm{S}}\right)^{-1} \tag{12}$$

where

$$\mathbf{M}_{\mathrm{S}} = \left(\mathbf{Z}^{\mathrm{B}}\right)^{-\frac{1}{2}} \mathbf{K}^{\mathrm{v}} \left(\mathbf{Z}^{\mathrm{A}}\right)^{\frac{1}{2}} \tag{13}$$

Note that this special case is the one most widely used for the transformation of single-ended S-parameters to mixed-mode S-parameters.

3 SIX-CONDUCTOR TRANSMISSION LINE

The application of the transformation to reduce model complexity can be illustrated by the use of multi-line transmission lines, such as the six-line transmission line shown in Fig. 2. The lines are arranged as three pairs, positioned at a fixed radius from the centre. A dielectric half-cylinder is inserted in the centre of the line to provide a discontinuity. For this example, the line is 50mm in length, with an outer radius of 5mm. Each conductor has a radius of 0.5mm, and is positioned at a radius of 3mm from the centre, with each pair at a 120 degree angle with respect to the other pairs. The dielectric disc has a permittivity of 20, and is 1mm in thickness.

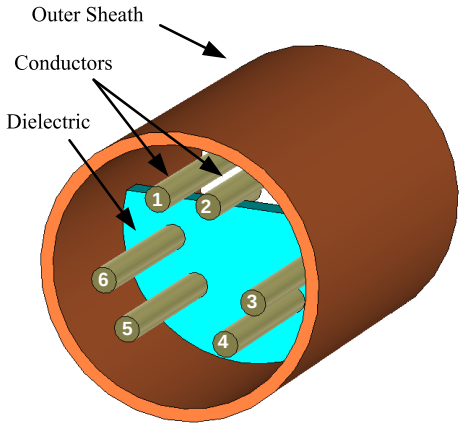


Figure 2: Six-conductor transmission line

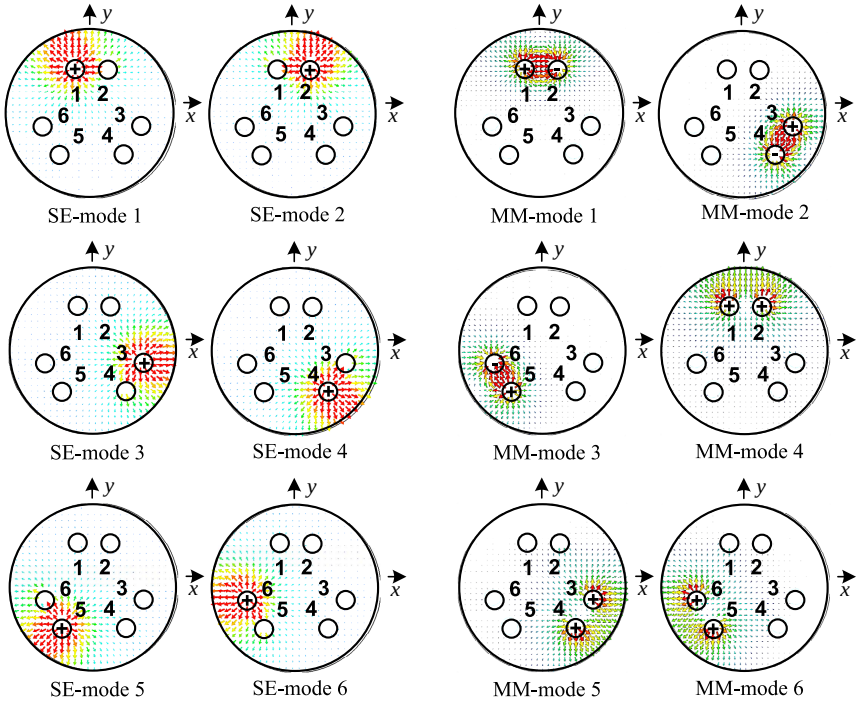
An infinite number of sets of S-parameters can be calculated for this structure. For the purposes of this paper, only two will be shown. Firstly, each conductor can be excited at each of its ports, with all the others terminated in 50 ohm loads. Such an analysis can readily be performed using a single-ended multipin waveguide feed in CST Microwave Studio, with a fast two-dimensional analysis to solve the port impedance for each mode, and a full three-dimensional analysis for the S-matrix. The six port modes established in this way at each port will be denoted as *SE-modes*. The electric field patterns for this set are shown in Fig. 3(a). For six lines at each of two ports, this results in a 12x12 complex S-matrix. For simplicity, only a dependence on frequency will be used here. To model this problem using surrogate models, 144 one-variable models are therefore required. While this can be reduced substantially using reciprocity and symmetry in the longitudinal direction, the dielectric disc reduces the number of symmetries in the transverse plane.

Using the generalized multi-mode S-parameter transformation, various other sets of S-parameters can be computed from the single-ended set, without any additional full-wave analyses. A typical set for this structure is the one obtained by exciting each set of two conductors with both common-mode and differential-mode signals at each port. These excitations as well as their respective electric field distributions are shown in Fig. 3(b) and denoted as *MM-modes*. To calculate the MM S-matrix, any set of port impedances can in principle be used. Here, the port impedance calculated for each MM-mode by a fast two-dimensional analysis in CST is used. For this analysis, each mode of the set is excited using a multi-pin feed with the indicated polarity.

From the two-dimensional CST analysis, and with the single-ended case represented by network A (denoted *SE*) and the multi-mode case by network B (denoted *MM*), the impedance matrices necessary for the transformation in (10) are shown in (14).

$$\begin{aligned} \mathbf{Z}^{\mathrm{SE}} &= diag\,(50) \\ \mathbf{Z}^{\mathrm{MM}} &= diag\,(138, 138, 138, 138, 138, 138, 70, 70, 70, 70, 70, 70) \end{aligned} \tag{14}$$

From standard definitions for differential and common-mode voltages and currents, the matrices



(a) SE port modes (b) MM port modes

Figure 3: Electric field distributions for port modes

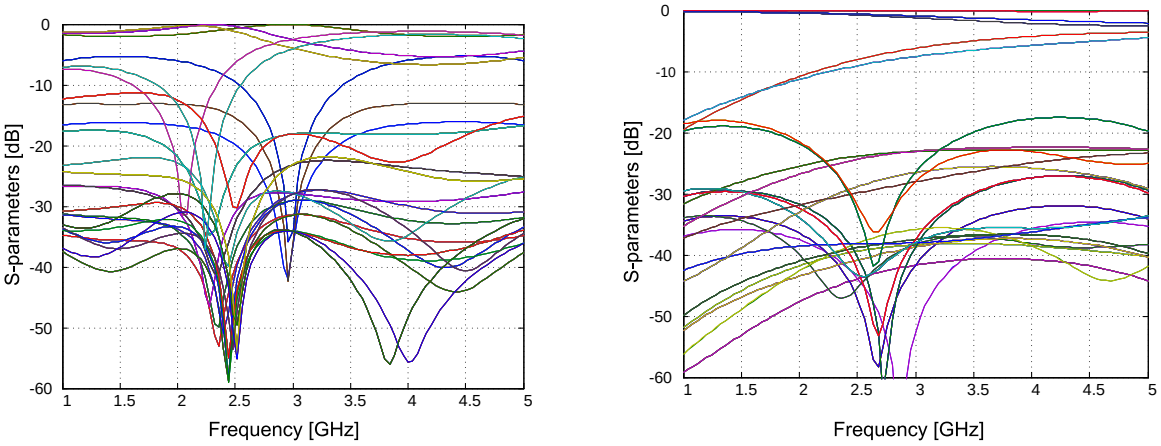
$\mathbf{K}^{\mathrm{v,i}}$ can be constructed as in (15)

$$\mathbf{K}^{\mathrm{v}} = \begin{bmatrix} \mathbf{K}^{\mathrm{p}} \\ \frac{1}{2}|\mathbf{K}^{\mathrm{p}}| \end{bmatrix} \qquad \mathbf{K}^{\mathrm{p}} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{15}$$

with $\mathbf{K}^{\mathrm{i}} = \left(\mathbf{K}^{\mathrm{v}\dagger}\right)^{-1}$ from (5).

The magnitudes of the two sets of 144 S-parameters are shown in Fig. 4. It is clear that in the multi-mode set, a number of S-parameters are small enough to be neglected (¡-60dB across the range), resulting in fewer parameters which need to be modelled. In addition, the non-zero S-parameters frequently occur in identical sets, and in general show less variation as functions of frequency than the single-ended set. This requires lower order surrogate functions for a similar modelling accuracy. Taken in combination, the change in the choice of excitation functions will result in a significant simplification of the required multiport surrogate model.

The given multi-mode set is of course only one example, and in general, each problem can be analyzed in order to find the optimum transformation matrix which will result in the smallest S-parameter set which can be modelled with the lowest order functions. The transformation technique has very few limitations mathematically, and can be included into most surrogate algorithms very easily as a data pre-conditioning step.



(a) Single-ended (b) Multi-mode

Figure 4: Single-ended and multi-mode S-parameters

4 CONCLUSION

This paper shows how the generalized multi-mode scattering parameter transformation can be used to reduce the required number and order of modelling functions required to create an accurate surrogate model. A twelve-port, six-line transmission line problem is used here as an example, but the technique is not limited to specific numbers or lines, or any specific sets of excitations.

ACKNOWLEDGEMENT

The authors thank CST for providing software and support.

REFERENCES

[1] P. Meyer, D.S. Prinsloo, Generalized Multi-Mode Scattering Parameter and Antenna Far-Field Conversions. *IEEE Trans. Antennas and Propagation*, **63**, 4815-4825, 2015.

[2] P. Meyer, Multi-Conductor Transmission Line Analysis Using the Generalized Multi-Mode S-parameter Transformation. *19th IEEE Workshop on Signal and Power Integrity (SPI 2015)*, Berlin, Germany, March 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

RECENT ADVANCES IN SURROGATE MODELING OF REFLECTOR ANTENNA SYSTEMS

Dirk I. L. de Villiers[1]

Department of Electrical and Electronic Engineering
Stellenbosch University, Stellenbosch, South Africa
e-mail: ddv@sun.ac.za

Keywords: Reflector Antennas, Surrogate Modeling.

Abstract. *This paper discusses some of the recent advances in the surrogate based modeling and optimization of reflector antenna systems, and presents several examples. Much of the focus is on the design of reflector systems for radio astronomy applications, where especially tight specifications are placed on some of the important performance metrics. In several cases the response surfaces of these performance metrics are computationally expensive to compute, and, as an added complication, they are often conflicting and need to be optimized in a multi-objective (MO) or Pareto framework. The large numbers of function evaluations required in these optimization loops make direct full wave simulation in the loop intractable, and surrogate based methods may in these cases lead to reliable global MO optimizations. Two main characteristics of reflector antenna responses of interest are exploited, namely their slow (or periodic) variation with design parameters, as well as the availability of a computationally cheap geometric optics approximation.*

1 INTRODUCTION

Reflector antenna systems are widely used in personal, commercial, scientific, and military applications, with ever increasing demands on their performance. Due to their typically large electrical size, the design of such systems traditionally relied heavily on approximated electromagnetic (EM) methods - most commonly Geometric Optics (GO) and Physical Optics (PO). These methods tend to display increasing accuracy with electrical size of the analyzed system, with the computational time normally scaling as N. An obvious drawback of using these methods is their inability to effectively and accurately model some of the (often important) finer details of the structures, including feed/dish interactions, strut and supporting structure influence, and large curvatures in the dish surfaces. With the enormous increase in availability of powerful personal and cluster computing facilities, as well as commercial computational electromagnetic (CEM) codes capable of solving a wide range of EM problems using full wave methods, it has become tractable in recent times to, if not design, at least simulate the performance of reflector systems much more accurately. An obvious drawback of full wave methods is the often prohibitively slow simulation times when it is to be used in an optimization loop.

Given this framework of the availability of well-developed fast approximate methods, as well as accurate full wave solutions, the use of surrogate based optimization (SBO) schemes for the design of reflector antennas is an obvious extension to the classical design methods. In this context SBO refers to a variety of methods used to shift the optimization task onto some surrogate model which is based on the fast approximated model (and is therefore fast to evaluate), but aligned in some way to the slow but accurate model (and is therefore more accurate than the approximate model – even if just in a specific region of interest in the parameter space).

This paper will discuss some of the recent advances in this field and provide examples of how SBO has been used to develop design methodologies for advanced reflector antenna systems. Much of the focus will be on the design of reflector systems for radio astronomy applications, where especially tight specifications are placed on some of the important performance metrics. In several cases the response surfaces of these performance metrics are computationally expensive to compute, and, as an added complication, they are often conflicting and need to be optimized in a multi-objective (MO) or Pareto framework. The large numbers of function evaluations required in these optimization loops make direct full wave simulation in the loop intractable, and SBO methods may in these cases lead to reliable global MO optimizations.

The paper is organized to first discuss the basics of reflector antenna systems analysis and some typical performance metrics. Given this background, the difficulty in producing accurate optimized designs will be highlighted, and discussions presented on surrogate modeling based methods that can be used to accelerate the design process. Several examples of different design applications will be presented, which include the design of the dish surfaces for a given feed radiation pattern, as well as the design of feed antennas for given reflector configurations.

2 REFLECTOR ANTENNA SYSTEMS BACKGROUND

In the most general terms, a reflector antenna system may be seen (in the transmit sense) as a transducer which transforms a spherical wavefront primary pattern, from some feed antenna, to a collimated plane wave which is radiated as the secondary pattern. This description is helpful in understanding the simplest approximation method used to analyze and design such systems, namely GO. The main assumption of GO is that the reflector surfaces are electrically very large and smooth, leading to specular reflection conditions for the ray tube approximations used to describe the radiation patterns. From here it is straight forward to analyze and describe the basic

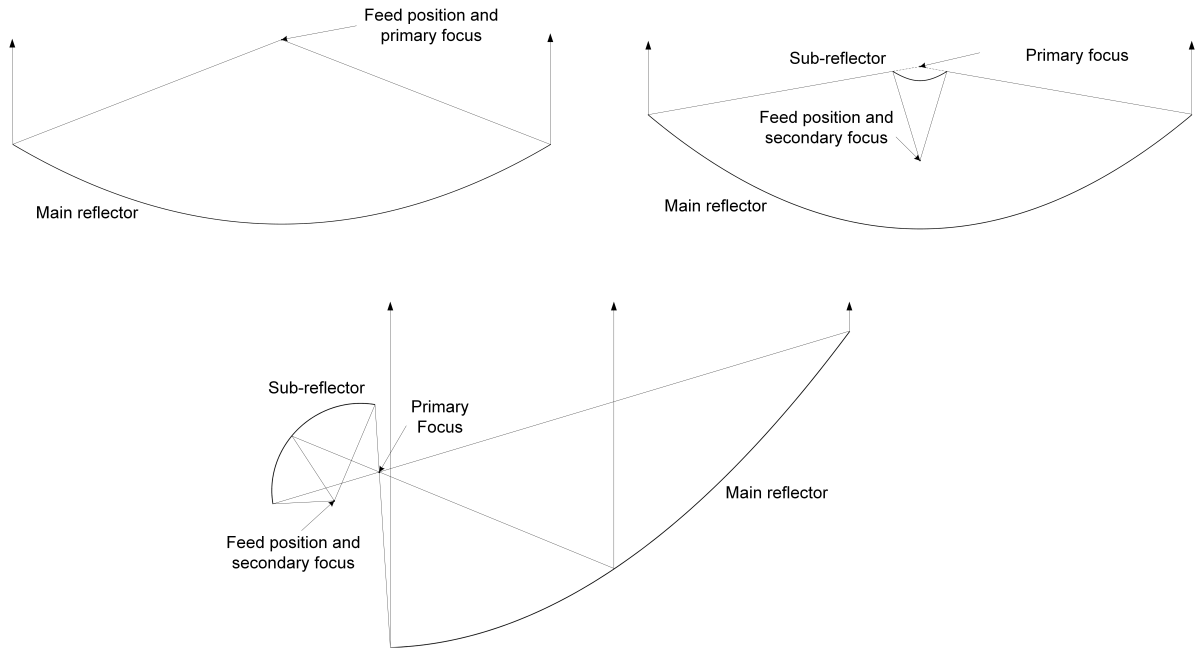


Figure 1: Outline geometry of some popular reflector configurations. Top left: Symmetrical paraboloid; Top right: Symmetrical Cassegrain; Bottom: Offset Gregorian. Ray paths, feed positions and focii for conic section (unshaped) systems are also indicated.

mechanisms of reflector systems – the simplest of which are possibly the family of conic section reflectors. These include the focus fed paraboloid, as well as Cassegrain and Gregorian dual reflector systems [1]. All these systems may be symmetrical or offset, with GO based design details available in the open literature [2, 3, 4, 5, 6], and sketches of some of the configurations shown in Fig. 1.

To achieve optimal performance, which typically involves maximizing the gain or receiving sensitivity (the ratio of effective aperture area and system temperature) while maintaining acceptable sidelobe levels (SLL) and cross polarization isolation (XPI), much effort has gone into developing reflector feed antennas with suitable radiation pattern characteristics [7, 8].

Another method used to improve performance is reflector shaping. Here the reflector antenna surfaces are shaped to provide a desired secondary pattern (normally specified as an aperture distribution) for a given primary radiation pattern. This strategy is most often employed in dual reflector systems, where a constant path length from the focus to the aperture can be enforced, along with specular reflection and power conservation [9, 10, 11, 12].

When designing very high performance reflector systems, and when the reflector surfaces become electrically smaller, the GO approximation no longer provides adequate accuracy. Some examples of such designs include the design of feeds for modern radio telescopes [13], PO synthesis of shaped beam reflectors for satellite applications [14], and design of small microwave link antennas for maximum sensitivity [15]. In all of these designs some other, more accurate, analysis method is employed in the optimization loop – mostly with a significantly design time penalty. The slow analysis time problem becomes especially pronounced when full wave simulations are required in the optimization loop, such as when the feed antenna parameters are varied or the reflector system is tightly coupled with the feed and not electrically large. Another situation that impedes the simulation time is when the receiving sensitivity is a goal function of the design. To calculate this response, the antenna radiometric noise temperature is required, which requires integration of the secondary pattern over the full 4π steradian sphere. Some

methods have been presented to accelerate this calculation, but at a cost in accuracy [16, 17]. The main bottleneck here is the requirement to calculate the secondary pattern over a fine angular grid, which becomes slow even for asymptotic methods such as PO. Some recent advances have accelerated this process as well, but it is most effective for extremely large problems [18].

3 SURROGATE MODELLING METHODS IN REFLECTOR DESIGN

In many reflector antenna applications, especially where pencil beams are required as the secondary pattern, the performance of the system is described by some derived quantities from the secondary pattern. These are normally the gain (described in normalized form as aperture efficiency), the receiving sensitivity, the SLL and XPI. Furthermore, these performance metrics are almost always conflicting, which makes Pareto base multi-objective (MO) optimization in the design process very attractive to quantify the trade-offs. MO optimization, however, typically requires a large number of function evaluations to converge, which may lead to unacceptably long run times if the full reflector analysis is included in the optimization loop. Faster, surrogate models are thus required to make such optimization based designs tractable.

3.1 Interpolation based models

Recently, interpolation based models have been used with some success in analysis and design of high performance reflector antenna systems. Interpolation of the response (hyper)surfaces may be required as functions of frequency, of physical parameters, or both.

A well known effect in many reflector systems is that of chromatic aberration, where a ripple with frequency is observed in the secondary radiation pattern amplitude close to broad-side. The cause of the ripple is interference of radiated fields with different path lengths which arise for a variety of reasons including coupling between the feed and the reflector or sub-reflectors in symmetrical systems [19], or back radiation from the feed or diffraction from the sub-reflector in offset systems [20]. Directly resolving the ripple in, for instance, the aperture efficiency will require a large number of samples in frequency. A Gaussian Process (GP) based interpolation scheme was suggested in [21], which is able to resolve the frequency ripple in offset Gregorian systems using only a fraction of the samples required for Nyquist convergence. An example of the possible performance is shown in Fig. 2, where the specific example is the offset Gregorian reflector system described in [21].

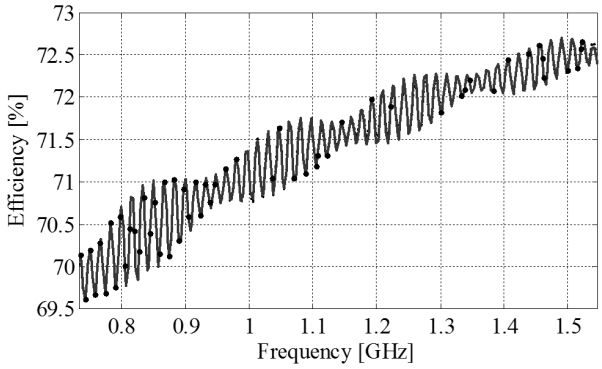


Figure 2: GP interpolation results for the aperture efficiency of an offset Gregorian system as a function of frequency. The training points are indicated as dots, the predicted response as a solid line, and the target response as a dashed line. The target response is mostly not visible as it is behind the interpolated response. Example taken from [21].

Frequency response interpolation is mostly used for detailed analysis of different responses using full wave simulations to model all the interactions between die reflector surfaces and feed. In the optimization based design context, some scalarization is normally performed over the frequency axis. It turns out that, for most design parameters of interest, the scalarized responses are often smooth functions. Some examples from the literature include aperture efficiency as a function of subtended angle in focus fed paraboloids [22], as well as receiving sensitivity, aperture efficiency, SLL and XPI of classical and shaped offset Gregorian systems [23, 24]. An example of some typical response surfaces is given in Section 4. Given these smooth responses as functions of design variables, interpolation based surrogate models can easily be constructed. All optimizations, including MO ones, are then performed on the interpolant instead of the actual simulation models [25].

3.2 Multi-fidelity models

In the above section on interpolation models, the allocation of a base set to ensure a converged interpolant is an important part of the model design, and may still require significant simulation time. This is because all the simulations are performed at the level of a fine model – meaning here the simulation model is as accurate as required for the specific design. The fine model may be GO or PO for the main beam if the reflectors are very large electrically, but in smaller systems, or when receiving sensitivity is required, these models are no longer sufficiently accurate. In this case they may be used as a faster coarse model, to augment the (limited) information from the accurate fine models obtained via, for instance, full wave simulations. Surrogate modeling, in this context, attempts to align the fully sampled coarse model to the sparsely sampled fine model over some region of interest in the design space. This is the multi-fidelity modeling scenario, and it is very well suited to reflector design problems – mainly due to the availability of a physics based (often GO) coarse model which is very fast to evaluate.

Examples of multi-fidelity modeling and design of reflector systems again include modeling of frequency and geometric parameters. In the case of frequency variations, the ripple mentioned in the previous sub-section may also be modeled using a multi-fidelity approach. Since the physical cause of the ripple is well understood, the period of the ripple may be estimated from the physical configuration of the system. This idea was implemented for the aperture efficiency of an offset Gregorian system in [26], where the amplitude of the ripple was estimated using PO simulations which did not include the main reflector. The response constructed in this way therefore used a multi-fidelity model description – the fine model is the full system simulated at a few points in frequency, and the coarse model is constructed using simulations of only the feed and sub-reflector, along with GO, to estimate the expected fields in the direction of the main beam. The models were aligned using a regression scheme, as well as band-limited sampling theory. An example of the possible performance is shown in Fig. 3, where the specific example is the offset Gregorian reflector system described in [26].

Multi-fidelity models have also been successfully used in MO optimization and entire domain modeling of reflector systems in terms of their geometric parameters. Here the correlation between coarse GO and fine PO or full wave models is exploited by aligning the coarse models to the fine models, in regions of interest in the design space, using two general methods. The alignment methods used are space mapping (SM) [27] as well as regression through the residuals between the coarse and fine models. SM based methods are demonstrated for MO optimization problems in [28], where the shaped surface of an offset Gregorian system is designed for maximum aperture efficiency and minimum SLL, and in [29], where a feed antenna for a

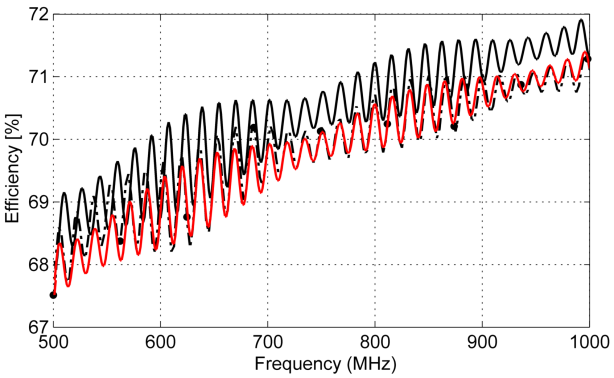


Figure 3: Multi-fidelity modeling results for the aperture efficiency of an offset Gregorian system as a function of frequency. The training points are indicated as dots, the coarse model as a solid black line, the fine model as a dashed line, and the aligned surrogate model as a solid red line. Example taken from [26].

classical Cassegrain system is designed for the same goals. Implicit SM has also been used to accurately model the SLL and aperture efficiency of blocked aperture symmetrical reflector systems over the full design space [30]. The residual regression method was used to model the radiometric noise temperature, used in the receiving sensitivity calculation, of a variety of reflector systems in [31], and for the MO optimization of feed antennas for an offset Gregorian system in [25].

In all these examples interpolation of the coarse model over the parameter space is also employed. An illustrative example is shown in the following section to highlight the main steps in a multi-fidelity MO optimization design procedure.

4 EXAMPLE

An example that illustrates both the interpolation and multi-fidelity modeling schemes, in the MO optimization context, is that discussed in [28]. In this problem the shape of the dishes, described by two parameters σ and b, is to be optimized for maximum aperture efficiency and minimum SLL. A fixed feed pattern is assumed, and details of the geometry is provided in [28]. GO is used as a coarse model, and a PO simulation using the commercial package GRASP [32] is used as the fine model. The optimization algorithm involves the following steps:

1. Allocate a base set X'_B, as a grid over the full parameter space, and evaluate the coarse model $\mathbf{c}(X'_B)$.

2. Construct a Kriging interpolant [33] through $\mathbf{c}(X'_B)$ to obtain the coarse model over the full parameter space $\mathbf{x}$ as $\mathbf{c}(\mathbf{x})$.

3. Allocate an initial fine model base set X_B and evaluate the fine model $\mathbf{f}(X_B)$.

4. Construct a SM based surrogate model $\mathbf{s}(\mathbf{x})$ using multi-point parameter extraction [27] at X_B.

5. Identify the Pareto front on $\mathbf{s}$ using some multi-objective evolutionary algorithm (or a grid search – the model evaluation is fast enough).

6. Sample the Pareto front obtained in step 4 to obtain a new base set X_B; evaluate the fine model at X_B and compare to the corresponding surrogate model samples.

7. If the termination condition is not satisfied, return to step 4.

Termination of the algorithm is based on the least squares distance between the Pareto set representations obtained in consecutive iterations. The important point here is that the surrogate model is aligned to the fine model only in the region of the estimated Pareto front – it may be quite inaccurate away from the front. Therefore the method is best suited to problems with smooth response surfaces, such as those encountered here. The method is illustrated in Figs. 4 and 5, where the improved alignment of the surrogate model is obvious along the Pareto front for the second iteration. Also note how the Pareto set representation lies on the surrogate model surface after the second iteration, indicating very good alignment with the actual fine model in this region.

5 CONCLUSIONS AND FUTURE WORK

This paper provided a summary of some recent advances in surrogate modeling of reflector antenna systems. Two main characteristics of reflector systems that allow relatively simple, accurate and low cost surrogate models to be developed, are the slow (or periodic) variation of the most important performance metrics as functions of the design parameters, as well as the availability of simple GO models of these responses. These characteristics have been exploited for some simple design cases, and a proof of concept established. Future work include a multitude of extensions to the current design and analysis methods, mostly working towards extending the MO optimization algorithm to more realistic, higher dimensional, design problems. Here more sophisticated sampling schemes should be employed to establish and update the interpolants, while allowing exploration of the search space away from the Pareto front [34].

REFERENCES

[1] W. V. T. Rusch and P. D. Potter, *Analysis of Reflector Antennas*. Academic Press, Inc., 1970.

[2] S. Rao, L. Shafai, and S. K. Sharma, Eds., *Handbook of Reflector Antennas and Feed Systems: Theory and design of reflectors*. Boston, MA, USA: Artech house, 2013, vol. 1.

[3] C. Granet, "Designing axially symmetric Cassegrain or Gregorian dual-reflector antennas from combinations of prescribed geometric parameters," *IEEE Antennas Propag. Mag.*, vol. 40, no. 2, pp. 76–82, Apr. 1998.

[4] W. V. T. Rusch, A. Prata Jr., Y. Rahmat-Samii, and R. A. Shore, "Derivation and application of the equivalent paraboloid for classical offset cassegrain and gregorian antennas," *IEEE Trans. Antennas Propag.*, vol. 38, no. 8, pp. 1141–1149, Aug. 1990.

[5] K. W. Brown and A. Prata Jr., "A design procedure for classical offset dual reflector antennas with circular apertures," *IEEE Trans. Antennas Propag.*, vol. 42, no. 8, pp. 1145–1153, Aug. 1994.

[6] C. Granet, "Designing classical offset cassegrain or gregorian dual-reflector antennas from combinations of prescribed geometric parameters," *IEEE Antennas Propag. Magazine*, vol. 44, no. 3, pp. 114–123, Jun. 2002.

[7] A. D. Olver, P. J. B. Clarricoats, A. A. Kishk, and L. Shafai, *Microwave horns and feeds*. IEEE Press, 1994.

[8] S. Rao, S. K. Sharma, and L. Shafai, Eds., *Handbook of Reflector Antennas and Feed Systems: Feed Systems*. Boston, MA, USA: Artech house, 2013, vol. 2.

[9] V. Galindo, "Design of dual-reflector antennas with arbitrary phase and amplitude distributions," *IEEE Trans. Antennas Propag.*, vol. 12, no. 4, pp. 403–408, Jul. 1964.

[10] V. Galindo-Israel, R. Mittra, and A. G. Cha, "Aperture amplitude and phase control of offset dual reflectors," *IEEE Trans. Antennas Propag.*, vol. AP-27, pp. 154–164, 1979.

[11] V. Galindo-Israel, W. A. Imbriale, and R. Mittra, "On the theory of the synthesis of single and dual offset shaped reflector antennas," *IEEE Trans. Antennas Propag.*, vol. AP-35, no. 8, pp. 887–896, Aug. 1987.

[12] P.-S. Kildal, "Synthesis of multireflector antennas by kinematic and dynamic ray tracing," *IEEE Trans. Antennas Propag.*, vol. 38, no. 10, pp. 1587–1599, Oct. 1990.

[13] R. Lehmensiek and I. P. Theron, "L-band feed horn and orthogonal mode transducer for the KAT-7 radio telescope," *IEEE Trans. Antennas Prapag.*, vol. 59, no. 6, pp. 1894–1901, Jun. 2011.

[14] D.-W. Duan and Y. Rahmat-Samii, "A generalized diffraction synthesis technique for high performance reflector antennas," *IEEE Trans. Antennas Propag.*, vol. 43, no. 1, pp. 27–40, Jan. 1995.

[15] P.-S. Kildal, S. A. Skyttemyr, and A. A. Kishk, "G/T maximization of a paraboloidal reflector fed by a dipole-disk antenna with ring by using the multiple-reflection approach and the moment method," *IEEE Trans. Antennas Propag.*, vol. 45, no. 7, pp. 1130–1139, Jul. 1997.

[16] W. A. Imbriale, "Faster antenna noise temperature calculations using a novel approximation technique," in *Proc. IEEE Int. Symp. Antennas and Propag.*, Toronto, ON, Canada, Jul. 2010, pp. 1 – 4.

[17] D. I. L. de Villiers and R. Lehmensiek, "Rapid calculation of antenna noise temperature in offset Gregorian reflector systems," *IEEE Trans. Antennas Propag.*, vol. 63, no. 4, pp. 1564–1571, Apr. 2015.

[18] O. Borries, H. H. Viskum, P. Meincke, E. Jørgensen, P. C. Hansen, and C. H. Schmidt, "Analysis of electrically large antennas using fast physical optics," in *Proc. Eur. Conf. Antennas and Propag. (EuCAP)*, Lisbon, Portugal, Apr. 2015.

[19] D. Morris, "Chromatism in radio telescopes due to blocking and feed scattering," *Astron. Astrophysics*, vol. 67, pp. 221–228, Jul. 1978.

[20] D. I. L. de Villiers, "Gain ripple in small offset gregorian antennas," in *Proc. IEEE Int. Symp. Antennas and Propag. (APSURSI)*, Jul. 2011, pp. 2172 –2175.

[21] J. P. Jacobs and D. I. L. de Villiers, "Gaussian process modeling of aperture efficiency ripple in reflector antennas," in *Proc. Loughborough Antennas Propag. Conf. (LAPC)*, Nov. 2015, pp. 1–4.

[22] C. A. Balanis, *Antenna Theory: Analysis and Design*, 3rd ed. Wiley, 2005, ch. 15, pp. 893–915.

[23] D. I. L. de Villiers and R. Lehmensiek, "Analytical evaluation of the efficiency improvement of shaped over classical offset dual-reflector antennas including sub-reflector diffraction," in *Proc. Int. Conf. Electromagn. Adv. Appl. (ICEAA)*, Turin, Italy, Sep. 2011, pp. 191–194.

[24] ——, "Sensitivity performance of several mappings in shaped offset Gregorian reflectors," in *Proc. IEEE Int. Symp. Antennas and Propag. (APSURSI)*, Memphis, TN, Jul. 2014, pp. 1477–1478.

[25] A. A. Vermeulen, "The design of a dual reflector feed using surrogate modeling techniques," Master's thesis, Stellenbosch University, Stellenbosch, South Africa, Dec. 2015.

[26] D. I. L. de Villiers, "Prediction of aperture efficiency ripple in clear aperture offset Gregorian antennas," *IEEE Trans. Antennas Propag.*, vol. 61, no. 5, pp. 2457–2465, May 2013.

[27] S. Koziel, J. W. Bandler, and K. Madsen, "A space-mapping framework for engineering optimization – theory and implementation," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 10, pp. 3721 – 3730, Oct. 2006.

[28] D. I. L. de Villiers and S. M. Koziel, "Fast multi-objective optimization of shaped offset Gregorian reflector systems," in *Proc. IEEE Int. Symp. Antennas and Propag. (APSURSI)*, Jul. 2015, pp. 1342–1343.

[29] ——, "Multi-objective optimization of Cassegrain reflector feeds using space mapping surrogate models," in *IEEE-APS Topical Conf. Antennas and Propag. in Wireless Comm. (APWC)*, Sep. 2015, pp. 813–816.

[30] D. I. L. De Villiers, "Accurate parametric modeling of gain and sidelobe levels in blocked aperture reflector systems using implicit space mapping," in *Proc. Eur. Conf. Antennas and Propag. (EuCAP)*, Davos, Switzerland, Apr. 2016, pp. 1–4.

[31] ——, "Fast parametric modeling of radio astronomy reflector antenna noise temperature," *IEEE Trans. Antennas Propag.*, 2016, in print.

[32] TICRA, Copenhagen, Denmark. GRASP10, Version 10.4.0. [Online]. Available: www.ticra.com

[33] S. Koziel, J. W. Bandler, and Q. S. Cheng, "Reduced-cost microwave component modeling using space mapping-enhanced electromagnetic-based kriging surrogates," *Int. J. Numer. Model.*, vol. 26, no. 3, pp. 275–286, 2013.

[34] I. Couckuyt, D. Deschrijver, and T. Dhaene, "Fast calculation of multiobjective probability of improvement and expected improvement criteria for pareto optimization," *Journal of Global Optimization*, vol. 60, no. 3, pp. 575–594, 2014.

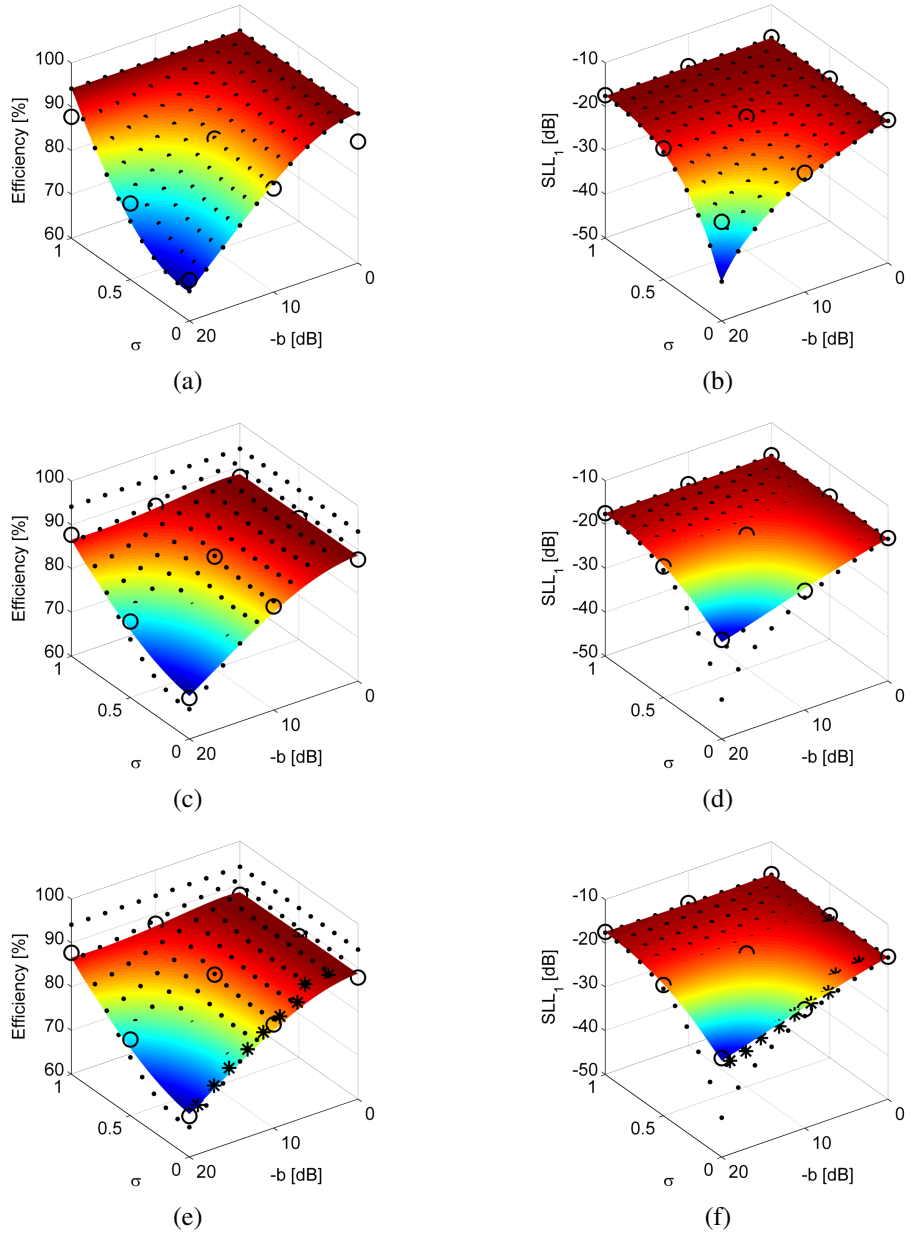


Figure 4: The evolution of the first iteration of the MO optimization algorithm using interpolation and multi-fidelity based models. The left column shows aperture efficiency, and the right the SLL. Dots indicate the coarse model base set X'_B, circles the fine model base set X_B, and stars the sampled Pareto set representation. The surface shows the smooth surrogate model. Steps 1-3 are illustrated in (a,b), step 4 in (c,d) and step 5 in (e,f).

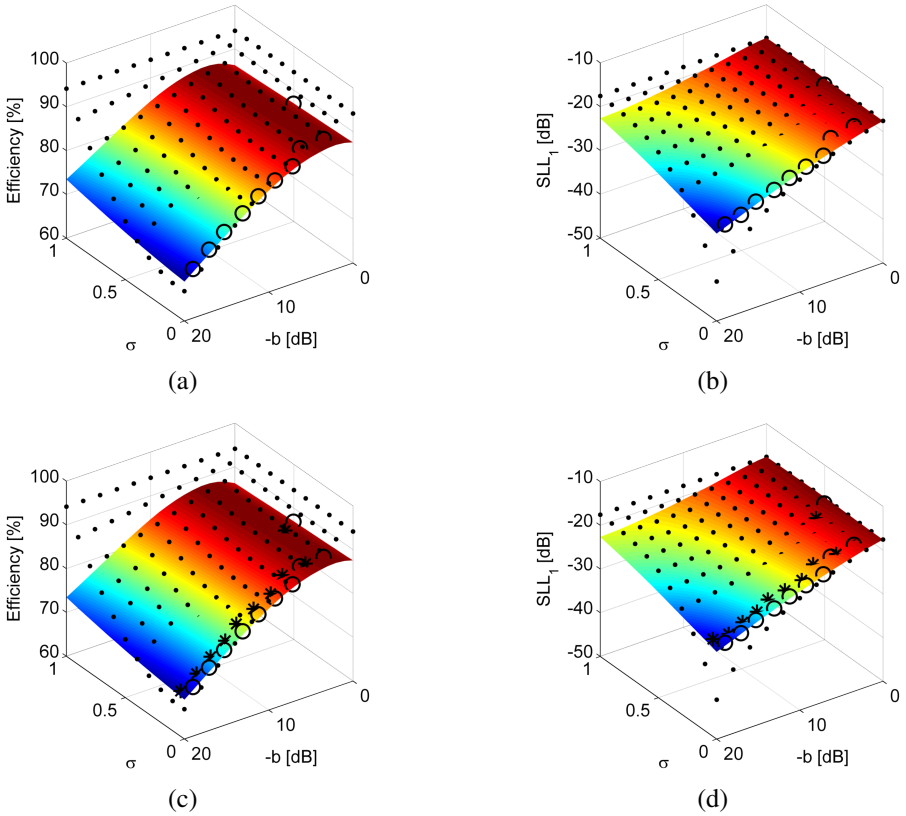


Figure 5: The evolution of the second and final iteration of the MO optimization algorithm using interpolation and multi-fidelity based models. The left column shows aperture efficiency, and the right the SLL. Dots indicate the coarse model base set X'_B, circles the fine model base set X_B, and stars the sampled Pareto set representation. The surface shows the smooth surrogate model. Step 4 is illustrated in in (a,b) and step 5 in (c,d).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A SHORT-TIME PRONY METHOD FOR THE DETECTION OF TRANSIENTS

Annie Cuyt[1], **Wen-shin Lee**[1], **and Min-nan Tsai**[1,2]

[1]Department of Mathematics and Computer Science
University of Antwerp (CMI), Belgium
e-mail: {annie.cuyt, wen-shin.lee}@unantwerpen.be

[2] Department of Computer Science and Information Engineering
National Taiwan Normal University, Taiwan
e-mail: edwen0224@gmail.com

Keywords: Exponential analysis, sparse interpolation, Prony's method, transients, power system.

Abstract. *Many parametric spectral methods are based on the classical algorithm of the French engineer G. de Prony for exponential analysis. A drawback of this method is that it cannot take into consideration any discontinuities due to the starting and ending of the exponential components at different instants.*

We introduce a short-time Prony method that allows to extract the characteristics from such a signal and we illustrate the new method on a number of power system signals. All parameters in the signals can be extracted with high accuracy and we show how to monitor the occurrence of the transients dynamically.

1 INTRODUCTION

A transient is used to refer to any signal or wave that is short lived. Transient detection has applications in many highly dynamic signals such as power line analysis, speech and image processing, turbulent flow applications, to name just a few. In a power system signal, transients are observed as short lived high-frequency oscillations superimposed on the voltages or currents of the fundamental frequency which is 50/60 Hz, as well as exponential components. They can be caused by lightnings, equipment faults, switching operations and so on.

Transient detection is a relatively common problem in many applications. They are often sinusoidal in nature and a lot of research has gone into the automation of their detection. We analyze how a sparsity based method can contribute to the power system analysis. A drawback of the standard Prony method for exponential analysis, is that it does not take into consideration the discontinuities due to the switching, in other words due to the exponential components to start and end at different instants.

In Section 2 the standard Prony method is recalled, while in Section 3 a short-time version of it is developed. The latter is applied successfully to a number of power signal simulations in Section 4. We use a damped sinusoidal model that is related to the phenomena typically observed at power system plants.

2 THE STANDARD PRONY METHOD

Let $\psi_i, \omega_i, \beta_i$ and γ_i respectively denote the damping, frequency, amplitude and phase in each component of the signal

$$\Phi(t) = \sum_{i=1}^{n} \alpha_i \exp(\phi_i t), \qquad \alpha_i = \beta_i \exp(\mathtt{i}\,\gamma_i), \qquad \phi_i = \psi_i + \mathtt{i}\,2\pi\omega_i. \tag{1}$$

Let us assume that the frequency content in (1) is limited by

$$|\Im(\phi_i)/(2\pi)| = |\omega_i| < \Omega/2, \qquad i = 1, \ldots, n,$$

as required by the Shannon-Nyquist theorem, and let $\Phi(t)$ be sampled at the equidistant points $t_j = j\Delta$ for $j = 0, 1, \ldots, 2n-1, \ldots$ with $\Delta \leq 1/\Omega$. In the sequel we denote

$$f_j := \Phi(t_j), \qquad j = 0, 1, \ldots, 2n-1.$$

The aim now is to find $n, \phi_1, \ldots, \phi_n, \alpha_1, \ldots, \alpha_n$ from the measurements $f_0, \ldots, f_{2n-1}, f_{2n}, \ldots$ and the form (1) of the model for $\Phi(t)$. The inherent structure present in (1) allows to separate the computation of the nonlinear parameters ϕ_i from that of the linear parameters α_i, as already dsicovered by the French engineer de Prony [7].

Let us define the Hankel matrices

$$H_n^{(r)} := \begin{bmatrix} f_r & \cdots & f_{r+n-1} \\ \vdots & \ddots & \vdots \\ f_{r+n-1} & \cdots & f_{r+2n-2} \end{bmatrix}, \qquad r \geq 0. \tag{2}$$

It is known [3, p. 603] that

$$\det H_\nu^{(r)} = 0, \qquad \nu > n, \tag{3}$$

and it is proved in [5] that, in the absence of noise,

$$\begin{aligned} &\det H_n^{(r)} \neq 0, \\ &\det H_\nu^{(r)} = 0 \text{ only accidentally}, \qquad \nu < n. \end{aligned} \tag{4}$$

From these statements the number of components n can be obtained as the (numerical) rank of $H_\nu^{(r)}$ for $\nu > n$. In practice, when the signal $\Phi(t)$ is affected by noise, the numerical rank is determined as the number of singular values of $H_\nu^{(r)}$ that rise above the noise level. In (4) this means that one finds n singular values clearly above the noise level and the remaining $\nu - n$ at or below the noise level.

We further denote

$$\lambda_i := \exp(\phi_i \Delta), \qquad i = 1, \ldots, n.$$

The λ_i can be retrieved [4] as the generalized eigenvalues of the problem

$$H_n^{(1)} v_i = \lambda_i H_n^{(0)} v_i, \qquad i = 1, \ldots, n \tag{5}$$

where v_i are the generalized right eigenvectors. Then from the values λ_i, the ϕ_i can uniquely be retrieved because of the restriction $|\Im(\phi_i \Delta)| < \pi$.

To conclude, one finds the α_i from the interpolation conditions

$$\sum_{i=1}^{n} \alpha_i \exp(\phi_i t_j) = f_j, \qquad j = 0, \ldots, 2n-1, \tag{6}$$

either by solving the system in the least squares sense in the presence of noise or by solving a subset of n consecutive interpolation conditions in case of a noisefree $\Phi(t)$. Note that

$$\exp(\phi_i t_j) = \lambda_i^j$$

and that the coefficient matrix of (6) is therefore a Vandermonde matrix.

We now present a reformulation of the exponential analysis problem using tools from rational approximation theory [1].

With $f_j = \Phi(t_j)$ we define the noisefree

$$F(t) = \sum_{j=0}^{\infty} f_j t^j. \tag{7}$$

Since

$$f_j = \sum_{i=1}^{n} \alpha_i \exp(j \phi_i \Delta) = \sum_{i=1}^{n} \alpha_i \lambda_i^j,$$

we can rewrite

$$F(t) = \sum_{i=1}^{n} \frac{\alpha_i}{1 - t\lambda_i}. \tag{8}$$

So we see that $F(t)$ is a rational function of degree $n-1$ in the numerator and n in the denominator, with poles $1/\lambda_i$. From Padé approximation theory we know that the Padé approximant $r_{n-1,n}(t)$ to $F(t)$ of degree $n-1$ in the numerator and n in the denominator, reconstructs $F(t)$, in other words

$$r_{n-1,n}(t) = F(t).$$

The partial fraction decomposition (8) is related to the Laplace transform of the exponential model (1), which explains why this approach is known as the Padé-Laplace method. When the signal $\Phi(t)$ is noisy, then the samples equal $f_j + \epsilon_j$ instead of f_j where ϵ_j denotes a noise term. Let us denote the noisy series by

$$F(t) + \epsilon(t) = \sum_{j=0}^{\infty} (f_j + \epsilon_j) t^j.$$

It is clear that the Padé approximant $r_{n-1,n}(t)$ does not reconstruct $F(t)+\epsilon(t)$ now as the latter is not a rational function anymore. For a novel way to make use of the Padé-Laplace reformulation in this situation, we refer the reader to [2].

3 A SHORT-TIME PRONY METHOD

When the model (1) changes dynamically across the time window used to collect the samples f_j, due to the fact that exponential components are switched on or off during the sampling, then the standard Prony method or its Padé-Laplace reformulation cannot be applied. In this section we adapt the model to allow for the components to start and/or end during the time of observation, to

$$\Psi(t) = \sum_{i=1}^{n} \alpha_i \exp(\phi_i(t)) \left(u(t - s_i) - u(t - e_i)\right), \tag{9}$$

where s_i and e_i are the start and end times of the i-th signal component and $u(t)$ is the unit step function. For ease of notation we still denote f_j for the time sample at time $t_j = j\Delta$. From the context it will always be clear whether f_j comes from $\Phi(t)$ as in (1) or $\Psi(t)$ as in (9). We also retain the notation $\lambda_i = \exp(\phi_i \Delta), i = 1, \ldots, n$.

We now combine the sampling of (9) with a window function $w(j - r)$ that is only nonzero in the time interval $[r\Delta, (r + 2\nu - 1)\Delta]$ with $\nu \geq n(r)$ where $n(r)$ denotes the number of exponential components switched on somewhere in the course of the time interval $[r\Delta, (r + 2\nu - 1)\Delta]$. Obviously $n(r) \leq n$. In our experiments we used the rectangular window function

$$w(j - r) = \begin{cases} 1, & r \leq j \leq r + 2\nu - 1, \\ 0, & \text{elsewhere.} \end{cases}$$

The Hankel matrix $H_\nu^{(r)}$ as defined in (2), makes use of the samples $f_r, \ldots, f_{r+2\nu-1}$ of (9). We consider its (numerical) rank stable when it does not alter while increasing r to $r + \rho$ with $\rho > 0$. Note that during the inspection of the (numerical) rank of the matrices $H_\nu^{(r)}, \ldots, H_\nu^{(r+\rho)}$, the early samples shift out of the matrix while they are being replaced by new samples. When we observe that the (numerical) rank of $H_\nu^{(r)}$ does not change in the time window $[r\Delta, (r + \rho + 2\nu - 1)\Delta]$, then the standard Prony method can be applied to extract the characteristics of the $n(r)$ components active in that time window. When the (numerical) rank of $H_\nu^{(r)}$ fluctuates while increasing r, we know that components are being switched on or off, and the extraction of their characteristics using Prony's method should be postponed to a stable time window because the signal is not following a model of type (1) in the observed window.

The width of the time window here is related to the number $n(r)$ of components in (9) that are switched on in the considered time interval. A smaller time window allows more easily to find stable intervals, while a larger time window allows to work with more components. This is very similar to a similar conclusion for the short-time Fourier transform, where a narrow

window gives good time resolution but poor frequency resolution and a wide window gives better frequency resolution but poor time resolution.

Then some words on the numerical computation of the ϕ_i and α_i in practice, from noise corrupted samples f_j. When the largest time window of stability for the computation of ν generalized eigenvalues in total (comprising the $n(r)$ active ones representing the signal space and $\nu - n(r)$ additional ones modeling the noise space), is $[r\Delta, (r+\rho+2\nu-1)\Delta]$, then the $\lambda_i = \exp(\phi_i \Delta)$ present in that time window are computed in the least squares sense rather than from (5),

$$H^{(r+1)}_{\nu+\lfloor \rho/2 \rfloor, \nu} v_i = \lambda_i H^{(r)}_{\nu+\lfloor \rho/2 \rfloor, \nu} v_i, \qquad i = 1, \ldots, \nu, \tag{10}$$

where the subscript of the Hankel structured matrix now indicates its dimension. Several popular exponential analysis implementations that can be used for this purpose are given in [9, 8, 4].

To adapt the Padé-Laplace formulation to work with a window function, we consider

$$F(t - r\Delta) = \sum_{i=1}^{n(r)} \frac{\alpha_i}{1 - (t - r\Delta)\lambda_i}, \qquad t_r \leq t \leq t_{r+\rho+2n(r)-1}.$$

Since this time interval delivers enough samples f_j for the computation of the Padé approximant $r_{n(r)-1,n(r)}(t - r\Delta)$, namely at least $2n(r)$, the reformulation still holds.

4 SIMULATION RESULTS

We describe some experiments that illustrate the use of the short-time Prony method. The considered synthesized electric signals are taken from [6]. Since these are a linear combination of switched cosine functions, more precisely

$$\Psi(t) = \sum_{i=1}^{n/2} \beta_i \cos(2\pi\omega_i(t) + \gamma_i) \exp(\psi_i(t - s_i)) \left(u(t - s_i) - u(t - e_i)\right), \tag{11}$$

each term is represented by 2 exponential terms with complex conjugate values ϕ_i. The characteristics of the test signals are given in Table 1. Besides the start and end moments s_i and e_i of each component in seconds, we also list the sample numbers S_i and E_i at which the switch occurs. All signals are monitored for a total time span of 128 samples, numbered from 0 to 127. The sampling rate is given by $1/\Delta$ (in Hz) where Δ is the time step. Each signal is then corrupted by 32 dB white Gaussian noise, which for some test signals is quite a lot more than in [6]. After extracting the frequencies ω_i and damping factor ψ_i, the computed coefficient $\beta_i \exp(-\psi_i s_i)$ needs to be corrected using the s_i obtained from the singular value plots, in order to have the correct amplitude β_i.

In Figure 1 the four test signals are graphed over the entire observation window. We remark that all signals are dead at $t = 127\Delta$, in other words $f_{127} = 0$. For each noise corrupted signal we graph in Figure 2 the dynamic evolution of the numerical rank of a particular Hankel matrix by displaying how the ν singular values of $H^{(r)}_{\nu}$ change when r is ranging from 0 to $129 - 2\nu$ (the value of r is on the x-axis). We display respectively the singular values of $H^{(r)}_4$ for $\Psi_1(t)$, $H^{(r)}_6$ for $\Psi_2(t)$, $H^{(r)}_6$ for $\Psi_3(t)$ and $H^{(r)}_8$ for $\Psi_4(t)$, and this for $0 \leq r \leq 129 - 2\nu$. Let us discuss these graphs before turning to the reconstruction of the signal parameters.

In $\Psi_1(t)$ the Hankel matrices $H^{(0)}_4$ to $H^{(30)}_4$ indicate a numerical rank of zero. Then there is a switch from no components at all to one component at sample 37 which enters the Hankel matrix $H^{(31)}_4$ for which neither statement (3) nor statement (4) hold. Nothing can be deduced

	n	β_i	γ_i	ω_i	ψ_i	s_i	S_i	e_i	E_i	$1/\Delta$
$\Psi_1(t)$	1	1.000	$-\pi/2$	480	-0.080	0.0308	37	0.1059	127	1200
$\Psi_2(t)$	3	1.000	$-\pi/2$	60	0.000	0.0000	0	0.0308	37	1200
		1.000	$3\pi/4$	60	0.000	0.0308	37	0.0625	75	
		1.000	$-\pi/2$	60	0.000	0.0625	75	0.1059	127	
$\Psi_3(t)$	3	1.000	0	50	0.000	0.0000	0	0.0200	30	1500
		0.250	$\pi/2$	100	-0.100	0.0200	30	0.0533	80	
		0.100	$38\pi/100$	300	0.030	0.0447	67	0.0847	127	
$\Psi_4(t)$	4	1.000	0	60	0.000	0.0000	0	0.0333	40	1200
		0.500	$\pi/2$	60	0.000	0.0333	40	0.0917	110	
		0.050	$-37\pi/100$	180	0.000	0.0417	50	0.0833	100	
		0.200	$-\pi/2$	360	-0.100	0.0500	60	0.1059	127	

Table 1: Signal parameters for the four test signals.

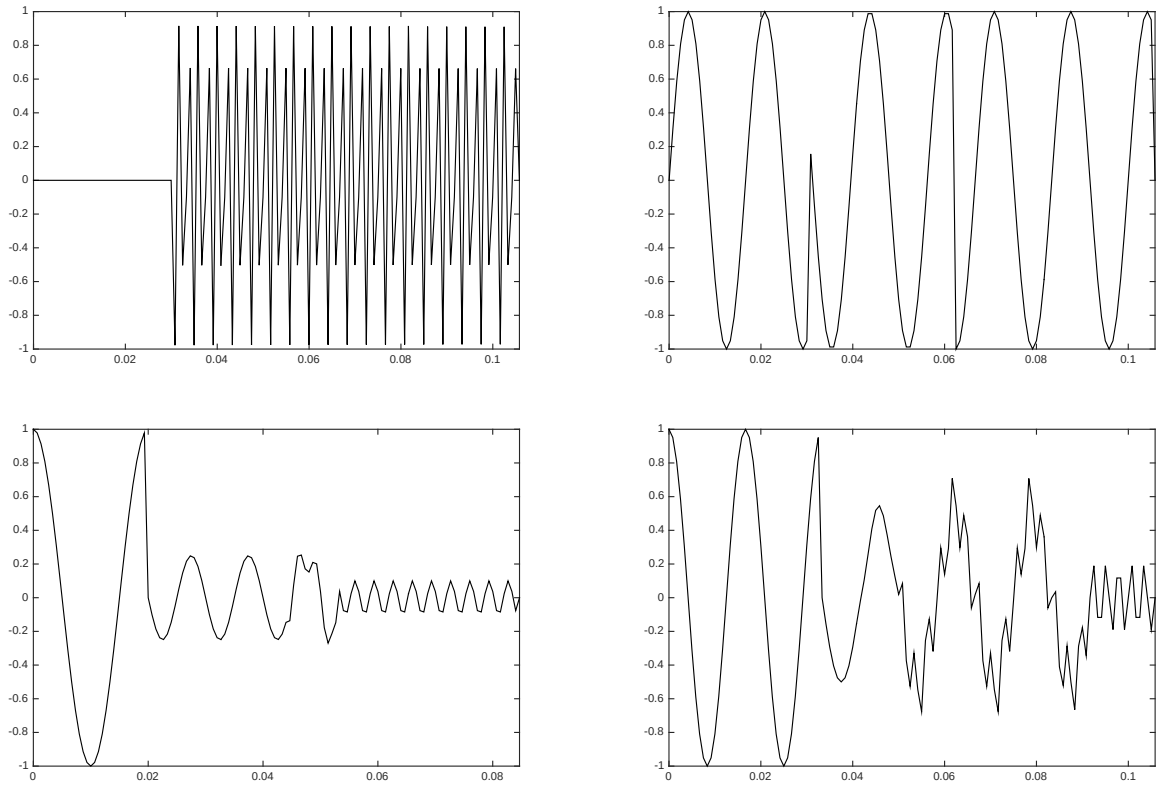


Figure 1: The signals $\Psi_1(t)$ (top left), $\Psi_2(t)$ (top right),$\Psi_3(t)$ (bottom left), $\Psi_4(t)$ (bottom right).

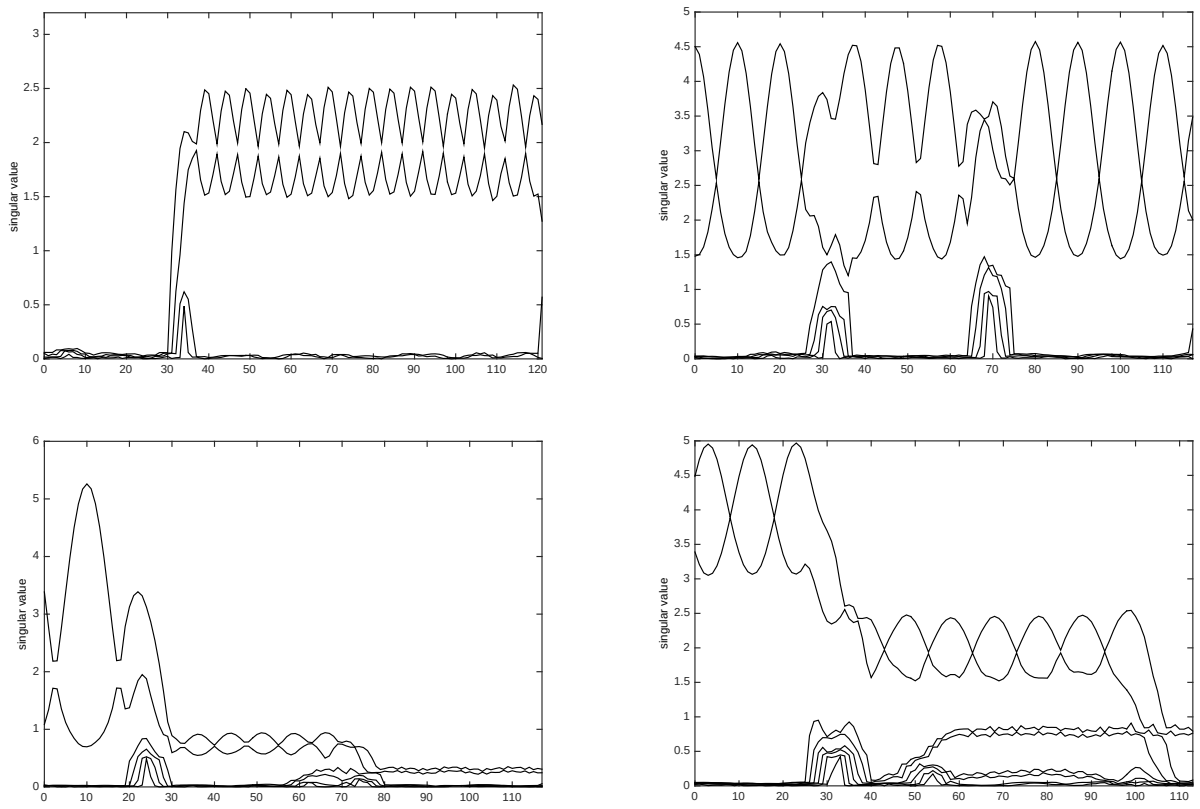


Figure 2: The dynamic singular value plots for $\Psi_1(t)$ (top left), $\Psi_2(t)$ (top right), $\Psi_3(t)$ (bottom left), $\Psi_4(t)$ (bottom right).

from the singular value behaviour to $H_4^{(36)}$. From $H_4^{(37)}$ on only one component is present and this for the remainder of the interval. The numerical rank is stable and equal to 2. The single component can be identified from the signal samples $f_r, 37 \leq r \leq 126$.

In $\Psi_2(t)$ three different components show up, but consecutively: the switches happen at the time samples indexed 37, 75 and 127. So the numerical rank of the observed Hankel matrix is unstable while samples from different signals move out and in because the conclusions (3) and (4) do not hold during the short time span of the instability. To be more precise: sample 37 enters $H_6^{(r)}$ for $r + 10 = 37$, sample 75 enters $H_6^{(r)}$ for $r + 10 = 75$ and sample 127 enters $H_6^{(r)}$ for $r + 10 = 127$. So the numerical rank instability is observed for $27 \leq r < 37, 65 \leq r < 75$ and at $r = 117$. Except for these unstable rank windows, the numerical rank equals 2. Each of the components can be identified separately: the first from $H_\nu^{(r)}$ with $0 \leq r$ and $r+2\nu-2 \leq 36$, the second one from $H_\nu^{(r)}$ with $37 \leq r$ and $r+2\nu-2 \leq 74$, the third one from $H_\nu^{(r)}$ with $75 \leq r$ and $r + 2\nu - 2 \leq 126$.

The results for $\Psi_3(t)$ and $\Psi_4(t)$ are more interesting. In $\Psi_3(t)$ the exponential model (1) is interrupted from $H_6^{(20)}$ to $H_6^{(29)}$ for a first time. But the first exponential component can be computed from the samples 0 to 29. Likewise the second component can be computed from the samples 30 to 66. In $H_6^{(57)}$ sample 67 sneaks in. From $H_6^{(67)}$ to $H_6^{(69)}$ components two and three strictly follow (1) but this window involving the samples 67 to 79 is barely enough to allow their identification. However, exponential component three can be computed from the samples 80 to 126.

In $\Psi_4(t)$ a similar situation arises, but now there is a sufficiently large window of stability,

	n	β_i	γ_i	ω_i	ψ_i
$\Psi_1(t)$	1	1.006	-1.565	479.99	-0.250
$\Psi_2(t)$	3	1.013	-1.570	60.05	-0.784
		1.013	2.339	60.05	-0.764
		1.016	-1.597	60.05	-0.735
$\Psi_3(t)$	3	1.004	0.015	50.02	-0.377
		0.257	1.563	100.04	-2.347
		0.117	1.224	299.92	-6.417
$\Psi_4(t)$	4	1.006	0.009	60.03	-0.391
		0.503	1.557	60.04	-0.319
		0.058	-1.613	181.08	10.250
		0.202	-1.567	360.00	-0.519

Table 2: Reconstructed signal parameters for the four test signals.

namely from $H_8^{(60)}$ to $H_8^{(85)}$ when $r + 14 = 99$, to allow for the identification of the last three components while they strictly adhere to model (1). The numerical rank is stable and equal to 6 in that window. The remaining first component can be identified from the samples 0 to 39 since $H_8^{(r)}$ indicates a stable numerical rank of 2 for $0 \leq r \leq 25$.

All reconstructions of α_i and ϕ_i have been made from the largest stable time windows, as mentioned in (10) and can be found in Table 2. The displayed values are an average of reconstructions over a 100 different noise realizations. We remark that the damping factors ψ_i are more noise sensitive than the other parameters.

5 CONCLUSION

We have introduced a short-time Prony method that allows to extract the characteristics from a signal in which the exponential components are switched on and off, a situation that the standard Prony method is unable to deal with. The new method was illustrated on a number of power system signals taken from [6]. All parameters in the sinusoidal model (9) could be extracted with high accuracy and the occurrence of the transient could be monitored from the dynamics of the singular values of particular Hankel matrices.

REFERENCES

[1] Z. Bajzer, A.C. Myers, S.S. Sedarous, and F.G. Prendergast. Padé-Laplace method for analysis of fluorescence intensity decay. *Biophys J.*, 56(1):79–93, 1989.

[2] A. Cuyt, W.-s. Lee, G. Lo Bianco, and M.-n. Tsai. How to pick up faint components in exponential analysis. Technical report, Universiteit Antwerpen, 2016.

[3] P. Henrici. *Applied and computational complex analysis I.* John Wiley & Sons, New York, 1974.

[4] Y. Hua and T.K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. Acoust. Speech Signal Process.*, 38:814–824, 1990.

[5] E. Kaltofen, W.-s. Lee, and A.A. Lobo. Early termination in Ben-Or/Tiwari sparse interpolation and a hybrid of Zippel's algorithm. In *Proceedings of the 2000 International Symposium on Symbolic and Algebraic Computation*, 192–201, ACM Press, New York, NY, USA, 2000.

[6] L. Lovisolo, E.A.B. da Silva, M.A.M. Rodrigues, and P.S.R. Diniz. Efficient coherent adaptive representations of monitored electric signals in power systems using damped sinusoids. *IEEE Trans. Signal Proc.*, 53(10):3831–3846, 2005.

[7] R. de Prony. Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures. *J. Ec. Poly.*, 1:24–76, 1795.

[8] R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoustics, Speech and Signal Proc.*, 37(7):984–995, 1989.

[9] R.O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propagation*, AP-34(3):276–280, 1986.

\ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

SURROGATE MODELING OF ANTENNA RADIATION CHARACTERISTICS BY GAUSSIAN PROCESSES

J. Pieter Jacobs[1], **and Dirk I.L. de Villiers**[2]

[1] Department of Electrical, Electronic and Computer Engineering
University of Pretoria, Pretoria, 0002, South Africa
e-mail: jpjacobs@up.ac.za

[2] Department of Electrical and Electronic Engineering
Stellenbosch University, Stellenbosch, South Africa
ddv@sun.ac.za

Keywords: antennas, Gaussian processes, modeling.

Abstract. *An overview is given on recent developments in the modeling of radiation properties of antennas by means of Gaussian process (GP) regression. Two widely differing applications are considered, namely modeling of the (possibly detrimental) effects of using a finite substrate/ground plane on the gain pattern of a microstrip patch antenna; and characterizing the highly oscillatory aperture efficiency responses of an offset Gregorian reflector system. It is shown that GP modeling performs well with respect to both problems, with interesting generalization behaviour observed in the case of the reflector system.*

1 INTRODUCTION

Gaussian process (GP) regression [1] has in recent years been shown to be suitable for the modelling of highly nonlinear input characteristics of microwave antennas, notably input reflection coefficient as a function of design variables and frequency [2]. The robustness of these models has been verified through incorporating them into optimization frameworks that included space mapping [3] and genetic algorithms [4]**.**

A Gaussian process is a stochastic process that can be viewed as a generalization of the Gaussian probability distribution to functions, and is characterized in full by its mean and covariance functions. During GP regression, the calculations required for learning and inference can be carried out using standard Gaussian probability distribution mathematics. Advantages over neural networks – a widely used modeling approach in electromagnetics – include easier implementation and interpretation since the optimization of far fewer parameters (in the order of the dimension of the input vectors) is required compared to the number of weights in a multi-layer perceptron. Furthermore, GP regression can automatically take into account the relative importance of input features when making predictions (a degree of flexibility not available under kernel methods such as standard support vector regression).

The present paper aims to give an overview of developments in the relatively recent area of GP modeling of antenna radiation characteristics. Two problem types – significantly different in nature – are considered. In the first instance, GP regression is applied to modeling the effects of finite substrate/ground plane size on the gain patterns of a microstrip patch antenna (finite substrate effects are often of practical interest when dealing with microstrip antennas). The operating frequency is fixed, and gain is predicted as a function of spatial direction. It is found that GP regression employing a suitable covariance function is well suited to this task, and important ripples that arise in principal plane patterns can be accurately accounted for. The second problem involves modeling the aperture efficiency responses of offset Gregorian reflector systems against frequency. These responses are of a quasi-periodic nature and highly oscillatory, and it would be expected that many training points would be needed to yield models with good predictive accuracy. It is shown that GP regression using a suitably constructed composite covariance function can significantly lessen the expected training data burden. Furthermore, it is demonstrated that a disproportionally large part of the training data can be concentrated in a section of the input space where data are relatively inexpensive to obtain (i.e., at lower frequencies). The resulting GP model can capture the underlying response structure to such an extent that accurate predictions can be made in (higher-frequency) sections of the response containing multiple adjacent cycles and no training points.

2 BACKGROUND TO MODELING WITH GAUSSIAN PROCESSES

2.1 Standard Gaussian Process Regression

A Gaussian process (GP) is a set that consists of an infinite number of random variables of which any finite subset has a jointly Gaussian distribution; it can be viewed as a distribution over functions. It is a natural extension of the jointly Gaussian distribution to the case where the mean vector is infinitely long and the covariance matrix is of infinite dimension. Notationally, $f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where $f(\mathbf{x})$ is a sample function, $\mathbf{x}$ and $\mathbf{x}'$ are positions in R^D space, and $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are mean and covariance functions defining the GP [1].

Hyperparameters determine the properties of the mean and covariance functions (this is the only parameterisation that is involved). The covariance function $k(\mathbf{x}, \mathbf{x}')$ gives the covariance

between the random output variables $f(\mathbf{x})$ and $f(\mathbf{x}')$. Standard covariance functions include the squared-exponential (SE) covariance function

$$k_{SE}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}r^2\right); \tag{1}$$

and the rational quadratic covariance function [1]:

$$k_{RQ}(\mathbf{x},\mathbf{x}') = \sigma_f^2 (1+\frac{1}{2\alpha}r^2)^{-\alpha} \tag{2}$$

In (1)—(2),

$$r = \left(\sum_{k=1}^{D}\frac{(x_k - x'_k)^2}{\tau_k^2}\right)^{1/2} \tag{3}$$

where x_k and x'_k are the kth components of input vectors $\mathbf{x}$ and $\mathbf{x}'$ of dimension D, and $\{\tau_k | k = 1, ..., D\}$ are indicative of how quickly change occurs along the corresponding dimensions of the input space. σ_f^2 governs the overall variance of the process and α is a scaling parameter. σ_f^2, α, and $\{\tau_k | k = 1, ..., D\}$ are referred to as hyperparameters; they can be optimized through a process which finds the hyperparameters for which the negative log marginal likelihood is a minimum. The log marginal likelihood is given by [1]

$$\log p(\mathbf{y} | X) = -\frac{1}{2}\mathbf{y}^T K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi \tag{4}$$

where $K = (X, X)$ is the $n \times n$ matrix of covariances evaluated between all possible pairs of n training outputs using the covariance function, X is the $D \times n$ matrix of training input vectors $\mathbf{x}_i$, $|K|$ is the determinant of K, and $\mathbf{y}$ is the training target vector.

Making predictions requires constructing the *prior* distribution, which is a jointly Gaussian distribution over the n random variables that represent the training outputs (in column vector $\mathbf{f}$), and the n_* random variables representing the test outputs (in $\mathbf{f}_*$); in this work the distribution is assumed to have zero mean:

$$\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_*\end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix}K(X,X) & K(X,X_*)\\ K(X_*,X) & K(X_*,X_*)\end{bmatrix}\right) \tag{5}$$

In (5), $N(\mathbf{a},\mathbf{b})$ denotes a multivariate normal distribution with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{b}$; $K(X, X_*)$ is the $n \times n_*$ matrix of covariances evaluated between all possible pairs of n training and n_* test outputs; and X_* is a matrix containing the test input vectors (other submatrices have similar definitions). The *posterior* distribution (also multivariate Gaussian) is then obtained by conditioning the test outputs on the known training outputs **y.** Hence $\mathbf{f}_* | X_*, X, \mathbf{y} \sim N(\mathbf{m}, \Sigma)$, with mean vector $\mathbf{m}$ and covariance matrix Σ [1]

$$\mathbf{m} = K(X_*, X)K(X, X)^{-1}\mathbf{y} \tag{6}$$

$$\Sigma = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*) \tag{7}$$

The mean vector $\mathbf{m}$ contains the most likely output predictions associated with the test input vectors in X_*, while the diagonal of the covariance matrix Σ gives the predictive variances.

2.2 Gaussian Process Modeling of Quasi-Periodic Responses

Standard covariance functions, e.g., (1)—(2), are not sufficiently expressive to readily model highly oscillatory responses of the sort considered in Section 3.2. However, it was demonstrated [5] that a composite covariance function $k_C(x,x')$ defined as

$$k_C\left(x,x'\right)=k_{SE}\left(x,x'\right)\times k_{PER}\left(x,x'\right) \tag{8}$$

had the requisite flexibility. In (8) – where a one-dimensional input space is assumed – $k_{SE}(x,x')$ is the squared-exponential function (1), and $k_{PER}(x,x')$ is a periodic covariation function, suitable for modeling functions that are comprised of exact repetitions of a basic function [1]:

$$k_{PER}\left(x,x'\right)=\sigma_f^2\exp\left(-\frac{2\sin^2[\pi\left|x-x'\right|/\lambda]}{\theta^2}\right) \tag{9}$$

In (9), the hyperparameter λ determines the intervals between repetitions, θ is a length-scale parameter, and σ_f^2 is defined as above.

3 MODELING EXAMPLES

3.1 Modeling of finite substrate size effects on gain patterns of microstrip patch antennas

Fig. 1 illustrates the geometry of a probe-fed rectangular patch antenna on a single-layer dielectric substrate backed by a ground plane of the same planar footprint as the dielectric substrate. The patch had length L and width W, and the planar dimensions of the dielectric substrate/ground plane were D_x and D_y in the x and y directions respectively. The substrate height was fixed at h = 0.8 mm, its dielectric constant ε_r was 4.34, and its loss tangent was 0.02. The probe feed was positioned at y_f = 3.6 mm.

The aim was to establish whether E (φ = 90°) and H (φ = 0°) plane frontal gain patterns (corresponding to the half-space z > 0) could be accurately modeled as a function of four variables: substrate/ground plane dimensions D_x and D_y, and patch size L and W; hence the design vector was $\mathbf{u} = [D_x\ D_y\ L\ W]^T$. The design space was specified by the variable ranges $0.5\lambda_0 \le D_x \le 3\lambda_0$, $0.5\lambda_0 \le D_y \le 3\lambda_0$, 12.6 mm $\le L \le$ 15.4 mm, and 8.64 mm $\le W \le$ 10.56 mm (λ_0 was the free-space wavelength at the operating frequency of 5.02 GHz). Since GP regression allows for single model outputs only, it was necessary to set up separate models for the gain in each of the principal planes. Of particular interest was the GP models' ability to predict the pattern distortion that is well known to occur for finite substrate/ground planes (as opposed to the ideal case of the patch being supported by an infinite substrate).

In order to construct a training data set, 70 design vectors $\boldsymbol{u}$ were selected from the design space using Latin hypercube sampling (LHS). For each principal plane model, 10 elevation angles (i.e., θ values) per geometry were randomly selected from the range -90° $\le \theta \le$ 90° such that elevation angles in general differed from geometry to geometry. This yielded n = 700 training input vectors $\{\mathbf{x}_i=[\mathbf{u}_i^T\ \theta_i]^T=[D_{xi}\ D_{yi}\ L_i\ W_i\ \theta_i]^T \mid i=1,...,n\}$. The corresponding target outputs y_i were gain values in the E or H plane – denoted below as G_E and G_H; these were obtained by means of simulation with the time-domain solver in CST Microwave Studio [7]. Test data consisted of 20 new LHS-selected geometries with 61 equally spaced elevation angles per geometry, resulting in n^* = 1220 test data points. (In order to verify the simulation setup, the radiation patterns of a related case study [8] were reproduced using the CST solver.)

Next, four candidate Gaussian process models were trained for each of the G_E and G_H patterns, each model using a different covariance function that included the squared exponential (1), rational quadratic (2), and two Matérn type covariance functions [1]. A Gaussian likelihood function and a prior with a zero mean function was assumed throughout. From each set of four models, the model which produced the lowest negative log marginal likelihood (4) was selected as the optimal model, and used to make predictions on the test data. In both the

G_E and G_H cases, the rational quadratic covariance function (2) models produced the lowest negative log marginal likelihood.

The models' predictive accuracies are summarized in Table 1 in terms of percentage root-mean-square-errors (RMSEs) normalized to the test target ranges, and linear correlation coefficients R [6]. In spite of the significant variations in planar dimensions and consequently shape of the substrate, high accuracies were obtained, with RMSEs of about 1.75% for each of G_E and G_H. Figure 2 gives simulated and predicted patterns (E and H plane) for a geometry from the test data set, namely $D_x = 2.57\lambda_0$, $D_y = 1.85\lambda_0$, $L = 13.52$ mm, $W = 9.16$ mm. The agreement between simulated and predicted patterns is good, and the E-plane pattern ripples are well represented.

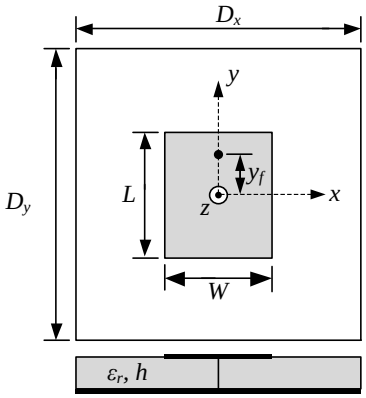
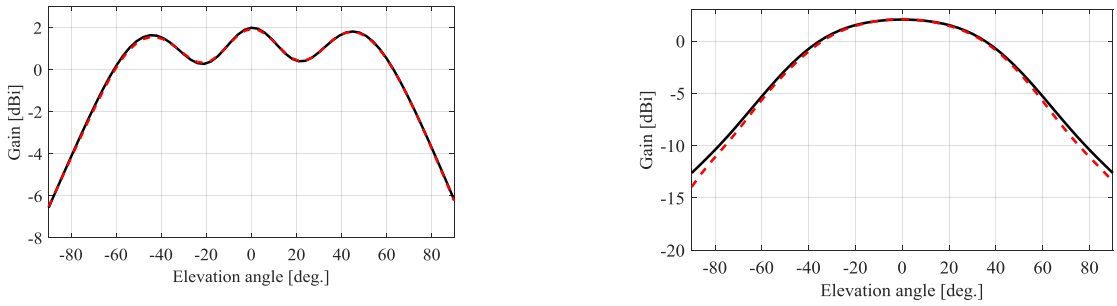


Figure 1: Probe-fed microstrip patch antenna on ground-plane-backed single-layer dielectric substrate.

	G_E (n = 700, n^* = 1220)	G_H (n = 700, n^* = 1220)
%RMSE	1.751	1.750
R	0.9977	0.9958

Table 1. Predictive errors of GP models for gain patterns on test data (n: number of training points; n^*: number of test points)



(a) (b)

Figure 2: Comparison of simulated (——) and predicted (- - -) radiation patterns for microstrip patch antenna test geometry with $D_x = 2.57\lambda_0$, $D_y = 1.85\lambda_0$, $L = 13.52$ mm, $W = 9.16$ mm.
(a) E-plane pattern (b) H-plane pattern.

3.2 Modeling of aperture efficiency ripple in reflector antennas

Consider the offset Gregorian antenna with geometry shown in Fig. 3, which may be designed using the equations given in [9].

Using physical optics and physical theory of diffraction in the commercial code GRASP [6], we simulated the aperture efficiency of a system with D_s = 3.8 m, D_m = 15.5 m, d_g = 1.0 m, θ_e = 50°, fed by an ideal Gaussian feed pattern with a -10 dB edge taper at θ_e, over the frequency range 0.7325 GHz to 1.5475 GHz. The efficiency responses, as function of frequency, typically display rapidly varying oscillations modulated by a slowly varying envelope. For example, the response in Fig. 4 shows four 'periods' of the slowly-varying envelope. The oscillatory response of the aperture efficiency (or directivity anywhere in the vicinity of the main beam) of the system is caused by interference between the direct reflected rays (shown in Fig. 3 at the edges of the reflectors as $\rho_{S1,2} + \rho_{M1,2} + \rho_{A1,2}$) and the diffracted rays from the sub-reflector (shown in Fig. 3 at the edges of the reflectors as $\rho_{D1,2} + \rho_{D1,2}$). The period of the oscillation is dependent on the specific geometry and separation between the dishes. A full description of this mechanism is provided in [9], [10].

A training data set was generated by selecting 28 equally-spaced points within the interval spanning the first 'period' of the slowly-varying envelope, i.e., 0.7325 GHz $\leq f \leq$ 0.95 GHz (Interval 1), as well as 26 points randomly selected from the interval spanning the remaining three periods, i.e., 0.95 GHz $< f \leq$ 1.5475 GHz (Interval 2). We postulated that thoroughly describing the first period by means of training points would enable the regression – via use of the composite covariance function (8) – to capture most of the essential rapidly varying structure of the response and that fewer data points would therefore be necessary for subsequent repetitions along the remainder of the input space (i.e., frequency).

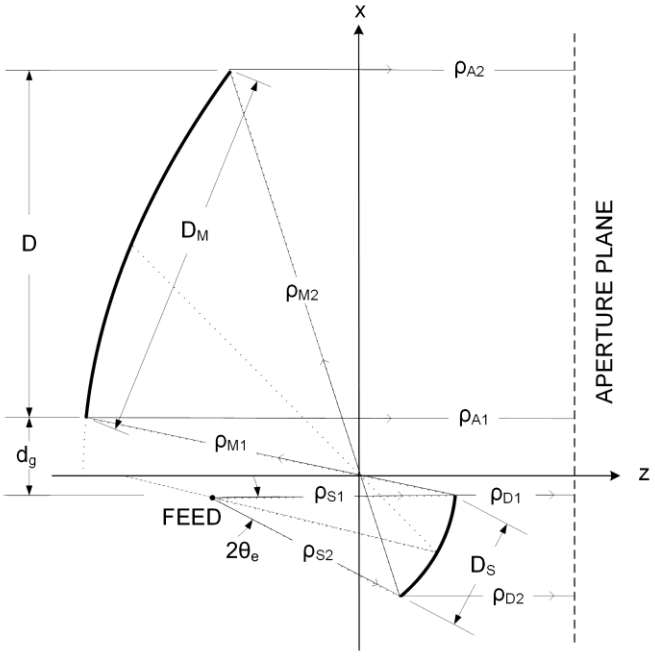


Figure 3. Symmetry plane cut view of a typical offset Gregorian reflector system.

Fig. 4 shows predictive results for GP models trained as described above. In spite of the generally non-uniform distribution of points in Interval 2, with significant stretches of the responses receiving minimal training points in some cases (e.g., the third period in Fig. 4), predictive results were excellent, with normalized root-mean-square-errors (RMSEs) of 0.74%

achieved (RMSEs were normalized to the target range). Similar results were also found for a variety of other distributions of training points, with more details given in [5].

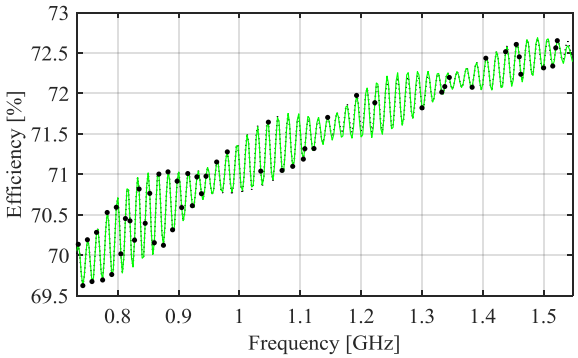


Figure 4. Target (- - -) and predicted responses (—), and training data points (•) for a randomly selected training data set. There were 28 training data points in the interval 0.7325 GHz $\leq f \leq$ 0.95 GHz, and 26 randomly selected training data points in the interval 0.95 GHz $\leq f \leq$ 1.5475 GHz. The test prediction normalized root-mean-square error was 0.74%.

4 CONCLUSIONS

GP regression has been shown to yield very accurate predictions for two diverse antenna modeling problems, namely gain patterns of a microstrip patch antenna subject to finite substrate/ground plane effects; and highly oscillatory aperture efficiency responses of an offset Gregorian reflector system. In the reflector system case, computations were extremely efficient as far as training data requirements are concerned. With respect to the reflector system, results are shown only for a 1-D case here, but work is ongoing on extending the method to larger dimensional spaces, and testing it on more reflector system examples. Applications may include global optimization as well as sensitivity analysis with respect to variations in the mechanical structure of the system – an important problem in high performance systems such as radio telescopes and space antennas which are required to operate in very harsh environmental conditions.

REFERENCES

[1] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[2] J. P. De Villiers, J. P. Jacobs, Gaussian process modeling of CPW-fed slot antennas. *Progress in Electromagnetics Research,* **98**, 233-249, 2009.

[3] J.P. Jacobs, S. Koziel, Two-Stage Framework for Efficient Gaussian Process Modeling of Antenna Input Characteristics. *IEEE Transactions on Antennas and Propagation*, **62**, 706-713, 2014.

[4] J. P. Jacobs, J. P. De Villiers, Gaussian-process-regression-based design of ultrawideband and dual-band CPW-fed slot antennas. *Journal of Electromagnetic Waves and Applications,* **24**, 1763-1772, 2010.

[5] J.P. Jacobs, D.I.L. de Villiers, Gaussian process modeling of aperture efficiency ripple in reflector antennas. *Loughborough Antennas & Propagation Conference (LAPC 2015)*, Loughborough, UK, November 2015.

[6] J.P. Jacobs, Characterization by Gaussian processes of finite substrate size effects on radiation patterns of microstrip antennas. Submitted for publication.

[7] CST Microwave Studio, ver. 2015, CST AG, Bad Nauheimer Str. 19, D-64289 Darmstadt, Germany, 2015.

[8] Bokhari, S.A., Mosig, J.R., and Gardiol, F.E, Radiation pattern computation of microstrip antennas on finite size ground planes. *IEE Proceedings H: Microwaves, Antennas and Propagation*, **139**, 278 – 286, 1992.

[9] C. Granet, Designing classical offset Cassegrain or Gregorian dualreflector antennas from combinations of prescribed geometric parameters. *IEEE Antennas and Propagation Magazine*, **44**, 114–123, 2002.

[10] D.I.L. de Villiers, Prediction of Aperture Efficiency Ripple in Clear Aperture Offset Gregorian Antennas. *IEEE Transactions on Antennas and Propagation,* **61**, 2457 - 2465, 2013.

[11] TICRA, GRASP10, Version 10.0.3, Copenhagen, Denmark. [Online]. Available: http://www.ticra.com.

[12] D.I.L. de Villiers, Gain ripple in small offset Gregorian antennas. *IEEE Int. Symp. Antennas Propag. (APSURSI 2011)*, July 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

ACCELERATION OF MESH-BASED PHYSICAL OPTICS FOR ELECTROMAGNETIC SCATTERING ANALYSIS

Dao P. Xiang[1], **and Matthys M. Botha**[2]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University
Private Bag X1, Matieland 7602, Stellenbosch, South Africa
e-mail: {17745349,mmbotha}@sun.ac.za

Keywords: Computational electromagnetics, Electromagnetic scattering, Physical optics (PO), Shadowing determination, Recursive update, Multilevel fast multipole method (MLFMM).

Abstract. *This work is concerned with scattering analysis of electrically large, conducting objects, using the physical optics (PO) approximation. Such analysis is important for radar cross section (RCS) calculations. Scattering objects are represented by a mesh of sub-wavelength sized triangle elements. Upon the mesh, the induced surface current density is represented by standard divergence-conforming basis functions. This is the mesh-based PO approach. It can be applied to take only a single reflection off the scatterer into account, as well as to take multiple, internal reflections into account. Both plane wave and point source illuminations are considered. The paper discusses computational challenges with regards to the fast and robust implementation of mesh-based PO analysis methods. These challenges are shadowing determination, in which the visibility of mesh elements with respect to source locations must be determined, as well as the fast evaluation of observed fields. A multi-level, buffer-based algorithm for efficient, single-reflection shadowing determination is outlined. It is also shown how the recursive field calculations required in the multiple-reflection PO, can be accelerated with the aid of the multilevel, fast multipole method (MLFMM), when the scattering object does not support any internal shadowing. Preliminary results illustrate the current status of this ongoing work on the acceleration of mesh-based PO analysis methods.*

1 INTRODUCTION

This work is concerned with scattering analysis of electrically large, conducting objects, using the physical optics (PO) approximation [1]. Such analysis is important for radar cross section (RCS) calculations. Scattering objects are represented by a mesh of sub-wavelength sized triangle elements. Upon the mesh, the induced surface current density is represented by standard divergence-conforming basis functions also used in the Method of Moments (MoM). This is the mesh-based PO approach [2, 3]. This approach can be applied to take only a single reflection off the scatterer into account (when illuminated by a plane wave or point source [Hertzian dipole]), as well as to take multiple, internal reflections into account. In the multiple-reflection case, the PO approximation is applied iteratively. In the interest of brevity, the paper will only be concerned with the standard RWG-type of mixed first-order, edge-associated basis functions, as employed in e.g. [2].

The paper starts off by briefly reviewing the above mesh-based PO analysis methods in Section 2. The computational challenges for the efficient (fast and robust) computer implementation of these methods are also discussed in Section 2. These include shadowing determination and the fast evaluation of observed fields. The main contribution of this paper follows in Sections 3, 4 and 5. Solutions are proposed to some of these challenges. Preliminary results are included, to illustrate the current status of this ongoing work on the acceleration of mesh-based PO analysis methods.

2 REVIEW OF MESH-BASED PHYSICAL OPTICS FOR PEC SCATTERERS

2.1 The PO approximation

Firstly, it is noted that for a magnetic field $\boldsymbol{H}^{\text{inc}}$, incident upon an infinite perfect electrically conducting (PEC) half space, the exact solution of the induced surface current is

$$\boldsymbol{J}_s = 2\hat{n} \times \boldsymbol{H}^{\text{inc}} \tag{1}$$

where $\hat{n}$ is the unit normal vector at the interface, pointing away from the PEC region. This simple result is the inspiration for the PO approximation [1, 4], which states that the surface current on a PEC object can be approximated in terms of the incident magnetic field and the outward-pointing unit vector to the surface, as follows:

$$\boldsymbol{J}_s = \begin{cases} 2\hat{n} \times \boldsymbol{H}^{\text{inc}} & \text{for visible surfaces} \\ 0 & \text{for shadowed surfaces.} \end{cases} \tag{2}$$

The status of a location as being visible or in shadow, is established by determining whether there is line-of-sight visibility to the source, or not. For an incident plane wave as source, this means visibility towards infinity, along the negative of the incident direction. For a point source, visible status means an unobstructed view of the source point, from the given position on the scatterer surface.

2.2 The single-reflection PO method

The single-reflection PO method (SRPO) employs the PO approximation directly, to establish an approximate solution for the induced current density on a PEC scatterer. In order to achieve this, two tasks must be performed: (i) shadowed and visible parts of the scatterer must be identified; and (ii) in the visible regions, the known incident field must be mapped to a description of the surface current. The second of these tasks is quite straightforward in the

mesh-based PO context: a local projection procedure is used to obtain the RWG basis function coefficient values, such that they best approximate the tangential components of the incident field according to (2) (see e.g. [2]). This is an $\mathcal{O}(N)$ task, where N denotes the number of mesh elements. The first task is much more challenging, with regards to computational efficiency.

2.3 The multiple-reflection PO method

The first step in the multiple-reflection PO method (MRPO) is to calculate the SRPO solution. This solution is denoted by $\boldsymbol{J}^{(0)}$. It is obtained in the illuminated regions of the scatterer's surface, according to the PO approximation and the local projection procedure:

$$\boldsymbol{J}^{(0)}(\boldsymbol{r}) = 2\hat{n} \times \boldsymbol{H}^{i}(\boldsymbol{r}). \tag{3}$$

In the case where multiple reflections are important, the cumulative radiation from parts of this current distribution is significant towards other parts of the object. Additional currents are induced according to the PO approximation, which represent a second reflection. In general, the current after l additional reflections is found by way of the following recursive expression [5]:

$$\boldsymbol{J}^{(l)}(\boldsymbol{r}) = \boldsymbol{J}^{(0)}(\boldsymbol{r}) + 2\hat{n} \times \fint_{S} \delta(\boldsymbol{r}')\boldsymbol{J}^{(l-1)}(\boldsymbol{r}') \times \hat{R}\frac{e^{-jkR}}{4\pi R}\left(jk + \frac{1}{R}\right) dS' \tag{4}$$

where $\boldsymbol{R} = \boldsymbol{r} - \boldsymbol{r}'$, $R = |\boldsymbol{R}|$, $\hat{R} = \boldsymbol{R}/R$; $\boldsymbol{r}'$ and $\boldsymbol{r}$ relate to the position vectors of the source and observation points, respectively, on the scatterer's surface. The scatterer's surface is denoted by S. The symbol $\fint$ represents the principle value integral. The factor $\delta(\boldsymbol{r}')$ is the shadowing coefficient, which takes on the value of either 1 or 0, depending on whether the current source point is visible to the observation point, or not.

In other words, the currents re-radiate iteratively, to excite a new set of PO currents at each reflection. The iteration process can be stopped after a pre-specified number of reflections based on the number of expected significant reflections supported by a given model, or until an adequately converged solution is obtained.

2.4 Computational challenges

For the mesh-based SRPO, the main challenge is to determine the shadowing status of each mesh edge, with respect to the illuminating source. A naïve approach would involve considering each edge in turn, and testing all elements for shadowing relative to the source. This would require $\mathcal{O}(N^2)$ operations ($\mathcal{O}(N)$ tests per edge, of which there are a total of $\mathcal{O}(N)$). Given that the other main task is projection, the cost of which is $\mathcal{O}(N)$, shadowing is clearly the bottleneck. This cost can be reduced by making use of the field-of-view buffer concept [6, 7, 3]. A buffer plane is defined upon an eikonal surface of the incident wave. This plane is subdivided into buffer boxes and each mesh element is listed in those boxes into which parts of it project. To test the shadowing status of a given point, it is also projected into the buffer to determine the box it falls into. Subsequently, only shadowing checks need to be performed against those elements which are listed for that particular buffer box. The present authors have been working on efficient shadowing algorithms which employ this general concept [8, 9]. The present status of this ongoing work is reviewed in Sections 3 and 4.

For the mesh-based MRPO, there are two main challenges. The first challenge is the calculation of subsequent incident fields due to the induced surface current, by way of (4). When using standard integration, the cost of this step is $\mathcal{O}(N^2)$, since at every mesh edge, the field must be calculated due to all other mesh elements. This step could be accelerated with a fast field

calculation method. Indeed, this has been done using the fast far field approximation (FaFFA) [10, 11], which reduces the cost to $\mathcal{O}(N^{5/3})$. However, seeing that field calculation at a specified set of points can be viewed as a Method of Moments (MoM) type of matrix-vector product, the multilevel, fast multipole method (MLFMM) [12] could be used to reduce the cost of this step even further, to $\mathcal{O}(N \log N)$. Ongoing work in this direction, is reviewed in Sections 5.

The second challenge for the mesh-based MRPO, is efficient shadowing determination. The shadowing status of all mesh edges, as observed from every mesh edge projection point, is required. A naïve approach would require $\mathcal{O}(N^2)$ operations for every projection point (as explained above) and hence $\mathcal{O}(N^3)$ operations in total. Furthermore, the shadowing data would require $\mathcal{O}(N^2)$ storage space, except if calculated on-the-fly. To reduce this cost to a practically feasible scale, is a challenging task. There are however various problems of practical interest where all projection points are visible to all induced current sources. In such cases the shadowing information is trivial and acceleration of the field calculations do suffice. In this paper, the application focus of the MRPO is on such cases.

3 ACCELERATION OF SHADOWING DETERMINATION FOR THE SRPO WITH PLANE WAVE ILLUMINATION

3.1 Method

In this case, the eikonal surfaces are flat planes transverse to the direction of propagation of the incident wave. The buffer plane is chosen to be such a surface. The steps followed in the shadowing determination algorithm are as follows:

1. Triangles are projected into the buffer plane and the overall buffer bounding box is established, as well as bounding boxes for each projected element.

2. The objective of the algorithm is to separate triangles into small groups in the buffer plane, according to a buffer plane subdivision. Therefore, divide the plane into buffer boxes of which the size is proportional to the average mesh size. If a box is found which contains an excessive number of elements, then that box is again subdivided according to the same procedure. This yields a multi-level subdivision, capable of handling strongly inhomogeneous meshes as well. In case of a homogeneous mesh, only the first-level subdivision is typically necessary.

3. Now project each edge mid-point into the buffer and do shadowing testing against only those elements listed in the specific buffer box. This yield the visibility status of each corresponding, edge-associated basis function.

Above, some technical details of the algorithm are omitted in the interest of brevity. Also, the above description assumes that no grazing incidence is present. Grazing incidence occurs when the propagation direction of the incident plane wave is close to tangential to certain constituent surfaces of the scatterer. In such cases many elements may end up being listed in individual buffer boxes. The grazing triangles must then be handled with a custom procedure, similar to the one noted in [8].

3.2 Results

Figure 1 shows examples of the meshes used for shadowing evaluation testing. All meshes represent the same simple model, but will serve to illustrate the capabilities of the algorithm. The series of meshes are of a sphere which is inhomogeneously meshed and progressively

refined, as explained in the figure. These progressively refined meshes are used to generate the shadowing determination runtime results shown in Figure 2, as a function of the number of mesh elements. On the same computational platform, results were also generated with the commercial mesh-based PO solver available as part of the FEKO software suite [13]. The direction of incidence is set to be aligned with the point around which the mesh is refined. The results show that the proposed, multilevel buffer subdivision scheme is significantly more efficient than the commercial solver. This is due the effective separation of mesh elements in the buffer, through appropriate localized subdivision. The observed time scaling for the proposed method is of $\mathcal{O}(N)$.

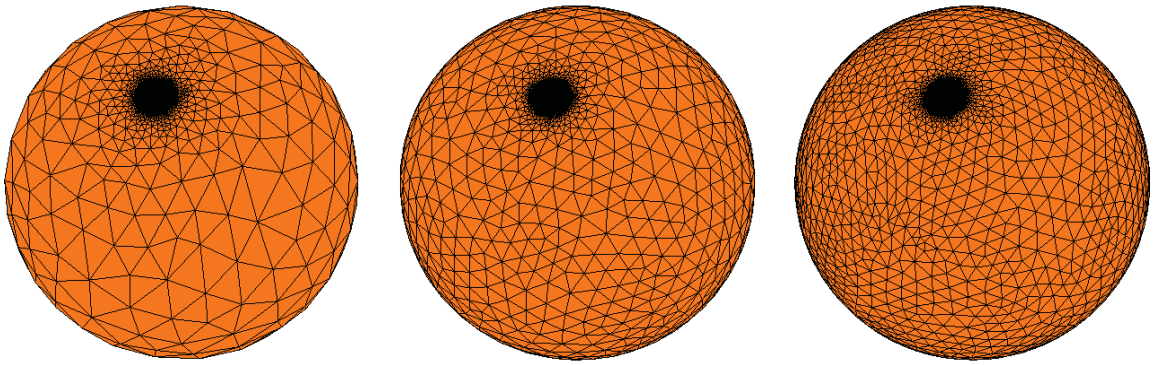


Figure 1: Progressively refined meshes of a conducting sphere. The sphere has a radius of $1\,\mathrm{m}$. The meshes are all inhomogeneous with a relative element size reduction factor of 100, around a single point.

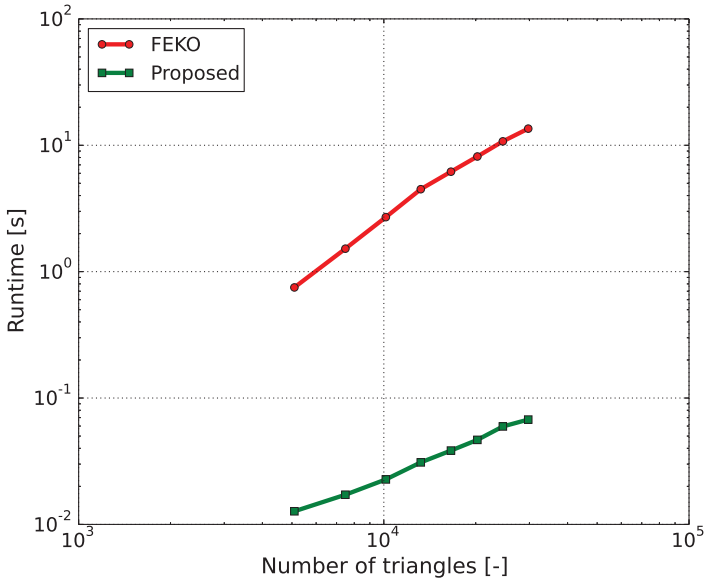


Figure 2: Shadowing determination runtime as a function of the number of mesh elements, for plane wave illumination of the inhomogeneous meshes shown in Figure 1. The proposed method is compared to a commercial solver.

4 ACCELERATION OF SHADOWING DETERMINATION FOR THE SRPO WITH POINT SOURCE ILLUMINATION

4.1 Method

In this case, the eikonal surfaces are spheres centred around the illuminating point source and the spherical coordinate system variables (θ, ϕ) are used to describe point locations in the buffer. Otherwise, the general procedure is the same as discussed above for the plane wave case. It is important to note that even in the case of a homogeneous mesh, the projected representation in the buffer plane is often inhomogeneous, due to the specific properties of the spherical coordinate system. This is a further important motivating factor for the recursive buffer subdivision approach. In the interest of brevity, a discussion on the case of grazing incidence for point source illumination will be deferred to a future publication.

4.2 Results

The set of progressively refined meshes described in Figure 1 is again used. The point source is placed normally above the point on the sphere around which the refinement is centred, at a height of $10^{-2}\,\mathrm{m}$. Figure 3 shows the shadowing determination runtime as a function of the number of mesh elements. Again, the results show that the proposed, multilevel buffer subdivision scheme is much more efficient than the commercial solver and that the cost scales as $\mathcal{O}(N)$.

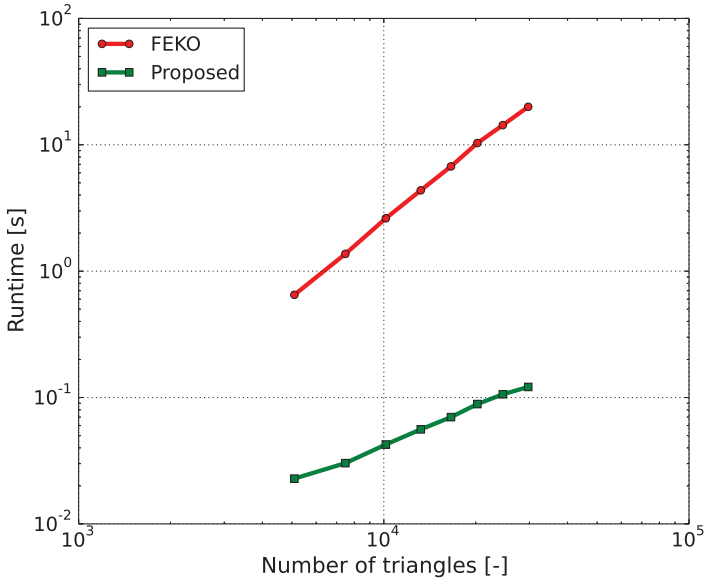


Figure 3: Shadowing determination runtime as a function of the number of mesh elements, for point source illumination of the inhomogeneous meshes shown in Figure 1. The proposed method is compared to a commercial solver.

5 MLFMM-BASED ACCELERATION OF INDUCED CURRENT CALCULATION IN THE MRPO

5.1 Method

To calculate the next solution according to the iterative MRPO expression (4), the observation point $\boldsymbol{r}$ must be placed at each edge mid-point in the mesh. Collectively, this can be expressed as a system of linear equations:

$$I^{(l)} = I^{(0)} + AI^{(l-1)} \tag{5}$$

where $I^{(l)}$ represents the vector of current coefficients after l additional reflections and $I^{(0)}$ represents the vector of current coefficients corresponding to the initial, single-reflection PO current. The dimension of square matrix A is $\mathcal{O}(N)$. The cost of conventionally evaluating the matrix-vector product $AI^{(l-1)}$ is thus $\mathcal{O}(N^2)$.

The matrix A in (5) can be split up into a sparse near-interaction matrix and the remaining far-interaction part, per the specifications of the MLFMM algorithm [12]:

$$I^{(l)} = I^{(0)} + \left(A^{\text{near}} + A^{\text{far}}\right) I^{(l-1)}. \tag{6}$$

The cost of the matrix-vector product $A^{\text{near}} I^{(l-1)}$ is $\mathcal{O}(N)$. Using the MLFMM, the matrix-vector product $A^{\text{far}} I^{(l-1)}$ has cost $\mathcal{O}(N \log N)$ (see e.g. [12] for further details). This can be compared with the scaling achieved using FaFFA acceleration, which as noted already, is given as $\mathcal{O}(N^{5/3})$ in [11].

5.2 Results

Consider the trihedral structure shown in Figure 4. It is illuminated by a plane wave with incident angles $\theta^{\text{inc}} = 45^\circ$ and $\phi^{\text{inc}} = 45^\circ$. It is homogeneously meshed with mesh size of $\lambda_0/10$, where λ_0 denotes the free space wavelength. On the illuminated side, all triangles are visible to each other, thus shadowing determination is trivial and consequently, the MRPO can be accelerated with the MLFMM in this case. Three reflections are incorporated with the MRPO ($l = 2$). Figure 5 shows how the runtime for the solution scales as a function of the number of mesh elements. Different data points are obtained by varying the excitation frequency (and thus wavelength) and remeshing (the frequency range used is $0.5\,\text{GHz}$ to $2.25\,\text{GHz}$). Clearly, the standard MRPO scales as $\mathcal{O}(N^2)$, due to the conventional evaluation of the matrix-vector product in (5). The MLFMM-accelerated MRPO has much lower runtime and as N becomes larger, the relative saving becomes increasingly significant. Note that $\mathcal{O}(N \log N)$ behaviour cannot be clearly observed in these preliminary accelerated results. This is ongoing work and further tests are still being conducted.

6 CONCLUSIONS

Computational challenges with regards to the fast implementation of mesh-based PO methods for scattering analysis of conducting objects, were discussed. Both single-reflection and multiple-reflection formulations were considered. These challenges are efficient shadowing determination and efficient calculation of observed fields. A multi-level, buffer-based algorithm for efficient, single-reflection shadowing determination was outlined and results were presented for a strongly inhomogeneous mesh. The results exhibit linear time scaling and are superior to the runtimes achieved with the commercial PO solver in FEKO. It was further shown how the recursive field calculations required in the MRPO, can be accelerated with the aid of the

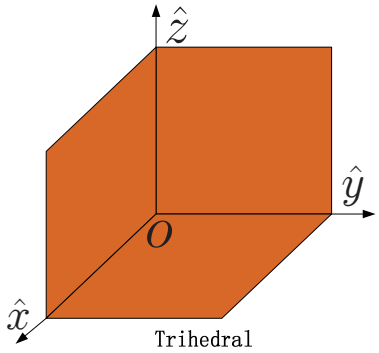


Figure 4: Conducting plate, trihedral model. The plate dimensions are all $1\,\mathrm{m}$.

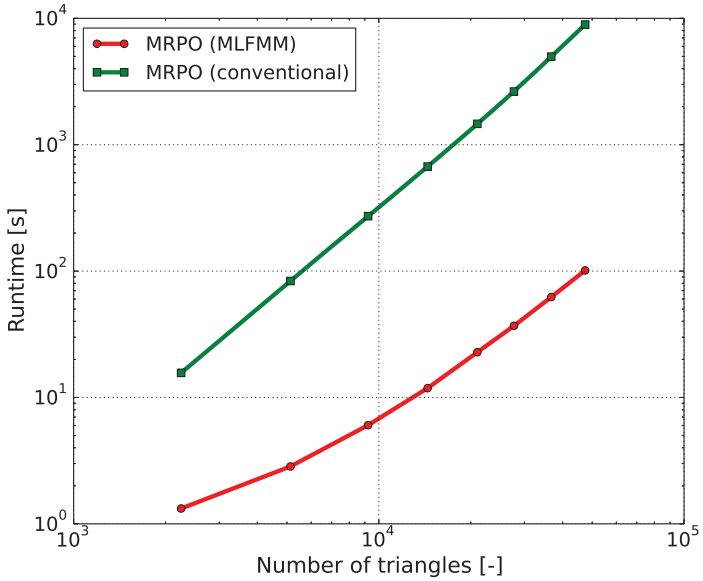


Figure 5: Runtime as a function of the number of mesh elements, for the conventional and MLFMM-accelerated MRPO implementations.

MLFMM. Preliminary results demonstrate significantly reduced runtime, when the scattering object does not support any internal shadowing.

ACKNOWLEDGMENT

This work was funded by Altair Development S.A. (Pty) Ltd, developers of the EM simulation software suite FEKO.

REFERENCES

[1] N. Bojarski, "A survey of the physical optics inverse scattering identity," *IEEE Trans. Antennas Propag.*, vol. 30, no. 5, pp. 980–989, 1982.

[2] U. Jakobus and F. M. Landstorfer, "Improved PO-MM hybrid formulation for scattering

from three-dimensional perfectly conducting bodies of arbitrary shape," *IEEE Trans. Antennas Propagat.*, vol. 43, no. 2, pp. 162–169, 1995.

[3] F.-T. Zha, S.-X. Gong, Y.-X. Xu, Y. Guan, and W. Jiang, "Fast shadowing technique for electrically large targets using z-buffer," *Journal of Electromagnetic Waves and Applications*, vol. 23, no. 2-3, pp. 341–349, 2009.

[4] W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*. New York: John Wiley and Sons, 2nd ed., 1998.

[5] F. Obelleiro-Basteiro, J. Luis Rodriguez, and R. J. Burkholder, "An iterative physical optics approach for analyzing the electromagnetic scattering by large open-ended cavities," *IEEE Trans. Antennas Propag.*, vol. 43, no. 4, pp. 356–361, 1995.

[6] E. A. Haines and D. P. Greenberg, "The light buffer: A shadow-testing accelerator," *IEEE Computer Graphics and Applications*, pp. 6–16, September 1986.

[7] J. M. Rius, M. Ferrando, and L. Jofre, "GRECO: Graphical electromagnetic computing for RCS prediction in real time," *IEEE Antennas and Propagation Magazine*, vol. 35, no. 2, pp. 7–17, 1993.

[8] D. P. Xiang and M. M. Botha, "Aspects of efficient shadowing calculation for physical optics analysis of meshed objects," in *Proceedings of the International Conference on Electromagnetics in Advanced Applications (ICEAA '14)*, pp. 492–495, Aug 2014. Aruba, Dutch Antilles.

[9] D. P. Xiang and M. M. Botha, "Further progress with fast and reliable shadowing determination for mesh-based po analysis," in *Proceedings of the International Conference on Electromagnetics in Advanced Applications (ICEAA '15)*, pp. 958–961, Sept 2015. Torino, Italy.

[10] C.-C. Lu and W. C. Chew, "Fast far-field approximation for calculating the RCS of large objects," *Micro. Opt. Tech. Lett.*, vol. 8, no. 5, pp. 238–241, 1995.

[11] R. J. Burkholder, "A fast and rapidly convergent iterative physical optics algorithm for computing the RCS of open-ended cavities," *Applied Computational Electromagnetics Society Journal*, vol. 16, no. 1, pp. 53–60, 2001.

[12] W. C. Chew, E. Michielssen, J. Song, and J. Jin, *Fast and efficient algorithms in computational electromagnetics*. Artech House, Inc., 2001.

[13] Altair Engineering Inc, "FEKO User's Manual Suite 7.0," August 2014. http://www.feko.info.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DESIGN OF AXIALLY SYMMETRIC POWER COMBINERS USING SURROGATE BASED OPTIMIZATION

Ryno D. Beyers[1], **Dirk I. L. de Villiers**[1]

[1]University of Stellenbosch
Cnr Banhoek Road & Joubert Street, Stellenbosch, South Africa
e-mail: rdb@sun.ac.za, ddv@sun.ac.za

Keywords: Power Combiners, Coaxial Combiners, Conical Combiners, Circuit Modelling, Surrogate Modelling.

Abstract. *The design and optimization of axially symmetric power combiners usually rely on computationally expensive full-wave simulations. Circuit models are often used to reduce the computational cost, however, full-wave parameter sweeps or optimization are usually still required. Surrogate based optimization utilizing circuit models offers an effective means to reduce the computational cost, since the circuit model can be aligned to the full-wave model, usually resulting in fewer full-wave simulations being required. This paper presents some of the recent advances in using Space Mapping, which is a surrogate based optimization technique, for the design and optimization of axially symmetric power combiners.*

1 INTRODUCTION

Power combiners and dividers have numerous applications and are frequently used in communications and radar systems as part of power amplifier and antenna configurations. Since the introduction of the modern personal computer and electromagnetic analysis software, the design of power combiners/dividers, especially axially symmetric types, has been facilitated by full-wave simulation [1–9]. However, the relatively high computational cost of full-wave simulations limits the amount of the parameter space that can be searched within a reasonable time and cost, thereby possibly limiting the performance that can be attained.

Circuit modelling can be used to various extents to accelerate the design process. For example, circuit modelling can be used for parts of the device with known circuit models, and the rest modelled by full-wave simulation, as is done in [7, 8]. This method reduces the size of the full-wave problem being solved, but it might not be straight-forward to divide the problem into appropriate parts, and full-wave parameter sweeps or optimization are still required to find optimal parameter values. Alternatively, circuit models for the entire device can be constructed, in which case full-wave parameter sweeps or optimization would likely be needed to relate physical dimensions to circuit element values [1, 3, 6]. Empirical data can be used to extract equations for the conversion of dimensions into element values, but these would also require full-wave parameter sweeps in addition to having a fixed and limited range of parameter values where adequate accuracy can be achieved [9].

The use of equivalent circuit models can be exploited, perhaps most effectively, by using Space Mapping [10,11], which is a surrogate based optimization method. By using this method, the optimization burden is moved from the accurate but slow full-wave simulation and placed on the circuit model, which is less accurate, but much faster to evaluate. The use of Space Mapping for the design of conical transmission line power combiners was recently demonstrated in [12], where an optimal design was obtained after only four iterations, requiring only four full-wave simulations. This paper presents guidelines for the Space Mapping based design and optimization of axially symmetric power combiners supporting transverse electromagnetic (TEM) mode propagation. A specific design example of a coaxial line power combiner with an improved fractional bandwidth of $133\ \%$ is presented.

2 EQUIVALENT CIRCUIT MODEL DESCRIPTION

Axially symmetric power combiners generally allow for a large number of devices to be combined with lower losses compared to other power combiner types with the same number of input ports [1, 13]. Non-resonant axially symmetric combiners typically have the added advantage of being able to support TEM modes, and can therefore offer high output port return losses over relatively wide bandwidths when the input ports are excited symmetrically. For these types of combiners the output port is generally referred to as the central port and the input ports as peripheral ports. The design of such a combiner is mostly focussed on obtaining a well-matched transition from the peripheral ports into the combining structure, as well as from the combining structure to the central port. Since ideal transmission lines can be used to model TEM transmission lines, the only non-trivial parts to be modelled are where discontinuities or gradual changes between different transmission line sizes or types occur in the physical structure of the combiner. Examples of some of these discontinuities and gradual changes and how they can be modelled can be found in [1, 5, 8, 14–16]. In most cases, some form of equivalent circuit model can be deduced or found in literature. The requirements of equivalent models used for design or optimization with Space Mapping are that they must be much faster to evaluate than

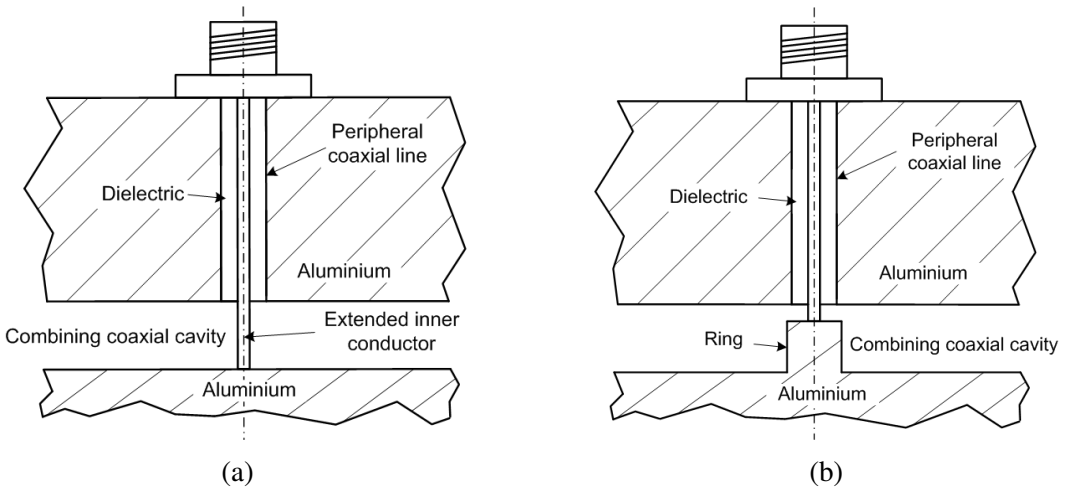


Figure 1: The uncompensated (a) and compensated (b) peripheral coaxial lines used for the input ports of a coaxial power combiner.

the full-wave models and that they can be linked to physical attributes of the structure, similar to how a shunt capacitance is linked to a step discontinuity in a coaxial line. As an example, it will be shown how a circuit model of a coaxial combiner can be obtained. This model will be used for the design example in Section 4.

2.1 Equivalent circuit model of a coaxial combiner

The configuration of the combiner must first be established before it is possible to construct an equivalent circuit model. An appropriate configuration will consist of parts that allow for well-matched transmission, and with equivalent circuit models that are either known or are easy to deduce. It has been demonstrated that the inductance of extended center conductor pins of peripheral coaxial lines (see Fig. 1a), which would normally adversely affect transmission, can be effectively compensated for by using a ring (see Fig. 1b) in the oversized combining coaxial line [6]. The ring-compensated peripheral port transition is therefore an appropriate candidate, since it allows for a well-matched transition, and it is relatively straight-forward to deduce an equivalent circuit model that corresponds well with the physical attributes of the transition, as will be shown in Section 2.1.1. The oversized combining coaxial line can simply be modelled by using ideal transmission lines and shunt capacitances where step discontinuities are introduced for impedance matching purposes. For this design the outer conductor radius will be kept constant and steps will only occur in the inner conductor. The only remaining part of the combiner to be modelled is a transition from the oversized combining coaxial line to the smaller central output port, usually with dimensions corresponding to a standard coaxial connector type. These types of transitions can be physically tapered coaxial lines with a constant impedance [6], or in some cases the impedance can also be tapered, possibly making more efficient use of the overall size of the combiner [3]. For this design, an impedance tapered transition will be used, for which it is straight-forward to deduce an equivalent circuit model, as will be shown in Section 2.1.2. The physical layout of the chosen combiner configuration is shown in Fig. 2, and its equivalent circuit model will now be developed.

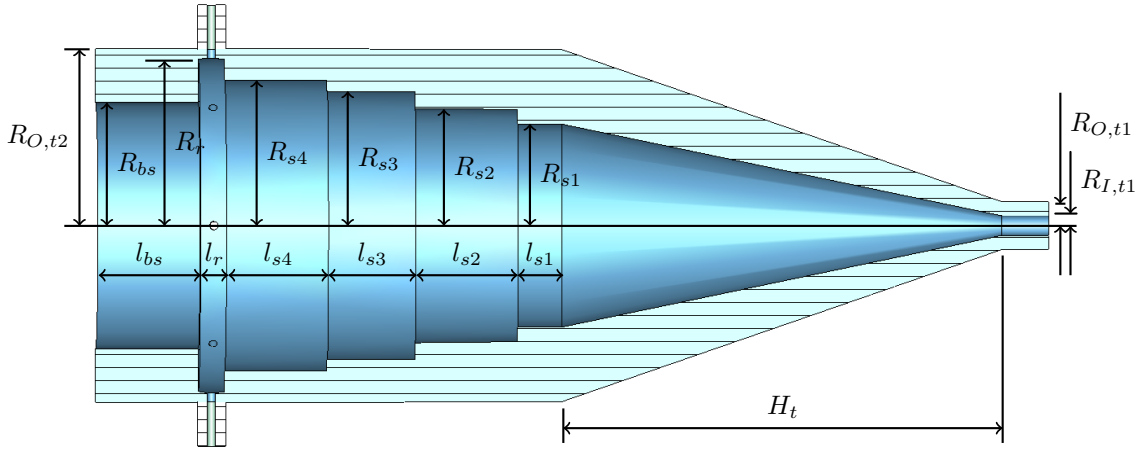


Figure 2: The physical layout of the coaxial combiner with the chosen configuration.

2.1.1 Peripheral coaxial line transition

The ring-compensated peripheral transition proposed in [6] and shown in Fig. 1b shortens the extended center conductor pins of the peripheral coaxial lines, thereby reducing their inductance. It also introduces two step discontinuities, modelled using shunt capacitances, on either side of the peripheral ports inside the oversized combining coaxial line. An additional effect, not taken into account in [6], is that there will be a change in characteristic impedance of the oversized combining coaxial line due to the holes that are made for the outer conductors of the peripheral coaxial lines. However, this perturbance can simply be modelled by using two short transmission lines, one on either side of the peripheral coaxial lines, with appropriately adjusted impedances. The resulting equivalent circuit model is shown in Fig. 3.

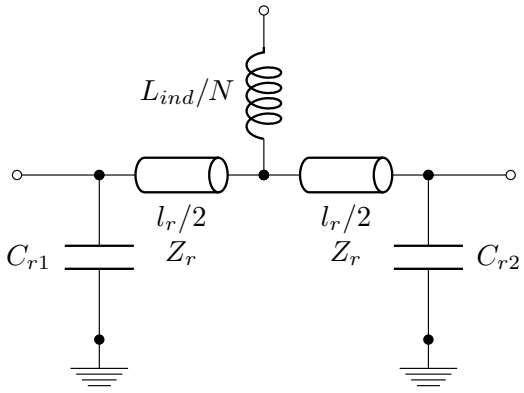


Figure 3: The equivalent circuit model of the peripheral transition.

Note that all circuit element values can be calculated directly from physical dimensions, except for L_{ind} and Z_r, which will be implicit parameters in the Space Mapping based optimization. The length l_r is simply equal to the length of the ring, and the shunt capacitances, C_{r1} and C_{r2}, can be determined by using the information presented in [14]. The values of C_{r1} and C_{r2} can in general be different, since the coaxial line sections on either side of the ring can have different inner conductor radii.

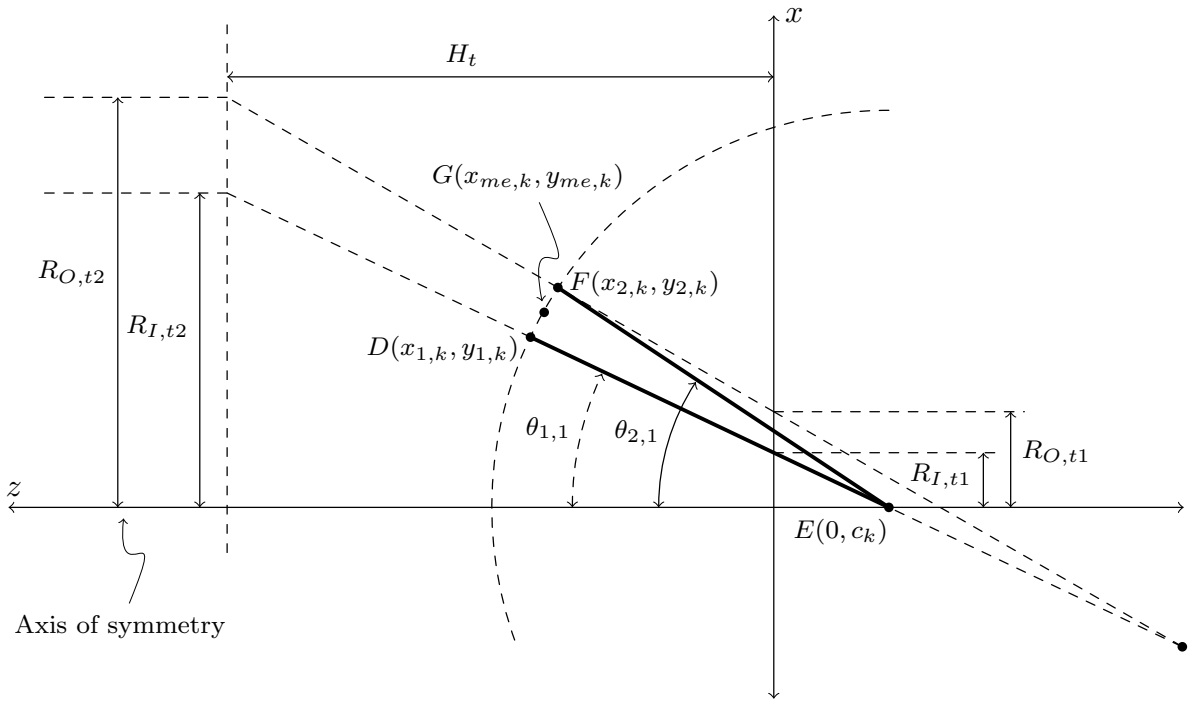


Figure 4: The impedance tapered transition between the central output coaxial line and the oversized combining coaxial line. Note that the figure is rotationally symmetric around the z-axis.

2.1.2 Central impedance tapered coaxial line transition

Even though the central port transition is usually referred to as a tapered coaxial line, it more closely resembles a conical transmission line, and can likely be more accurately modelled as such. A conical transmission line typically consists of two conductors that intersect on the axis of rotational symmetry, forming a transmission line with a constant impedance. An impedance tapered conical line with its conductors tapered at a constant rate (therefore containing no curves, as shown in Fig. 4) would have conductors that do not intersect on the axis of symmetry, and the standard formula for calculating the impedance of a conical transmission line cannot be used. Instead, the tapered transmission line can be approximated by dividing one of the conductors (the lower one, numbered with subscript 1, is chosen here) into K sections and using the simple algorithm described below, with reference to Fig. 4, to determine the impedance and length of each (k-th) section:

1. Draw a line (see line DE) tangential to the conductor from the current position, $D(x_{1,k},y_{1,k})$, to the axis of symmetry, $E(0,c_k)$.

2. Draw a circle centred at $E(0,c_k)$, the intersection of line DE with the axis of symmetry, with a radius equal to the distance from the circle origin to the current position on the conductor, $D(x_{1,k},y_{1,k})$.

3. Find the point of intersection between the circle and the other conductor, $F(x_{2,k},y_{2,k})$.

4. Draw a line (see line FE) from the point of intersection on the second conductor to the circle origin.

5. Use the angles of the two lines, $\theta_{1,k}$ and $\theta_{2,k}$, to calculate the approximate impedance of the current section using the formula for a conical transmission line:

$$Z_k = 60\log\left[\frac{\cot(\theta_{1,k}/2)}{\cot(\theta_{2,k}/2)}\right]. \tag{1}$$

6. Estimate the transmission length of the current section by computing the distance from the mean of the current two points on the conductors, $G(x_{me,k},y_{me,k})$ to the mean of the next two points on the conductors, $(x_{me,k+1},y_{me,k+1})$.

The approximated response of the transition can now be obtained by constructing a circuit with K cascaded ideal transmission lines with impedances and lengths as calculated using the above algorithm.

The accuracy of the circuit model is tested by considering a few different transitions and comparing the circuit model response with full-wave simulations. Fig. 5 shows the approximated impedance profiles of the transitions generated using the parameter values given in Table 1. Transitions 1 and 2 have standard 7/16 connector dimensions for the smaller coaxial line, and Transition 3 has standard N-type dimensions. Fig. 6 shows good agreement between the full-wave and simulated responses of the all the considered transitions.

Transition 1			Transition 2			Transition 3		
Parameter	Value	Units	Parameter	Value	Units	Parameter	Value	Units
$R_{I,t1}$	3.5	mm	$R_{I,t1}$	3.5	mm	$R_{I,t1}$	1.52	mm
$R_{O,t1}$	8	mm	$R_{O,t1}$	8	mm	$R_{O,t1}$	3.5	mm
$R_{I,t2}$	42	mm	$R_{I,t2}$	52	mm	$R_{I,t2}$	20.5	mm
$R_{O,t2}$	60	mm	$R_{O,t2}$	60	mm	$R_{O,t2}$	40	mm
H_t	120	mm	H_t	300	mm	H_t	180	mm

Table 1: Parameter values for the three transitions used to test the equivalent circuit model of the impedance tapered transition.

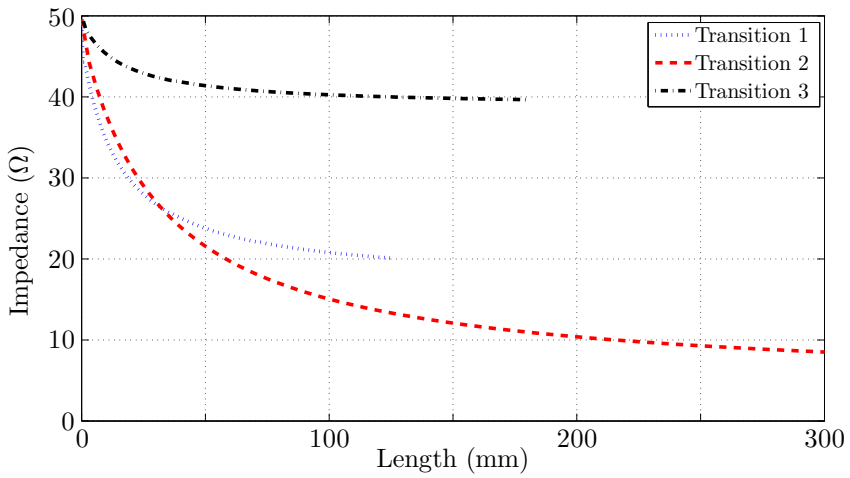


Figure 5: The approximated impedance profiles of the three transitions.

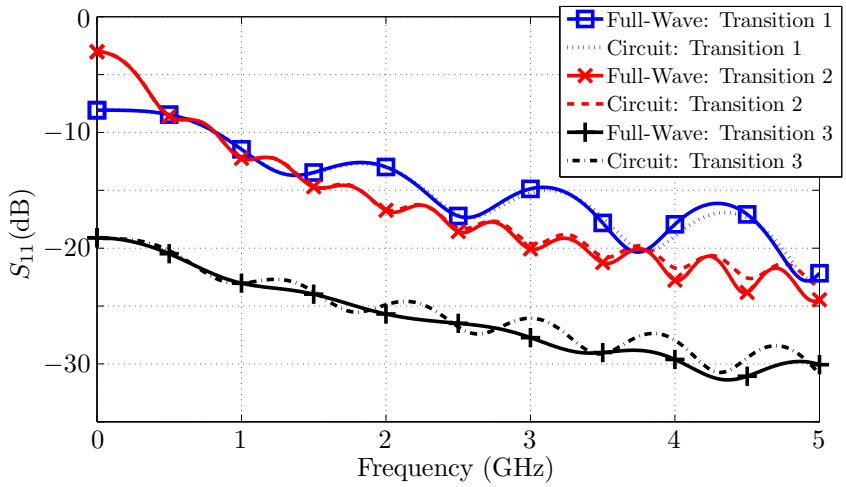


Figure 6: The responses of the circuit model and full-wave simulations.

2.1.3 Complete circuit model

The equivalent circuit models for all parts of the chosen combiner configuration have now been obtained, and can be used to model the entire combiner shown in Fig. 2. The resulting circuit is shown in Fig. 7. The transmission line lengths l_{s1}, l_{s2}, l_{s3}, l_{s4}, l_r, and l_{bs} can be used directly from the physical dimensions of the combiner. The impedances Z_{s1}, Z_{s2}, Z_{s3}, Z_{s4}, and Z_{bs} can be calculated from the inner and outer radii of the respective parts of the combiner using:

$$Z = 60\log\left[\frac{R_{\text{outer}}}{R_{\text{inner}}}\right]. \tag{2}$$

The step capacitances C_{s12}, C_{s23}, and C_{s34} can be calculated using the same method as for C_{r1} and C_{r2} by using the information in [14]. Note that the transmission length of the tapered section, l_t, is not equal to the parameter H_t shown in Figs. 2 and 4. The transmission length l_t and impedance profile Z_t of the central impedance tapered coaxial line transition can be determined from the physical dimensions of the combiner by using the algorithm described in Section 2.1.2. The only parameters with unknown values at this point are Z_r and L_{ind}, which will be implicit parameters in the Space Mapping based optimization.

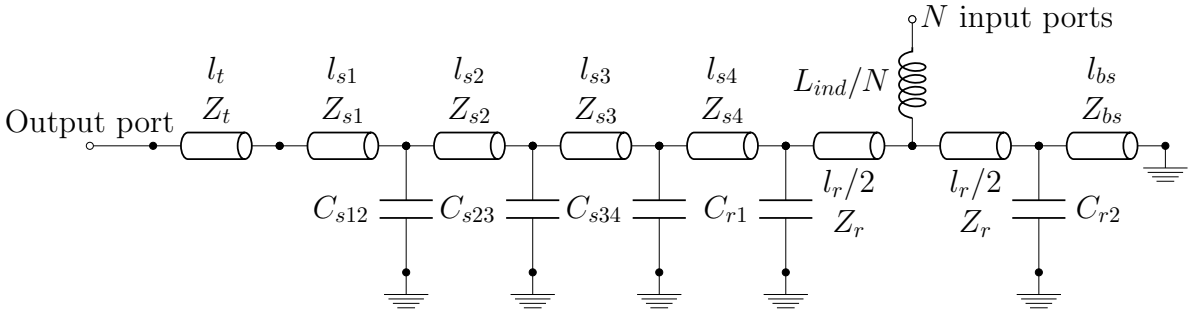


Figure 7: The equivalent circuit model of the entire combiner.

3 OPTIMIZATION STRATEGY

The advantage of using Space Mapping based optimization is that the optimization burden is shifted from the full-wave model (fine model) to the circuit model (coarse model). This is made possible through the alignment of the coarse model with the fine model before being optimized to reach any chosen design goals. The alignment of the coarse model with the fine model can be done using a combination of different methods, however, for this application only multiplicative and additive implicit space mapping (ISM) will be used.

The optimization strategy employed here is essentially the same as in [12]. The goal of the optimization is to obtain the vector of parameter values, $\boldsymbol{x}$, resulting in the lowest cost of the fine model response, $\boldsymbol{R}_f$, given the design objective function, U, and therefore

$$\boldsymbol{x}_f^* = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, U\left(\boldsymbol{R}_f(\boldsymbol{x})\right). \tag{3}$$

Since the main concern when designing axially symmetric combiners that support TEM mode propagation is impedance matching, the response, $\boldsymbol{R}_f$, is usually the central (output) port reflection coefficient, S_{11}, and the objective function, U, defines the desired return loss over the frequency band of interest. Instead of directly optimizing the fine model, ISM produces an approximate solution,

$$\boldsymbol{x}^{(i+1)} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, U\left(\boldsymbol{R}_s^{(i)}(\boldsymbol{x})\right), \tag{4}$$

which should lead to a solution of (3), where $\boldsymbol{R}_s^{(i)}(\boldsymbol{x})$ is the aligned coarse model (surrogate model) response at the i-th iteration. The surrogate model is constructed by optimizing the implicit parameters, $\boldsymbol{x}_p$, $\left(\boldsymbol{x}_p = [Z_r\ L_{ind}]^T \text{ for the coaxial combiner in Section 2.1}\right)$ to fit the coarse model response onto the fine model response. However, the implicit parameters are not optimized directly: The coefficients $\boldsymbol{G}$ and $\boldsymbol{H}$ are optimized and the implicit parameters calculated using

$$\boldsymbol{x}_p^{(i)}(\boldsymbol{x}) = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{H}, \tag{5}$$

where $\boldsymbol{G}$ is an mxn-matrix, and $\boldsymbol{H}$ is an mx1 matrix, with n the number of design parameters in $\boldsymbol{x}$ and m the number of implicit parameters in $\boldsymbol{x}_p$. The response of the surrogate model can then be obtained using

$$\boldsymbol{R}_s^{(i)}(\boldsymbol{x}) = \boldsymbol{R}_c\left(\boldsymbol{x}, \boldsymbol{x}_p^{(i)}(\boldsymbol{x})\right). \tag{6}$$

As explained in [12], equivalent circuit models for these types of devices are expected to be good representations of their physical structures, and information gained from previous iterations can be used to improve the surrogate models. The previous iterations are therefore taken into account by extracting values for $\boldsymbol{G}$ and $\boldsymbol{H}$ using

$$\boldsymbol{x}_p^{(i)}(\boldsymbol{x}) = \underset{\boldsymbol{G},\boldsymbol{H}}{\operatorname{argmin}} \sum_{k=0}^{i} w_k \left\| \boldsymbol{R}_f\left(\boldsymbol{x}^{(k)}\right) - \boldsymbol{R}_c\left(\boldsymbol{x}^{(k)}, \boldsymbol{x}_p^{(i)}(\boldsymbol{x}^{(k)})\right) \right\|, \tag{7}$$

where $w_k = w_0^{(k+1)}$. Generally $w_0 > 1$ to give the more recent iterations more weight, since it is expected that they will be closer to the optimum point in the parameter space, but $w_0 = 1$ will likely also produce acceptable results.

Each iteration consists of the optimization of the surrogate model, the evaluation of the fine model at the obtained parameter values, and the re-alignment of the coarse model with the fine model, generating a new surrogate model as outlined in (4) - (7). Note that while each iteration may consist of many coarse model evaluations, only a single fine model evaluation is

required. The iterations are repeated until the fine model response reaches the objective or a preset maximum number of iterations are completed. It should be noted that if the parameter extraction is performed using (7), the coarse model evaluation time increases with each iteration due to the increasing number of responses taken into account in (12).

4 DESIGN EXAMPLE: A COAXIAL COMBINER WITH IMPROVED BANDWIDTH

The design of an axially symmetric power combiner using surrogate based optimization will now be illustrated by using the coaxial combiner configuration shown in Fig. 2 and its equivalent circuit model shown in Fig. 7. The circuit model is analyzed using MATLAB, and the fine model evaluations are done using Computer Simulation Technology (CST) Microwave Studio (MWS) [17]. The inner and outer conductor radii of the central output coaxial port, $R_{I,t1}$ and $R_{O,t1}$, are chosen to be the same as the standard $50\ \Omega$ N-type connector, and the outer conductor radius of the oversized combining coaxial line, $R_{O,t2}$ is chosen to be 26 mm. Standard $50\ \Omega$ SMA connector dimensions are chosen for the peripheral feeding coaxial lines, and all the remaining physical dimensions are optimizable parameters. The input parameter vector is thus

$$\boldsymbol{x} = \left[H_t\ R_{s1}\ l_{s1}\ R_{s2}\ l_{s2}\ R_{s3}\ l_{s3}\ R_{s4}\ l_{s4}\ R_r\ l_r\ R_{bs}\ l_{bs}\right]^T, \tag{8}$$

and the implicit parameter vector is

$$\boldsymbol{x}_p = \left[Z_r\ L_{ind}\right]^T. \tag{9}$$

The parameter space is limited to

$$\boldsymbol{x}_{min} = \left[22\ 9\ 5\ 15\ 5\ 20\ 5\ 20\ 5\ 23\ 3\ 13\ 7\right]^T \tag{10}$$

and

$$\boldsymbol{x}_{max} = \left[130\ 20\ 30\ 22\ 30\ 25\ 30\ 25\ 30\ 25.5\ 3.5\ 23\ 25\right]^T, \tag{11}$$

with all values listed in the same order as in (8), and dimensions given in mm. The input parameter vector $\boldsymbol{x}$ contains 13 parameters, and according to (5) the matrix $\boldsymbol{G}$ takes all of these parameters into account in the calculation of $\boldsymbol{x}_p$. This is entirely unnecessary, and as a matter of fact unwanted, since it causes the implicit parameters to be dependent on parameter values they should obviously not be dependent on. For this design, the implicit parameters, Z_r and L_{ind}, are only dependent on the compensation ring radius, R_r. The matrices in (5) are thus reduced to take only R_r into account, resulting in a much more consistent surrogate model being constructed for each iteration. The design objective is to obtain a return loss of greater than 15 dB from 1.3 to 6.5 GHz. The objective function is defined as

$$U(\boldsymbol{R}_f(\boldsymbol{x})) = \frac{1}{i_M - i_m} \sum_{i=i_m}^{i_M} \left|\max\{\boldsymbol{R}_f(\boldsymbol{x}, f_i) - (-15), 0\}\right|, \tag{12}$$

where $\boldsymbol{R}_f(\boldsymbol{x}, f_i)$ is the central port reflection coefficient, S_{11}^{dB}, at frequency f_i, which is at index i in the frequency vector. The indices i_m and i_M correspond to the frequencies 1.3 and 6.5 GHz in f_i, respectively. For the ISM parameter extraction and the surrogate model optimization $w_0 = 1.1$ is chosen, and a global search is first performed using the Population Based Incremental Learning (PBIL) algorithm [18], followed by a local search using the Nelder-Mead simplex search [19]. Given the minimum and maximum values of R_r, some initial values for the implicit parameters are estimated as $Z_r = 4.5\ \Omega$ and $L_{ind} = 2$ nH, using the formula for

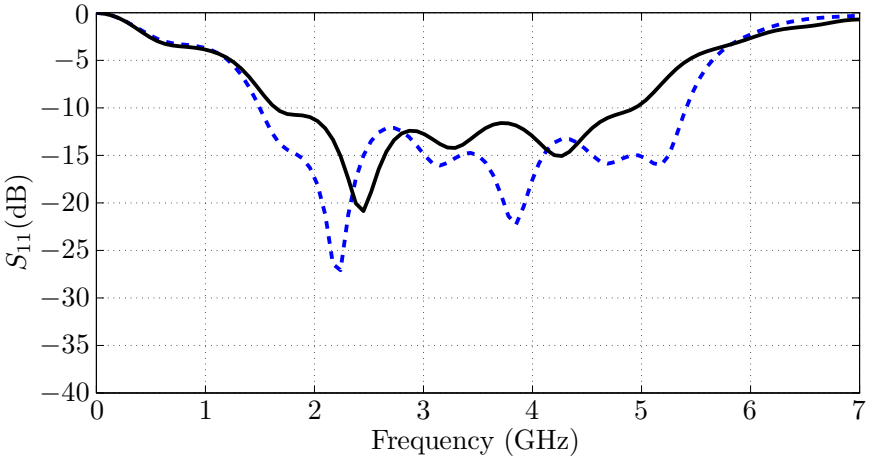


Figure 8: The responses of the surrogate (dashed) and fine (solid) models before alignment.

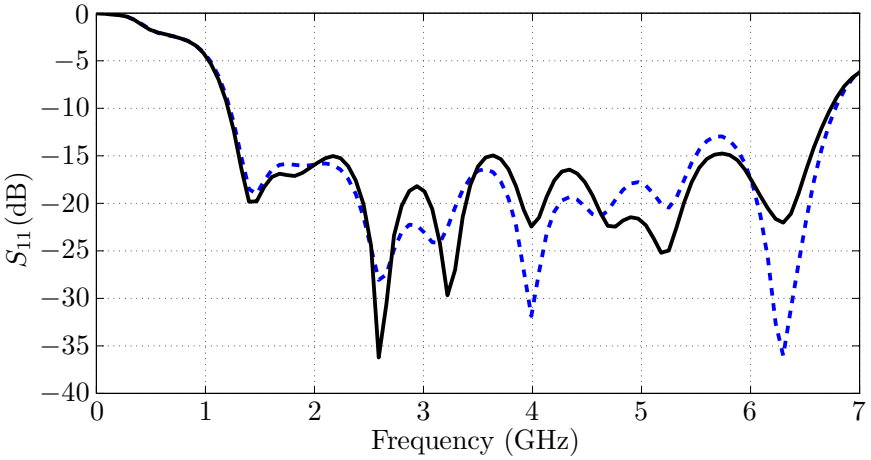


Figure 9: The responses of the surrogate (dashed) and fine (solid) models after 3 iterations.

the impedance of a normal unperturbed coaxial line as given in (2) and the 1 nH per mm rule of thumb. The optimization procedure is executed and the responses of the coarse and fine model are shown in Fig. 8, before the coarse model is aligned. The objective function defined in (12) is shown in Fig. 10 versus iteration number. After 3 iterations the optimization converged and produced the responses shown in Fig. 9, where the fine model response reaches the design objective of a reflection coefficient better than -15 dB over a 133 % bandwidth from 1.3 to 6.5 GHz. The resulting parameter values are listed in Table 2.

5 CONCLUSIONS

This paper presented guidelines for the design and optimization of axially symmetric power combiners. A complete equivalent circuit model of a coaxial combiner was derived using the guidelines. The derived equivalent circuit model was used as a coarse model in a Space Mapping based optimization procedure to design a combiner with improved bandwidth performance. The design objective was to obtain a central port reflection coefficient of better than -15 dB from 1.3 to 6.5 GHz. The optimization procedure converged after 3 iterations and a full-wave simulated

Parameter	Value	Units	Parameter	Value	Units	Parameter	Value	Units
H_t	117.66	mm	l_{s3}	17.34	mm	R_r	23.93	mm
l_{s1}	20.68	mm	R_{s3}	20.02	mm	l_r	3.07	mm
R_{s1}	15.11	mm	l_{s4}	18.59	mm	R_{bs}	14.08	mm
l_{s2}	19.42	mm	R_{s4}	22.38	mm	l_{bs}	9.09	mm
R_{s2}	17.25	mm						

Table 2: Parameter values for the three transitions used to test the equivalent circuit model of the impedance tapered transition.

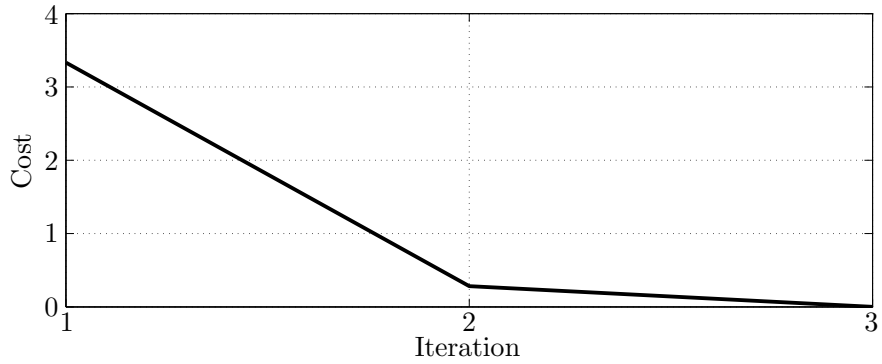


Figure 10: The objective cost after each iteration.

reflection coefficient of better than -15 dB over a 133 % bandwidth from 1.3 to 6.5 GHz was obtained, compared to a measured reflection coefficient of -12 dB over a 112 % bandwidth from 0.52 to 1.86 GHz obtained in [6].

REFERENCES

[1] A. E. Fathy, S. W. Lee, and D. Kalokitis, "A simplified design approach for radial power combiners," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 1, pp. 247–255, Jan. 2006.

[2] J. M. Denoual, A. Peden, B. Della, and J. P. Fraysse, "16-way radial divider/combiner for solid state power amplifiers in the K-band," in *38th European Microwave Conference Proceedings, (EuMC) 2008*, pp. 345–348, Oct. 2008.

[3] K. Song and Q. Xue, "Planar probe coaxial-waveguide power combiner/divider," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 11, pp. 2761–2767, Nov. 2009.

[4] X. Shan, Z. Shen, P. Kumaresh, and R. M. Jayasuriya, "A novel 8-way radial power combiner," in *Asia-Pacific Microwave Conference Proceedings, (APMC) 2009*, pp. 2625–2628, Dec. 2009.

[5] Q. Xue, K. Song, and C. H. Chan, "China: power combiners/dividers," *IEEE Microwave Magazine*, vol. 12, no. 3, pp. 96–106, May 2011.

[6] M. Amjadi and E. Jafari, "Design of a broadband eight-way coaxial waveguide power combiner," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 1, pp. 39–45, Jan. 2012.

[7] D. I. L. de Villiers, P. W. van der Walt, and P. Meyer, "Design of a ten-way conical transmission line power combiner," *IEEE Transactions on Microwave Theory and Techniques*, vol. 55, no. 2, pp. 302–308, Feb. 2007.

[8] D. I. L. de Villiers, P. W. van der Walt, and P. Meyer, "Design of conical transmission line power combiners using tapered line matching sections," *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, no. 6, pp. 1478–1484, Jun. 2008.

[9] R. D. Beyers and D. I. L. de Villiers, "Compact conical-line power combiner design using circuit models," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 11, pp. 2650–2658, Nov. 2014.

[10] S. Koziel, Q. S. Cheng, and J. W. Bandler, "Space mapping," *IEEE Microwave Magazine*, vol. 9, no. 6, pp. 105–122, Dec. 2008.

[11] S. Koziel, J. W. Bandler, and K. Madsen, "A space-mapping framework for engineering optimization: theory and implementation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 10, pp. 3721–3730, Oct. 2006.

[12] D. I. L. de Villiers and R. D. Beyers, "Design of conical transmission line power combiners using space mapping," in *International Conference on Electromagnetics in Advanced Applications, (ICEAA) 2015* , pp. 1437–1440, Sep. 2015.

[13] R. A. York, "Some considerations for optimal efficiency and low noise in large power combiners," *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, no. 8, pp. 1477–1482, Aug. 2001.

[14] P. I. Somlo, "The computation of coaxial line step capacitances," *IEEE Transactions on Microwave Theory and Techniques*, vol. 15, no. 1, pp. 48–53, Jan. 1967.

[15] R. D. Beyers and D. I. L. de Villiers, "Design and analysis of an impedance tapered conical to coaxial transmission line transition," in *44th European Microwave Conference Proceedings, (EuMC) 2014*, pp. 307–310, Oct. 2014.

[16] R. D. Beyers and D. I. L. de Villiers, "Analysis of shorted coaxial peripheral feeding networks for conical line power combiners," in *Asia-Pacific Microwave Conference Proceedings, (APMC) 2013*, pp. 285–287, Nov. 2013.

[17] Computer Simulation Technology AG, Darmstadt, Germany. CST version 2015. [Online]. Available: www.cst.com

[18] S. Baluja, "Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning," Carnegie Mellon University, Technical Report CMU-CS-94-163, 1994.

[19] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

FLEXURAL VIBRATION ANALYSIS OF GRAPHENE NANOPLATELETS REINFORCED NANOCOMPOSITE BEAMS

Jie Yang and Chuang Feng

School of Engineering, RMIT University
PO Box 71, Bundoora, VIC 3083, Australia
{j.yang, chuang.feng}@rmit.edu.au

Keywords: Flexural Vibration, Nanocomposites, Graphene Nanoplatelet, Timoshenko Beam Theory

Abstract. *Polymer nanocomposites with the addition of a small percentage of graphenes as reinforcement have demonstrated tremendous potential to serve as next generation functional or structural materials due to their combination of high specific surface area, strong nanofiller-matrix adhesion and the outstanding mechanical properties of the sp2 carbon bonding network in graphene. While the majority of the work on graphene based composites has been focused on quantifying their mechanical properties, the vibration characteristics of graphene composites has not yet been investigated. This paper presents an analytical study on the flexural vibration of polymer beams reinforced with uniformly distributed graphene nanoplatelets (GPL). Based on the assumption that each graphene nanoplatelet acts as an effective rectangular solid fiber, the Halpin-Tsai model for fiber-reinforced composites is modified to predict the Young's modulus of the GPL/epoxy nanocomposites. Governing equations of motion are derived within the framework of linear elastic strain-displacement relationship and Timoshenko beam theory to account for transverse shear strain. The influences of GPL weight fraction, beam slenderness ratio and boundary conditions on the natural frequencies and associated mode shapes are investigated in detail. It is found that adding a very small amount of GPLs into the polymer matrix can effectively lead to a remarkable increase in beam stiffness hence the natural frequencies of the nanocomposite beams, and that the rate of frequency change of the beam is independent of the slenderness ratio.*

1 INTRODUCTION

Compared to carbon nanotubes (CNTs), graphene and its derivatives have demonstrated favorable advantages when used as reinforcing fillers in composites. For example, the ultimate strength and Young's modulus of graphene and its derivatives can reach up to 1 TPa and 130 GPa, respectively [1]. A recent experimental study [2] showed that with the same weight fraction of additives, i.e. 0.1±0.002%, Young's modulus of graphene based epoxy nanocomposites was 31% greater than that of pristine epoxy while an increase of 3% was observed when CNT fillers are used. This remarkable reinforcing effect stems from graphene's excellent mechanical properties and very high specific surface areas [2-7]. In addition, graphene's derivatives, i.e. graphene nanoplatelets, are in abundance in nature. This significantly expedites the engineering applications of graphene based composites due to their superior mechanical properties but moderate cost [8].

Extensive theoretical and experimental research work has been done on the fabrication and characterization of mechanical properties of graphene based nanocomposites. Rafiee et al. [2] fabricated functionalized graphene sheet (FGS)/epoxy nanocomposite and found this composite offers significant improvements in mechanical properties compared to those of CNT reinforced epoxy composites. Lee et al. [9] successfully synthesized FGS/epoxy nanocomposite whose strength was increased by approximately 30-80% at a loading of 1.6 wt%. Martín-Gallego et al. [10] studied the reinforcement of graphene and observed the significant improvement in mechanical properties of FGS-epoxy composites. Graphene is usually present in the form of agglomerates due to van der Waals force and the dispersion state is crucial to the performance of the composites. Tang et al. [6] discussed the effect of graphene dispersion on the mechanical properties of graphene/epoxy nanocomposites. They found the nanocomposites with highly dispersed graphene exhibit higher strength and better fracture toughness. This significant enhancement was also observed in other previous studies [1, 5, 7]. Among the theoretical studies, Rahman and Haque [11] investigated the mechanical properties of graphene platelet (GPL)/epoxy nanocomposites by molecular mechanics (MM) and molecular dynamics simulations (MD). Significant improvement in Young's modulus and shear modulus of the composites was observed. Cho and co-workers [12] predicted the elastic constants of graphite using MM and determined the elastic constants of graphite/epoxy nanocomposites by using Mori–Tanaka micromechanical model. Xiang et al. [13] also developed a micromechanics model for graphene/polymer nanocomposite based on Mori–Tanaka method. Spanos et al. [14] used a micromechanical finite element approach to obtain the mechanical properties of the composites reinforced with uniformly distributed graphenes. They found that their mechanical properties depend on the dimensions and volume fraction of graphenes as well as on the stiffness of the interphase between the two constituents. Chandra et al. [15] presented a multiscale model for vibration frequency analysis of graphene/polymer composites and observed exceptional vibrational behaviour and large stiffness of graphene reinforced composites.

It should be pointed out that although there has been a fast growing interest in graphene based polymer composites, the majority of the previous work is primarily focused on the material fabrication and characterization. Studies on the structural behaviour are very limited in number. This paper presents an analytical study on the flexural vibration of polymer nanocomposite beam in which GPLs are uniformly dispersed in epoxy matrix. Theoretical formulations are established within the framework of Timoshenko beam theory and linear strain-displacement relationship. The modified Halpin-Tsai model is used to predict the Young's modulus of the GPL/epoxy nanocomposites. A detailed parametric study is

conducted to discuss the influences of GPL weight fraction, beam slenderness ratio and boundary conditions on the natural frequencies and associated mode shapes of the beam.

2 GOVERNING EQUATIONS

2.1 Equations of motion

Shown in figure 1 is a GPL/epoxy nanocomposite beam with length *l*, width *b* and thickness *h*, subjected to a distributed dynamic load *q*(*x*, *t*). According to Timoshenko beam theory, the bending moment *M* and shear force *Q* are given as

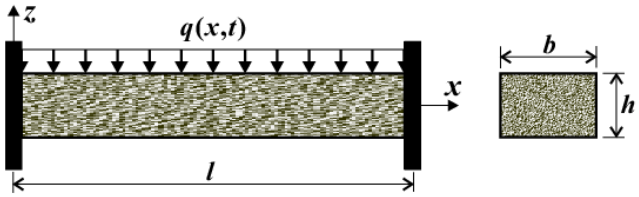


Figure 1: Schematic configuration of a GPL/epoxy nanocomposite beam.

$$M(x) = -E_c I \frac{\partial \varphi}{\partial x}, \quad Q(x) = \alpha A G_c (\frac{\partial w}{\partial x} - \varphi) \tag{1}$$

where shear correction factor α = 5/6, E_c and G_c are the effective Young's modulus and shear modulus of the beam, A and I are the area and second moment of the cross-section, and φ represents the cross-sectional rotation, respectively. The governing equations of the beam can be derived as

$$E_C I \frac{\partial^2 \varphi}{\partial x^2} - \alpha A G \varphi + \alpha A G \frac{\partial w}{\partial x} + \rho I \frac{\partial^3 w}{\partial x \partial t^2} = 0 \tag{2}$$

$$\alpha A G \frac{\partial^2 w}{\partial x^2} - \alpha A G \frac{\partial \varphi}{\partial x} + bq(x,t) = \rho A \frac{\partial^2 w}{\partial t^2} \tag{3}$$

where ρ is the mass density of the beam. Combining equations (2) and (3) and eliminating φ leads to the equation of motion

$$E_c I \frac{\partial^4 w}{\partial x^4} - \rho I \left(1 + \frac{E_c}{G_c \alpha}\right) \frac{\partial^4 w}{\partial x^2 \partial t^2} + \rho A \frac{\partial^2 w}{\partial t^2} = q_0 cos(\omega_e t) \tag{4}$$

where $q = q_0\cos(\omega_e t)$ with q_0 and ω_e being the amplitude and radian frequency of the transverse load. Introducing the following dimensionless quantities

$$X = \frac{x}{l}, \; W = \frac{w}{h}, \; \eta = \frac{l}{h}, \; \tau = \frac{t}{l}\sqrt{\frac{E_c}{\rho_c}}, \; Q_0 = \frac{12 q_0 \eta^4}{E_c} \tag{5}$$

equation (4) can be cast into the following dimensionless form

$$\frac{\partial^4 W}{\partial X^4} - [1 + 2(1+\nu)\frac{1}{\alpha}] \frac{\partial^4 W}{\partial X^2 \partial \tau^2} + 12\eta^2 \frac{\partial^2 W}{\partial \tau^2} = Q_0 \cos(\omega \tau) \tag{6}$$

where $\omega = \omega_e l \sqrt{\rho_c / E_c}$ is the dimensionless excitation frequency.

2.2 Halpin-Tsai model for GPL/polymer nanocomposites

Assuming that GPLs serve as effective rectangular solid fillers uniformly dispersed in the polymer matrix, the effective Young's modulus of the nanocomposite can be approximated by employing modified Halpin-Tsai micromechanics model as [2, 16-18]

$$E_{\mathrm{C}}=\frac{3}{8}\frac{1-\xi\eta_{\mathrm{L}}V_{\mathrm{eff,GPL}}}{1-\eta_{\mathrm{L}}V_{\mathrm{eff,GPL}}}\times E_{\mathrm{M}}+\frac{5}{8}\frac{1-\xi\eta_{\mathrm{W}}V_{\mathrm{eff,GPL}}}{1-\eta_{\mathrm{W}}V_{\mathrm{eff,GPL}}}\times E_{\mathrm{M}} \tag{7}$$

where E_{M} is Young's modulus of the matrix, $V_{\mathrm{eff,GPL}}$ is GPLs' volume fraction, $\xi=(l_{\mathrm{GPL}}+w_{\mathrm{GPL}})/h_{\mathrm{GPL}}$, $\eta_{\mathrm{L}}=\frac{(E_{\mathrm{eff,GPL}}/E_{\mathrm{M}})-1}{(E_{\mathrm{eff,GPL}}/E_{\mathrm{M}})+\xi}$ and $\eta_{\mathrm{W}}=\frac{(E_{\mathrm{eff,GPL}}/E_{\mathrm{M}})-1}{(E_{\mathrm{eff,GPL}}/E_{\mathrm{M}})+2}$ in which l_{GPL}, w_{GPL}, h_{GPL} and E_{GPL} are the length, width, thickness, and Young's modulus of GPLs, respectively. Given the mass density of GPLs ρ_{GPL} and that of polymer matrix ρ_{M}, GPLs' volume fraction in the nanocomposite can be approximated as

$$V_{\mathrm{GPL}}=\frac{f_{\mathrm{GPL}}}{f_{\mathrm{GPL}}+\left(\rho_{\mathrm{GPL}}/\rho_{\mathrm{M}}\right)\left(1-f_{\mathrm{GPL}}\right)} \tag{8}$$

where f_{GPL} is the weight fraction of the GPLs in the composite.

Equation (7) was used to estimate the effective Young's modulus of GPL/polymer nanocomposites and has been experimentally validated by Rafiee et al. [2]. It is assumed that Poison's ratio is $\nu = 0.4$ due to the fact that the mechanical behavior of the nanocomposite is weakly dependent on Poisson's ratio. This is verified in our numerical simulations which indicate that Poison ratio has a very limited effect on the vibration of the beam.

3 SOLUTION

3.1 Free vibration

For free vibration, the governing equation is reduced to

$$\frac{\partial^4 W}{\partial X^4}-[1+2(1+\nu)\frac{1}{\alpha}]\frac{\partial^4 W}{\partial X^2\partial\tau^2}+12\eta^2\frac{\partial^2 W}{\partial\tau^2}=0 \tag{9}$$

For harmonic vibration, the transverse displacement can be written as

$$W(X,\tau)=\emptyset(X)e^{i\omega_l\tau} \tag{10}$$

where $\omega_l=\Omega l\sqrt{\rho_c/E_c}$ is the dimensionless linear natural frequency with Ω being the natural frequency of the beam. Substituting equation (10) into (9), the solution to the resulting differential equation is

$$\emptyset(X)=C_1 sinh\,(\lambda_1 X)+C_2 cosh(\lambda_1 X)+C_3\,sin(\lambda_2 X)+C_4 cos(\lambda_2 X) \tag{11}$$

where

$$\lambda_{1,2}^2=\frac{-\left[1+2\left(1+\nu\right)\frac{1}{\alpha}\right]\omega_l^2\pm\sqrt{\left[1+2\left(1+\nu\right)\frac{1}{\alpha}\right]^2\omega_l^4+48\eta^2\omega_l^2}}{2} \tag{12}$$

and C_1 to C_4 are unknown constants to be determined by associated boundary conditions.

In this paper, clamped-clamped (C-C), clamped-hinged (C-H) and hinged-hinged (H-H) beams are considered, with the following boundary conditions

a) C-C beam: $\emptyset(0)=\emptyset(1)=\frac{d\emptyset(0)}{dx}=\frac{d\emptyset(1)}{dx}=0$;
b) C-H beam: $\emptyset(0)=\emptyset(1)=\frac{d\emptyset(0)}{dx}=M(1)=0$;
c) H-H beam: $\emptyset(0)=\emptyset(1)=M(0)=M(1)=0$.

The boundary conditions for each case will result in an algebraic equation, i.e. $[H(\omega_l)]\{\chi\}=\{0\}$. $H(\omega_l)$ is a matrix dependent on ω_l while χ is a vector consisting of the four

unknown coefficients. The non-trivial solution of χ requires the determinant of the $H(\omega_l)$ be zero, from which the natural frequencies and the associated mode shapes can be obtained.

3.2 Forced harmonic vibration

For forced harmonic vibration, the transverse displacement is approximated as

$$W(X,\tau) = S(\tau)\emptyset(X) \tag{13}$$

where $\emptyset(X)$ is the mode shape obtained from free vibration analysis and $S(\tau)$ denote the vibration characteristics to be determined. Substituting equation (13) into equation (6) and applying Galerkin method, the governing equation becomes

$$\ddot{S}(\tau)+\gamma S = Q_E\cos(\omega\tau) \tag{14}$$

where

$$\begin{cases} \gamma = -\dfrac{1}{\gamma_o}\displaystyle\int_0^1\left(\phi\frac{d^4\phi}{dX^4}\right)dX, \quad Q_E = \dfrac{Q_0}{\gamma_o}\displaystyle\int_0^1\phi dX \\ \gamma_o = \displaystyle\int_0^1\left[12\eta^2\phi^2-\left(1+\frac{2(1+\nu)}{\alpha}\right)\phi\frac{d^2\phi}{dX^2}\right]dX \end{cases} \tag{15}$$

The dynamic response of the beam under transverse loading is obtained as

$$S(\tau)=\left(S_0-\frac{Q_0}{\gamma-\omega^2}\right)\cos(\omega_l\tau)+\frac{\dot{S}_0}{\omega_l}\sin(\omega_l\tau)+\frac{Q_0}{\gamma-\omega^2}\cos(\omega\tau) \tag{16}$$

4 RESULTS AND DISCUSSION

Table 1 compares our results with the previous data [19] for the natural frequency of a homogeneous beam whose Young's modulus, mass density and Poisson's ratio are E = 70 GPa, ν = 0.33 and ρ = 2780 kg/m^3, respectively. Excellent agreement is observed.

C-C		H-H	
Present	Ref [19]	Present	Ref [19]
5.5933	5.59	2.4674	2.47

Table 1: Comparison of linear natural frequency.

In what follows, the dimensions, mass density, Young's modulus and weight fraction of GPLs are l_{GPL} = 2.5 μm, w_{GPL} = 1.5 μm, t_{GPL} = 1.5 nm, ρ_{GPL} = 1.06 g/cm^3, E_{GPL} = 1.01 TPa, and f_{GPL} = 1.5%, respectively [2]. Young's modulus of the epoxy matrix is E_M = 3.0 GPa. Table 2 lists the first three dimensionless frequencies of the nanocomposite beams with different end supports. The associated mode shapes are displayed in figure 2. As expected, the frequencies are the highest for a C-C beam and the lowest for an H-H beam. The significant reinforcing effect of GPLs on the dynamic characteristics of the beam is clearly evidenced by over 60% increase in the natural frequencies as compared to those of pristine epoxy beams.

Mode	C-C		C-H		H-H	
	Pure epoxy	Composite	Pure epoxy	Composite	Pure epoxy	Composite
1	0.3090	0.5057	0.2031	0.3426	0.1151	0.2108
2	0.8715	1.4060	0.7000	1.1351	0.5445	0.8918
3	1.7247	2.7665	1.4822	2.3823	1.2571	2.0263

Table 2: Dimensionless natural frequency of composite beam.

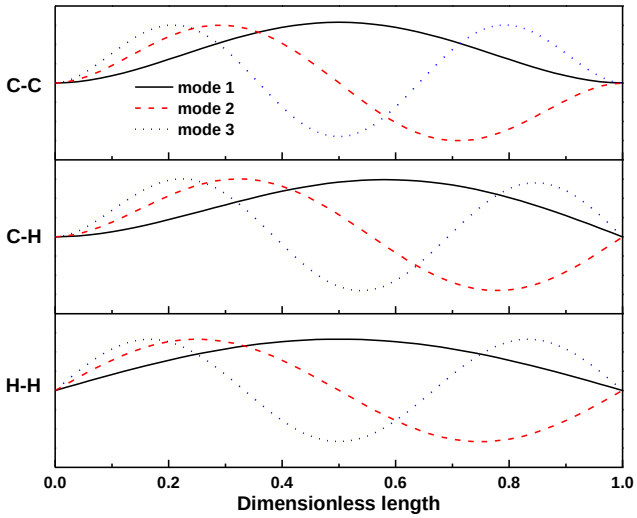


Figure 2: Mode shapes for nanocomposite beams with different boundary conditions.

The dimensionless fundamental frequency of a C-C beam versus slenderness ratio is shown in figure 3. It is easily understood that the frequency decreases as the slenderness ratio increases. At the same slenderness ratio, the natural frequency of the beam with GPL reinforcement is higher than that of the pure epoxy beam. This is due to the significantly increased effective Young's modulus of the beams when reinforced with GPLs, regardless of the end supports of the beams.

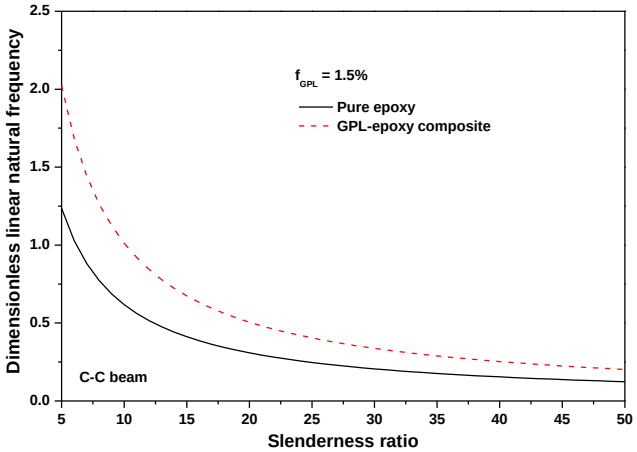


Figure 3: Variation of dimensionless fundamental frequency with beam slenderness ratio.

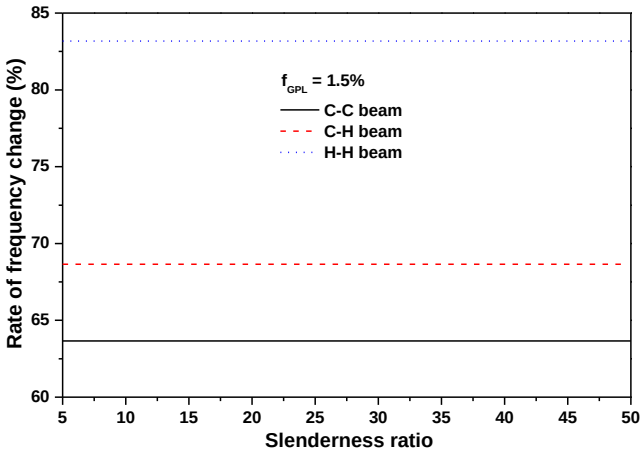


Figure 4: Effect of slenderness ratio on frequency change

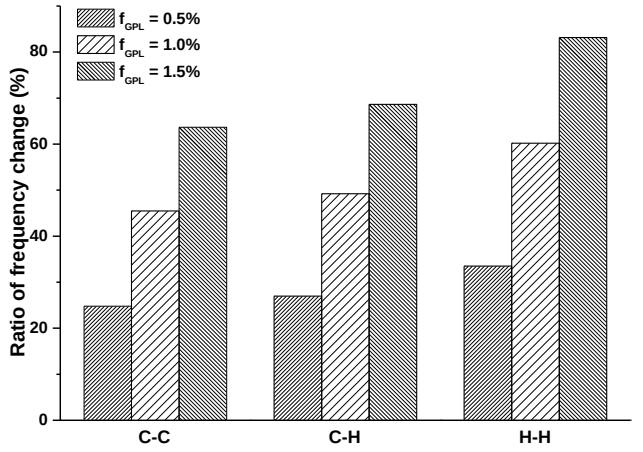


Figure 5: Effect of GPL weight fraction on frequency change.

Figure 4 presents the effect of slenderness ratio on the frequency change $(f_c–f_0)/f_0$, where f_c and f_0 denote the natural frequencies with and without GPLs. It is seen that the frequency gain is the highest for H-H beam (82.5%) whereas the lowest for C-C beam (63%). The rate of frequency change, however, is independent of the slenderness ratio for all beams. As revealed by figure 5, the frequency gain is almost doubled as the GPL weight fraction increases from 0.5% to 1.5%.

Figure 6 shows the dimensionless free vibration amplitudes of C-C, C-H and H-H beams under initial condition $W(0) = 0.1$ and $\dot{W}(0) = 0$. The vibration period, which is the lowest for the C-C beam among the three beams considered, becomes smaller when the beam is reinforced with GPLs. Figure 7 presents the dynamic responses of the same beams under a harmonic load Q_E = 0.001. As can be seen, the H-H beam has the highest vibration amplitude. Compared with the pure epoxy beam, the maximum vibration amplitude is considerably reduced with the addition of 1.5% w.t. GPLs into the epoxy matrix, confirming once again the remarkable reinforcing effect of GPLs on the dynamic performance of the beams. It should be also noted that this effect is most prominent in H-H beams. This is consistent with the observations obtained in natural frequencies.

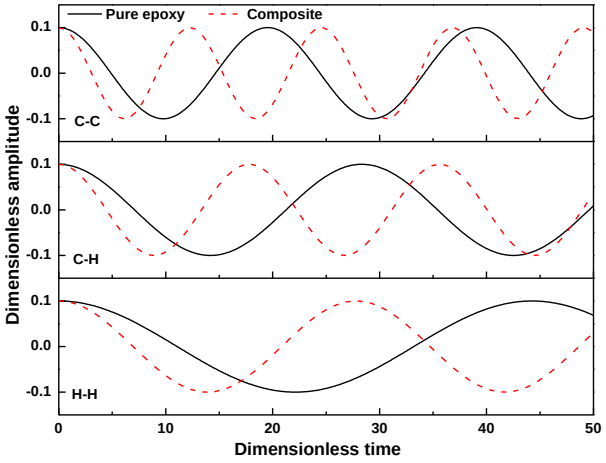


Figure 6: Dimensionless free vibration response of beams with different end supports.

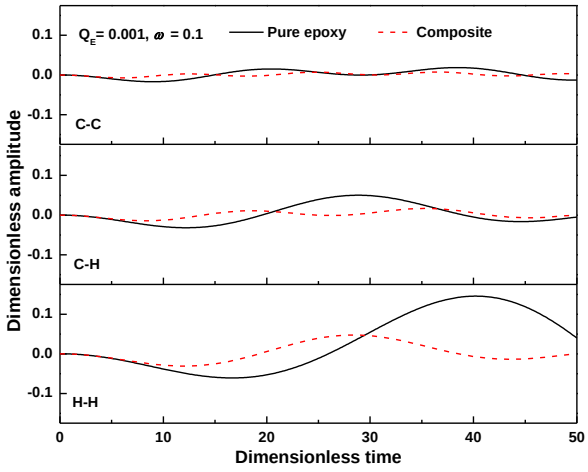


Figure 7: Dimensionless forced vibration response of beams with different end supports.

5 CONCLUSIONS

- Free and forced vibrations of GPL reinforced nanocomposite beam have been analytically investigated within the framework of Timoshenko beam theory.
- The addition of a very small amount of GPLs can significantly improve the free and forced vibration performance of the nanocomposite beams.
- The rate of frequency change increases at a higher GPL weight fraction but is independent of the slenderness ratio of the beam.
- Among the three boundary conditions considered, the reinforcing effect of GPLs is more significant for an H-H beam than the other two.

REFERENCES

[1] I. Zaman, T.T. Phan, H.-C. Kuan, Q. Meng, L.T.B. La, L. Luong, et al., Epoxy/graphene platelets nanocomposites with two levels of interface strength. *Polymer*, 52, 1603-1611, 2011.

[2] M.A. Rafiee, J. Rafiee, Z. Wang, H. Song, Z.Z. Yu, N. Koratkar, Enhanced mechanical properties of nanocomposites at low graphene content. ACS nano, 3, 3884-3890, 2009.

[3] M.M. Shokrieh, M. Esmkhani, Z. Shokrieh, Z. Zhao, Stiffness prediction of graphene nanoplatelet/epoxy nanocomposites by a combined molecular dynamics–micromechanics method. Computational Materials Science, 92, 444-450, 2014.

[4] T. Kuilla, S. Bhadra, D. Yao, N.H. Kim, S. Bose, J.H. Lee, Recent advances in graphene based polymer composites. Progress in polymer science, 35, 1350-1375, 2010.

[5] [5] X. Zhao, Q. Zhang, D. Chen, P. Lu, Enhanced mechanical properties of graphene-based poly (vinyl alcohol) composites. Macromolecules, 43, 2357-2363, 2010.

[6] L.C. Tang, Y.J. Wan, D. Yan, Y.B. Pei, L. Zhao, Y.B. Li, et al., The effect of graphene dispersion on the mechanical properties of graphene/epoxy composites. Carbon, 60, 16-27, 2013.

[7] S.K. Yadav, J.W. Cho, Functionalized graphene nanoplatelets for enhanced mechanical and thermal properties of polyurethane nanocomposites. Applied surface science, 266, 360-367, 2013.

[8] J.R. Potts, D.R. Dreyer, C.W. Bielawski, R.S. Ruoff, Graphene-based polymer nanocomposites. Polymer, 52, 5-25, 2011.

[9] J.K. Lee, S. Song, B. Kim, Functionalized graphene sheets- epoxy based nanocomposite for cryotank composite application. Polymer Composites, 33, 1263-1273, 2012.

[10] M. Martín-Gallego, R. Verdejo, M. Lopez-Manchado, M. Sangermano, Epoxy-graphene UV-cured nanocomposites. Polymer, 52, 4664-4669, 2011.

[11] R. Rahman, A. Haque, Molecular modeling of crosslinked graphene–epoxy nanocomposites for characterization of elastic constants and interfacial properties. Composites Part B: Engineering, 54, 353-364, 2013.

[12] J. Cho, J. Luo, I. Daniel, Mechanical characterization of graphite/epoxy nanocomposites by multi-scale analysis. Composites science and technology, 67, 2399-2407, 2007.

[13] X.Y. Ji, Y.P. Cao, X.Q. Feng, Micromechanics prediction of the effective elastic moduli of graphene sheet-reinforced polymer nanocomposites. Modelling and Simulation in Materials Science and Engineering, 18, 045005, 2010.

[14] K. Spanos, S. Georgantzinos, N. Anifantis, Mechanical properties of graphene nanocomposites: A multiscale finite element prediction. Composite Structures, 132, 536-544, 2015.

[15] Y. Chandra, R. Chowdhury, F. Scarpa, S. Adhikari, J. Sienz, C. Arnold, et al., Vibration frequency of graphene based composites: A multiscale approach. Materials Science and Engineering: B, 177, 303-310, 2012.

[16] P.K. Mallick, Fiber-reinforced composites: materials, manufacturing, and design. CRC press, 2007.

[17] E.T. Thostenson, T.W. Chou, On the elastic properties of carbon nanotube-based composites: modelling and characterization. Journal of Physics D: Applied Physics, 36, 573, 2003.

[18] J.C. Halpin, Effects of Environmental Factors on Composite Materials. DTIC Document, 1969.

[19] J. Yang, Y. Chen, Free vibration and buckling analyses of functionally graded beams with edge cracks. Composite Structures, 83, 48-60, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DYNAMIC RESPONSE OF REAL OFFSHORE WIND TURBINES ON MONOPILES IN STRATIFIED SEABED

Guillermo M. Álamo[1], Juan J. Aznárez[1], Luis A. Padrón[1], Alejandro E. Martínez-Castro[2], Rafael Gallego[2] and Orlando Maeso[1]

[1]Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería (SIANI), Universidad de Las Palmas de Gran Canaria
Edificio Central del Parque Científico y Tecnológico, Campus Universitario de Tafira, 35017 Las Palmas de Gran Canaria, Spain
e-mail: {guillermo.alamo, juanjose.aznarez, luis.padron, orlando.maeso}@ulpgc.es

[2] Departamento de Mecánica de Estructuras e Ingeniería Hidráulica, ETS de Ingenieros de Caminos, Canales y Puertos, Universidad de Granada
Avenida Fuentenueva s/n, 18002 Granada, Spain
e-mail: {amcastro, gallego}@ugr.es

Keywords: offshore wind turbines, soil-structure interaction, monopile impedance functions, modal analysis

Abstract. *This paper studies the influence of the superstructure and foundation parameters of offshore wind turbine (OWT) systems, as well as of the soil profile, on the magnitude of the associated soil-structure interaction (SSI) phenomena. The analyses are carried out by assuming characteristic properties of real OWTs and of soil profiles based on North Sea boreholes. To do so, a simplified substructuring model is proposed for the computation of the fundamental frequency and equivalent damping of OWTs founded on monopiles including the SSI effects. The whole superstructure is reduced to a three-degrees-of-freedom system through its modal mass and height, while the foundation stiffness is represented by impedance functions. The pile impedance functions are computed by a time-harmonic integral model that makes use of Green's Functions for the layered halfspace to represent the soil behaviour, while the pile is represented by finite elements as a Timoshenko beam and treated as a load-line within the soil.*

The obtained results confirm the necessity of considering the SSI effects for an accurate estimation of both the fundamental frequency and equivalent damping of the soil-structure system. Regarding the pile dimensions, the pile diameter plays a significant role on the magnitude of the SSI effects, while the pile length has almost no influence. On the other hand, the results highlight the importance of a good knowledge of the soil profile, as high differences are produced between the homogeneous and variable-with-depth profiles, even when both present the same mean shear velocity. The superficial soil layers are found to be the ones of crucial importance when evaluating the SSI effects on the dynamic properties of OWT systems.

1 INTRODUCTION

Wind energy has become in the last decades a viable and sustainable solution to the continuously increasing energy demand. Offshore wind turbines (OWT) present some advantages over their onshore counterparts: better wind conditions, terrains with less occupancy, and lower noise and visual impact on the population. This, together with the advances in the generators technology, has produced a significant increment of offshore wind farms in the recent times. Among the different foundations systems that are used for OWT constructions, the monopile is preferred over the jacket or gravity-based foundations due to its simplicity and cost-efficiency.

The accurate estimation of the fundamental frequency of the OWT system is a keystone of the design stage of these constructions. The fundamental frequency has to be kept away from the excitation frequency content in order to prevent resonance effects. There are three principal frequencies to avoid: the wind and wave frequencies ($<$0.1 Hz), the rotor frequency (1P) and the blade-passing frequency (2P or 3P). Recommendation codes advise to place the system fundamental frequency outside the $\pm 10\%$ range of the aforementioned frequencies [1]. In order to correctly estimate the tower-foundation system fundamental frequency, the soil stiffness has to be taken into account as the soil-structure interaction (SSI) effects can significantly reduce the system fundamental frequency value. The dynamic characterization, i.e. computation of the modified fundamental frequency and damping, of OWT structures including the SSI effects has been the object of study for numerous recent works [2–21].

In his early work, Zaaijer [2] compared different methodologies used to estimate the structural fundamental frequency taking into account the soil effects. A Finite Element Method (FEM) model including the American Petroleum Institute (API) p-y, t-z, Q-z curves was taken as reference method and different foundation systems, such as monopiles, gravity footings and jackets, were assumed. The best results were obtained by using impedance matrices, as inertia effects in the foundation and non-linear soil-structure interaction were negligible. Thus, stiffness matrix models are applicable for pile foundations under loading conditions relevant for the fatigue analysis. The obtained numerical results were compared with measures from two wind farms, resulting in acceptable predictions of the fundamental frequencies. However, Zaaijer found that, as known from offshore practice, the models tend to underpredict the foundation stiffness and, therefore, the system fundamental frequency. In the same work, Zaaijer also studied the sensitivity of the system fundamental frequency to variations in different parameters of soil, structure, foundation and environment, obtaining that soil parameters dominate the uncertainty of the natural frequency. The effect of the uncertainties in the soil profile was also studied by Carswell et al. [8] through probabilistic methods focusing on the Serviceability Limit State (SLS). They concluded that the system reliability shows the same sensibility to load as to soil uncertainty.

Adhikari and Bhattacharya [3] enhanced the Bernoulli model developed by Tempel and Molenaar [22] in order to include the effects of a flexible foundation and the tower axial load. The pile foundation was represented by two frequency-independent springs simulating the lateral and rotational stiffness. They illustrated the analytical results with numerical examples and applied their model to real turbines. Later Bhattacharya and Adhikari [9] evaluated the lateral and rotational foundation stiffness by direct measurement and compared their results with those obtained from a small-scale prototype and from a FEM model. They observed that analytical and FEM results overestimated, in general, the system natural frequency. More recently, Arany et al. [4] further developed their model by including the cross-coupling term in the foundation stiffness and using the Timoshenko theory to model the tower. They concluded

that the cross-coupling spring term has a significant effect on the natural frequency, while the Timoshenko beam model does not significantly improve the results, being the slender beam assumption accurate enough. The effect of including the cross-coupling term was also studied by Zania [5]. She presented an analytical iterative method to obtain the equivalent modified period and damping due to SSI effects based on the pile impedance functions from Novak and Nogami [23]. She concluded that disregarding the off-diagonal terms and the frequency dependency of the impedance matrix is inappropriate, since it results in a non-conservative overestimation of the fundamental frequency and underestimation of damping. This effect was more evident as the height of the system increases. This conclusion might explain why Bhattacharya and Adhikari [9] find that natural frequencies tend to be overestimated, disagreeing with what was previously exposed by Zaaijer [2].

One of the drawbacks of using the impedance functions is that they are intrinsically defined in the frequency-domain. Implementation into time-domain models allowing non-linear analysis can be done by adopting methodologies such as lumped-parameter models (LPM) [24], as done by Damgaard et al. [10, 11]. They developed different LPM to represent the impedance functions of gravity [10] and monopile [11] foundations and implemented such models in the aeroelastic code HAWC2. By studying the reference NREL 5MW OWT [25], they concluded that the side-side response is more affected by the SSI than the fore-aft vibration for gravity foundations; and that the SSI effects are critical in the design of OWT on monopiles as they have a great impact on the fatigue damage equivalent moment at seabed. Their LPM was used in a later work [12] to study the effect of changes in the soil properties on the system fundamental frequency, damping and fatigue loads in parked conditions. The changes of soil stiffness, soil damping and the presence of sediment transportation at seabed were shown to be critical.

Bisoi and Haldar [6] made use of the p-y curves to represent the soil-pile interaction in a FEM model that included wind and wave loads. They compared three soil profiles (homogeneous, linear and parabolic) obtaining that the natural frequency marginally changes between them. The three profiles presented the same properties at a depth equal to the pile diameter. They also found that the effect of soil non-linearity increases for higher wind speeds. Damgaard et al. [13] obtained the OWT fundamental frequency and damping by using both experimental data from rotor-stop tests and a Winkler approach based on the p-y curves. The Winkler approach together with a hysteresis loop method reasonably estimated the modal soil material damping.

Bhattacharya et al. [14] carried out small scale tests of OWT founded on monopiles and tripods. They illustrated that the natural frequency of the overall system shifts with the number of cycles of loading due to the softening or stiffening of the foundation. Lombardi et al. [7] further studied this relation for monopiles on clay soils, obtaining that the fundamental frequency decreases while the damping increases with the number of cycles of repeated loading. The drop in the natural frequency is higher when the forcing frequency is close to the system natural frequency.

Damgaard et al. [15] studied the influence of the water pore pressure on the estimation of the tower fundamental frequency by combining a Kelvin and a bidimensional FEM models. They compared the numerical results with experimental free-vibration tests, obtaining a better agreement when the permeability of the soil was considered. Yu et al. [16] also investigated the effect of the presence of water in the soil on the dynamic behaviour of OWT founded on monopiles and gravity foundations by executing earthquake centrifuge tests. They demonstrated that the SSI plays a significant role in the seismic behaviour of OWT, affecting the structural settlement, foundation response and fundamental frequency of the system. They remarked the difficulty of the observation and analysis, specially if soil liquefaction is produced.

Bisoi and Haldar [17] addressed the optimization of the structural mass for 2 and 5 MW OWT founded on monopiles in clay. The SLS, fatigue life and resonance avoidance criteria were checked and p-y curves were used to model the soil-pile interaction. They found that the rotor and nacelle mass and the tower height play a crucial role on design, while the embedded depth of the monopile beyond the critical length has a marginal impact. Myers et al. [18] analysed when the strength (resistance in operational and extreme conditions) or stiffness (resonance avoidance) criteria govern the design of monopiles for OWT, and presented optimum pile sections that satisfied these demands. If a fixed base was assumed, the strength criterion controlled the design; but when the soil flexibility was included, the stiffness criterion became important in two of the three studied sites, corresponding to deeper water depths.

Despite most of the studies focus on the tower lateral vibrations, there are several works related to other vibration modes. Kjørlaug and Kaynia [19] studied the vertical seismic response of the NREL 5MW OWT, showing that the tower could amplify up to two times the vertical accelerations at the seabed. On the other hand, Tibaldi et al. [20] showed that, in addition to the structural modes, the blade and additional aeroelastic modes can play a significant role in the structural response during operational conditions.

An interesting associated phenomenon is highlighted, for instance, by Hu et al. [21], who showed the tendency of the system to easily get stuck in resonance, exposing the necessity of a good estimation of the system fundamental frequency in generators with variable rotor speed.

2 PROBLEM DEFINITION

OWT structures founded on monopiles consist of the assembly of three parts: the tower, a transition piece and the pile. The tower is usually a conical hollow beam that supports the rotor and generator nacelle. The monopile is a cylindrical hollow beam that is driven into the seabed, so part of it lays under the soil while another part stays submerged and also can reach the sea level. Both the tower and monopile are connected through a transition piece, being the access point to the structure, usually employed for maintenance and repair activities.

The whole system can be divided into two different parts: the foundation and the (super)structure. The foundation corresponds to the pile length embedded in the seabed and is the part of the system that interact with the soil. On the other hand, the structure corresponds to the part of the system above soil. In some cases, only the structure is analysed during the design stage of the OWT system by considering an infinitely rigid base. If this assumption is made, the dynamic behaviour of the superstructure can be represented by its fixed-base fundamental frequency f_n and modal damping ratio ξ. However, as the soil does not behave as an infinitely rigid body, the SSI effects modify the foundation-structure dynamic properties: reducing the fundamental frequency and changing (increasing or decreasing) the damping ratio. This work aims to quantify these changes by computing the system fundamental frequency $\tilde{f}_n$ and equivalent damping ratio $\tilde{\xi}$ including the SSI effects.

2.1 Structure definition based on real OWT data

As mentioned before, the superstructure consists of three components: tower, transition piece and monopile (above soil). A set of real OWT structures that have been studied in the literature is taken as starting point in order to stablish the dimensions and properties of the systems under analysis. The first group (OWTs 1-12) corresponds to different constructions from wind farms built in the United Kingdom [26]; while OWTs 13-16 are four OWT systems that have been recursively studied by some authors [2–7]. As no detailed information was accessible about the

dimensions of the transition piece, in this study only the tower and pile lengths are considered.

The tower is defined by the tower heigh H_t, tower base D_{bot} and tip D_{top} external diameters, and tower thickness ratio δ_t, which represents the ratio between the inner and outer diameters of the tower cross-section. The thickness ratio is assumed to present the same value along the tower, so thicker walls are located at the tower base. The mass of the rotor and nacelle components is represented by a punctual mass M_{RNA}. Table 1 shows the values of the tower variables for the set of OWT systems analysed.

OWT		M_{RNA} (10^3 kg)	H_t (m)	D_{top} (m)	D_{bot} (m)	δ_t (%)
1-2	Vestas 2MW-V66	80	60-78	2.3	4.2	98.0
3-4	Vestas 3MW-V90	111	8-105	2.3	4.2	98.0
5-6	Vestas 2MW-V80	94	60-100	2.3	4.2	98.0
7-8	Siemens SWT-3.6-107	220	80-96	2.3	4.2	98.0
9-10	Vestas 2MW-V80	94	60-100	2.3	4.2	98.0
11-12	Siemens SWT-3.6-107	220	80-96	2.3	4.2	98.0
13	Lely A2 NM41-2bladed	32	41.5	1.9	3.2	99.0
14	North Hoyle 2MW-V80	100	70	2.3	4.0	97.6
15	Irene Vorrink 600kW	36	51	1.7	3.5	99.0
16	Walney S1 3.6MW	234	83.5	3.0	5.0	97.9

Table 1: Tower dimensions for the studied OWT systems.

The monopile is defined by the above-soil pile height H_p, the pile embedded length L_p, the pile external diameter D_p, and the pile thickness ratio δ_p, which represents the ratio between the inner and outer diameters of the pile cross-section. Table 2 shows the values of the pile variables for the set of OWT systems analysed. Due to lack of specifications, the pile height is assumed to be equal to the water level. Both the tower and monopile are assumed to be made of steel (Young's modulus $E = 210$ GPa, density $\rho = 7850$ kg/m^3 and Poisson's ratio $\nu_p = 0.25$).

OWT		H_p (m)	L_p (m)	D_p (m)	δ_p (%)
1-2	Vestas 2MW-V66	11	15	3.5	97.4
3-4	Vestas 3MW-V90	10	28	4.3	97.9
5-6	Vestas 2MW-V80	20	31	4.2	97.6
7-8	Siemens SWT-3.6-107	19	11	4.7	97.7
9-10	Vestas 2MW-V80	21	33	4.0	98.2
11-12	Siemens SWT-3.6-107	25	30	4.7	97.7
13	Lely A2 NM41-2bladed	4.6	21	3.7	98.1
14	North Hoyle 2MW-V80	11	33	4.0	97.5
15	Irene Vorrink 600kW	3.8	19	3.5	98.4
16	Walney S1 3.6MW	20	31	4.2	97.6

Table 2: Monopile dimensions for the studied OWT systems.

2.2 Foundation definition based on real OWT data

The foundation consists of the portion of the monopile embedded in the seabed. Thus, the same properties that are shown in Table 2 define the dimensions of the foundation of each OWT system.

As can be seen from the collected data, there is no apparent relation between the dimensions of the structure and foundation or between the pile diameter and embedment length. In order to contemplate different possibilities and to study the effect of the foundation dimensions on the SSI effects, a set of three pile diameter values ($D_p = 3, 4$ and 5 m) combined with a set of three pile embedment lengths ($L_p = 15, 25$ and 35 m) will be considered in the parametric analysis carried out in section 5. The pile thickness ratio is obtained as a function of the pile diameter following the API recommendations [27]:

$$\delta_p \approx 98 - \frac{1.274}{D_p} \qquad (\%) \tag{1}$$

Finally, as the Timoshenko beam theory is used to model the pile (see section 3.2), the shear coefficient $\alpha_p = 0.5$ for hollow circular cross-sections is assumed.

2.3 Soil profile definition based on real boreholes

Attempting to study the SSI effects for soil profiles close to real seabed, two typical boreholes (Nelson Field and Hutton TLP) of the North Sea [28] are selected as reference profiles in this work. As these soils consist of different layers of clay and sand, two additional soil profiles formed only by clay or sand are also included in the analyses. The profiles are presented in terms of the shear wave propagation velocity c_s, which is obtained as a function of the material and depth z through the expression provided by Ohta and Goto [29]:

$$c_s(z) = 78.98 z^{0.312} \begin{pmatrix} 1.000 & \text{clay} \\ 1.260 & \text{fine sand} \\ 1.286 & \text{medium sand} \end{pmatrix} \qquad (\text{m/s}) \tag{2}$$

Fig. 1 shows the evolution with depth of the aforementioned soil profiles together with their mean value $c_{s,30}$ [30], as this value is a common choice to characterize the soil. Additionally, a homogeneous profile with $c_s = 180$ m/s (close to the Nelson Field mean shear velocity) is also included in the study in order to contrast whether the homogeneous assumption is a valid option.

For depths greater than 80 m, the shear velocity is kept constant with depth (halfspace domain). For all profiles, depth-independent soil density $\rho_s = 1800$ kg/m^3, soil Poisson's ratio $\nu_s = 0.35$ and soil hysteretic damping ratio $\xi_s = 5\%$ are assumed. Nevertheless, in section 5.1 the influence of the soil Poisson's ratio is studied by varying its value between 0.35-0.49, as higher ratios are commonly used to represent saturated soils.

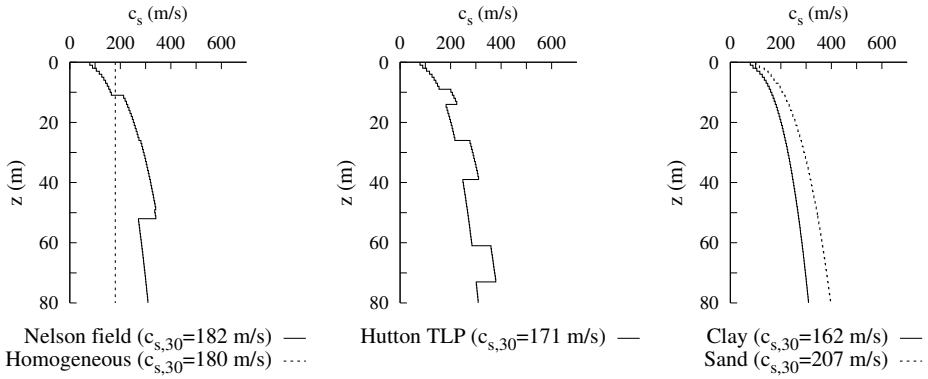


Figure 1: Soils profiles used in the study. Evolution with depth of the shear wave velocity and $c_{s,30}$ mean value.

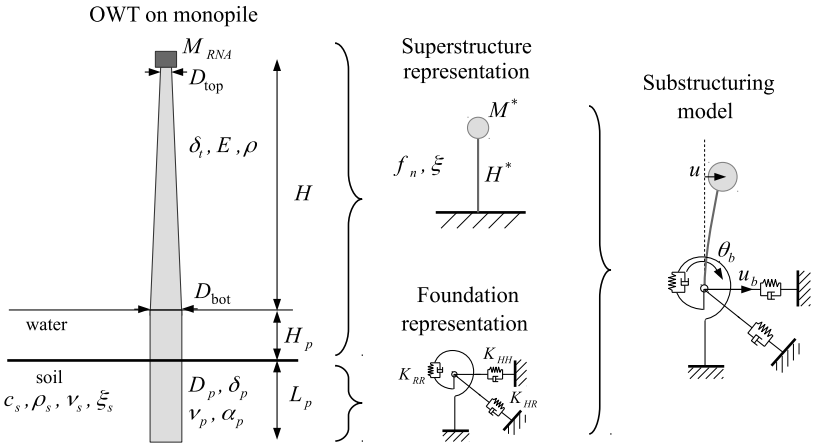


Figure 2: Three-step methodology illustration.

3 METHODOLOGY

Fig. 2 depicts the proposed three-step methodology for the dynamic characterization of the foundation-structure system: on one hand, the superstructure is represented by a single-degree-of-freedom whose mass, height and stiffness are determined by the modal properties of the OWT structure; on the other hand, the monopile foundation stiffness is represented by a set of impedance functions. The scope of this study is to analyse the lateral fundamental mode of vibration, so both steps can be reduced to two-dimensional problems. Finally, both the superstructure and foundation representations are coupled together into the final substructuring model.

3.1 FEM formulation for the modal characterization of the structure

In order to obtain the mass and height that will characterise the superstructure system a modal analysis is made based on a FEM representation of it. The choice of FEM instead of using analytical expressions is made due to the combination of the tapered (tower) and cylindrical (pile) sections of the OWT structure. Furthermore, with the FEM one can easily assume different properties for each element in order to include additional mass or stiffness due to constructive elements. As the effects of the foundations are not included at this step, the structure is assumed to be placed on an infinitely rigid base.

The structure is modelled as Bernoulli beams by using two-node Hermitian elements with four degrees-of-freedom corresponding to the lateral displacements and rotations. An enough number of constant-section elements are used for both the conical and cylindrical lengths, based on a convergence study. The rotor and nacelle punctual mass M_{RNA} is included at the tower tip node. For the submerged pile elements, the hydrodynamic water added mass and the mass of the water inside the pile are also included through a modified density. This mass inclusion does not affect the results of the fundamental mode of vibration [5], even though the system total mass is significantly increased.

Considering time-harmonic displacements and forces, the system equation of motion results in:

$$(\mathbf{K} - \omega\mathbf{M})\,\mathbf{u} = \mathbf{F_{ext}} \tag{3}$$

where $\mathbf{K}$ and $\mathbf{M}$ are the global stiffness and mass matrices obtained by the assembly of the elemental ones; $\mathbf{u}$ is the vector of nodal displacements and rotations and $\mathbf{F_{ext}}$ is the vector of

external nodal loads. The structure modal analysis is made by solving the eigenvalues problem $|\mathbf{K} - \omega^2\mathbf{M}| = 0$, so the fixed-base fundamental frequency ω_n and its modal shape $\boldsymbol{\phi}_n$ are obtained as the smallest eigenvalue and its eigenvector respectively. Then, the base shear effective modal mass M^* and height H^* of the structure to a base acceleration excitation [31] are computed as:

$$M^* = \frac{\left(\boldsymbol{\phi}_n^T\mathbf{M}\boldsymbol{\iota}\right)^2}{\boldsymbol{\phi}_n^T\mathbf{M}\boldsymbol{\phi}_n}; \qquad H^* = \frac{\mathbf{h}^T\mathbf{M}\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n^T\mathbf{M}\boldsymbol{\iota}} \tag{4}$$

being $\boldsymbol{\iota}$ the influence vector and $\mathbf{h}$ the height vector (having the node height in the displacement terms and unitary value in the rotation ones). The base shear effective modal mass represents the mass of a single-degree-of-freedom that is equivalent to the studied system in producing the same base shear force at the fundamental frequency. The base shear effective modal height corresponds to the height of the aforementioned modal mass for which the modal mass vibrating at the fundamental frequency produces the same base overturning moment as the studied system. This modal parameters are chosen to represent the whole structure as the base shear force and moment are the excitation forces that affect the foundation.

3.2 Foundation impedance functions

The foundation stiffness is modelled through a set of four impedance functions: the horizontal K_{HH}, rocking K_{RR} and coupled horizontal-rocking K_{HR}, K_{RH} terms. Each term K_{ij} relates the force or moment in direction i with the displacement or rotation in direction j. The impedance functions are frequency-dependent and its real and imaginary components correspond to the stiffness and damping terms, respectively.

For the computation of the monopile impedance functions, a previously developed three-dimensional time-harmonic model [32] is used. The model represents the soil behaviour through an integral formulation based on the dynamic reciprocal theorem and making use of Green's Functions for the multilayered halfspace [33], while the pile is modelled as a Timoshenko beam though FEM. Both formulations are coupled assuming that the pile is a load line acting within the soil and imposing compatibility (welded boundary condition in the pile-soil interface) and equilibrium (on the pile-soil interaction tractions) conditions, resulting in a system of linear equations where the pile displacements, pile head external forces and soil-pile interaction tractions are the system unknowns. In order to compute each impedance term, the corresponding pile head displacements are prescribed and the system of equations is solved for the pile head forces.

3.3 Substructuring simplified model for the dynamic characterization of the system

At this point, the structure and foundation have been reduced into a single-degree-of-freedom mass and a set of impedance functions respectively. Both are coupled together resulting in a three degrees-of-freedom model where the unknowns are the mass deflection u relative to the base and the base lateral displacement u_b and rotation θ_b (see Fig. 2). In order to find the dynamic characteristics of the system, a ground lateral acceleration is assumed as the external excitation yielding the following equations of motion:

$$\begin{cases} K^*\, u - M^*\omega^2\left(u + u_b + H^*\theta_b\right) = -M^*\ddot{u}_g \\ K_{HH}\, u_b + K_{HR}\, \theta_b - M^*\omega^2\left(u + u_b + H^*\theta_b\right) = -M^*\ddot{u}_g \\ K_{RH}\, u_b + K_{RR}\, \theta_b - M^*\omega^2 H^*\left(u + u_b + H^*\theta_b\right) = -M^*\ddot{u}_g H^* \end{cases} \tag{5}$$

The term $K^* = \omega_n^2 M^*(1 + 2\mathrm{i}\xi)$ represents the structural lateral stiffness corresponding to the fundamental mode. The structural damping is also represented by its complex component as the modal damping factor ξ is included. The lateral ground acceleration $\ddot{u}_g$ is used as the system excitation as the modal mass and height are obtained under the same assumption.

An equivalent single-degree-of-freedom oscillator reproducing the system response is searched for computing the flexible-base fundamental frequency $\tilde{\omega}_n$ and damping $\tilde{\xi}$ of the system [34, 35]. An hysteretically damped oscillator is assumed and the equivalence is established in terms of the maximum value of the base shear force per effective seismic force [31] transfer function :

$$Q(\omega) = \left| \frac{\omega_n^2 u}{\ddot{u}_g} \right| \approx \left| \frac{1}{(1 - \omega^2/\tilde{\omega}_n^2) + 2\mathrm{i}\tilde{\xi}} \right| \tag{6}$$

By iteratively solving eq. 5, the flexible-base fundamental frequency is obtained as the frequency at which the maximum value of the transfer function Q_m is produced, while the equivalent damping factor is calculated as $\tilde{\xi} = 1/(2Q_m)$.

The proposed simplified model is validated against an enhanced FEM model including the foundation stiffness. For that purpose, eq. 3 is modified by adding into the stiffness matrix the foundation impedance functions in the terms corresponding to the ground node, and by assuming an horizontal base acceleration as excitation $\mathbf{F_{ext}} = -M\boldsymbol{\iota}\ddot{u}_g$. Then, the flexible-base fundamental frequency is obtained as the frequency where the maximum value of $Q(\omega)$ takes place. Differences below 1% are produced for the studied soils between the proposed methodology and the enhanced FEM formulation, revealing the ability of the three-step procedure to correctly capture the foundation-structure system fundamental frequency.

4 MODAL CHARACTERIZATION OF THE STUDIED OWT STRUCTURES

Table 3 shows the modal parameters obtained for the set of studied OWT structures once the procedure presented in section 3.1 is applied. Most of the constructions present fixed-base fundamental frequencies between 0.2-0.5 Hz, agreeing with the typical range for medium-sized OWT constructions. Additionally, OWTs 14 and 16 present higher fundamental frequencies due to their limited size (see Table 1) that makes them more rigid structures. The modal height and mass are strongly related to the physical dimensions of the structure, being the first about 85-90% of the system total height and the later around 30% of the system total mass.

OWT	f_n (Hz)	H^* (m)	M^* (10^3 kg)	OWT	f_n (Hz)	H^* (m)	M^* (10^3 kg)
1	0.53	60.8	170	9	0.42	65.2	236
2	0.37	75.4	188	10	0.22	98.7	258
3	0.35	80.1	206	11	0.24	94.3	355
4	0.23	100	232	12	0.19	109	363
5	0.49	66.3	216	13	0.83	43.2	46
6	0.24	100	239	14	0.42	70.8	203
7	0.25	91.0	329	15	0.61	51.3	51
8	0.19	105	341	16	0.26	89.1	469

Table 3: Modal parameters for the studied OWT systems.

A strong relation between the modal parameters and the fixed-base fundamental frequency is found when analysing the obtained results. Thus, a second order polynomial fitting can be used to express the modal mass and height as functions of the fixed-base fundamental frequency

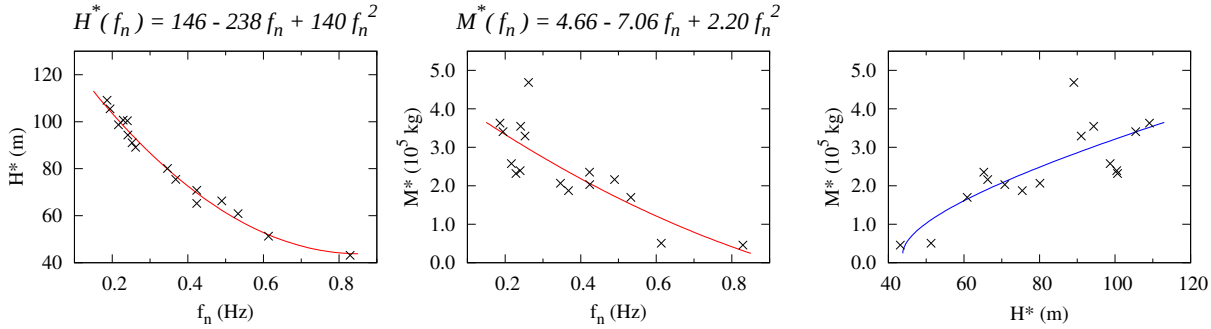


Figure 3: Modal parameters for the studied set of OWTs (crosses). Modal height and mass fitting expressions as a function of the fixed-base fundamental frequency (red lines). Relation between the modal height and mass through the fitting expressions (blue lines).

without significant errors, as shown in Fig. 3. The proposed polynomials allow to study a range of structures that follow the typical dimensions for OWT on monopile constructions instead of just a finite set of systems.

For illustration purposes, the variations in the system fundamental frequency and damping due to the SSI effects assuming a foundation consisting of a pile with $L_p = 25$ m and $D_p = 4$ m embedded in the Nelson Field profile are plotted in Fig 4 against the fixed-base fundamental frequency. The crosses are obtained by using the real modal parameters of the studied OWT structures, while the continuous lines are obtained by assuming the relation between the modal mass or height and the fixed-base fundamental frequency given by the fitting polynomials. The results show that the use of the fitting expressions adequately follows the behaviour of the real structures.

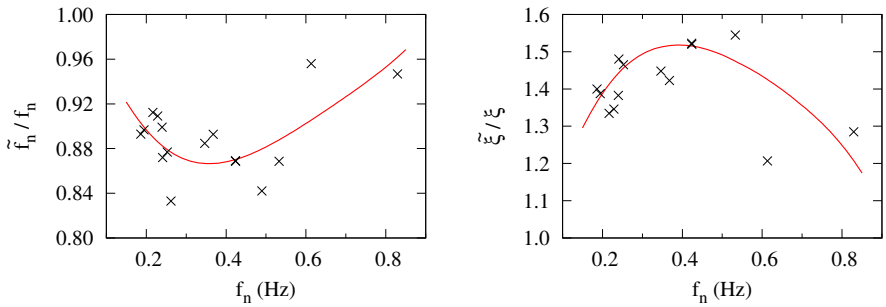


Figure 4: Variations in the system fundamental frequency and damping due to SSI effects. Comparison between the results of the fitted (line) and actual (crosses) modal parameters. Pile foundation with $D_p = 4$ m and $L_p = 25$ m embedded in the Nelson Field soil profile. Structural modal damping $\xi = 1\%$.

4.1 Influence of the modal parameters on the system dynamic properties

In order to explain the results obtained in Fig. 4, the effects of each one of the modal parameters have to be analysed individually. Fig. 5 presents the variations in the system fundamental frequency and damping due to the SSI effects assuming a fixed value of two of the modal parameters while varying the other. Different pairs of constant parameters are employed in order to contemplate different combinations and their values are chosen in order to be within the studied ones. The same foundation dimensions and soil profile as in Fig. 4 are considered.

Attending to Fig. 5, all modal parameters increase the SSI effects if their values augment. On one hand, an increment in the system stiffness is produced if the modal mass or fixed-base

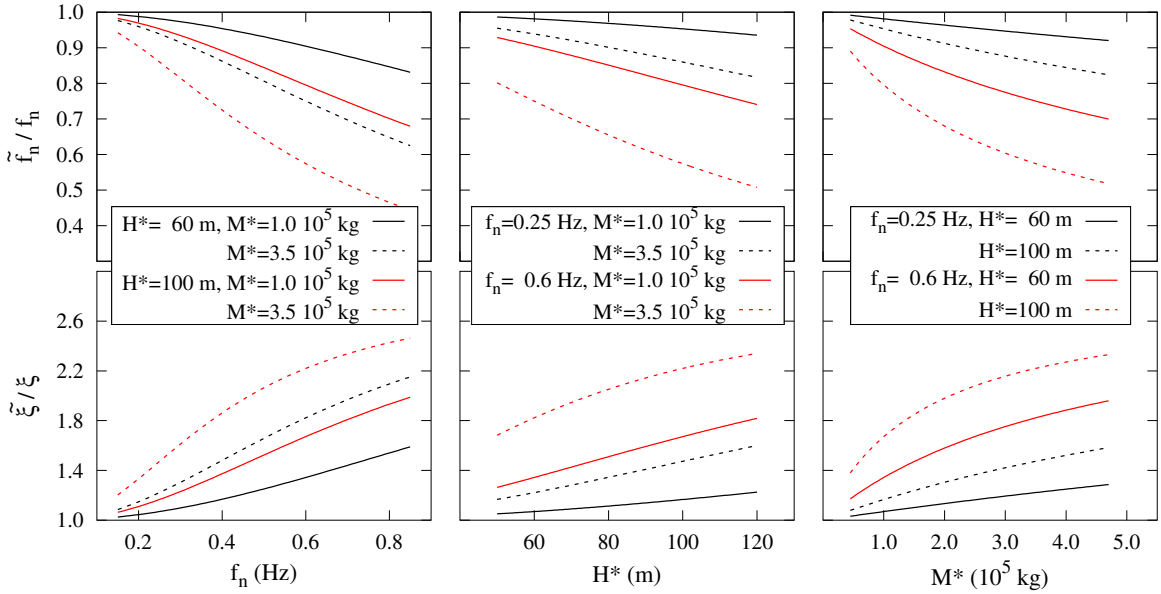


Figure 5: Influence of the modal parameters on the fundamental frequency and damping variations. Pile foundation with $D_p = 4$ m and $L_p = 25$ m embedded in the Nelson Field soil profile. Structural modal damping $\xi = 1\%$.

fundamental frequency increase while the other remains unaltered. Then, if the system stiffness augments, the soil will be relatively softer yielding higher frequency and damping variations. On the other hand, the increment of the SSI effects with the structural height has been reported in numerous studies, eg. [5, 34]. Now, the shape of the previous curves can be explained considering that in the their first part, the SSI effects increase owing to the increment in the fixed-base fundamental frequency, while in the last part the effect of the reduction in the modal mass and height has more importance.

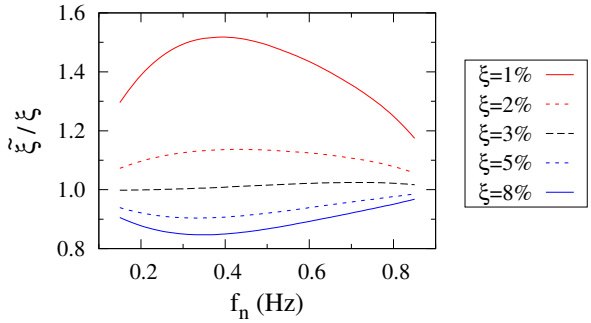


Figure 6: Influence of the fixed-base structural modal damping on the damping variations. Pile foundation with $D_p = 4$ m and $L_p = 25$ m embedded in the Nelson Field soil profile.

Finally, in Figs. 4 and 5 a structural modal damping $\xi = 1\%$ was assumed, as a typical value for studying OWT structures [36]. However, if another damping ratio is considered, the SSI effects does not change it to the same extent, as illustrated by Fig. 6. The figure shows the relation between the flexible and fixed-base damping ratios that is obtained by assuming different modal damping ratios (ξ = 1-8%). Again, the same foundation and soil as in the previous results are considered. For a modal damping $\xi = 3\%$ virtually no variations are seen. As the modal damping goes away from this value, the variations become stronger: increasing the flexible-base damping ratio for $\xi < 3\%$ and decreasing it otherwise. Nevertheless, the

frequency range at which the system is more sensitive to the SSI effects, i.e. stronger variations, is almost the same.

5 INFLUENCE OF THE SSI EFFECTS ON THE SYSTEM DYNAMIC PROPERTIES

Figures 7 and 8 present the variations in the fundamental frequency and damping of the system due to the SSI effects as a function of the structure fixed-base fundamental frequency. The pile dimensions and profiles introduced in section 2 are used in order to define the foundation characteristics, while the polynomial fittings of the modal parameters presented in section 4 together with a modal damping $\xi = 1\%$ are used to define the structures.

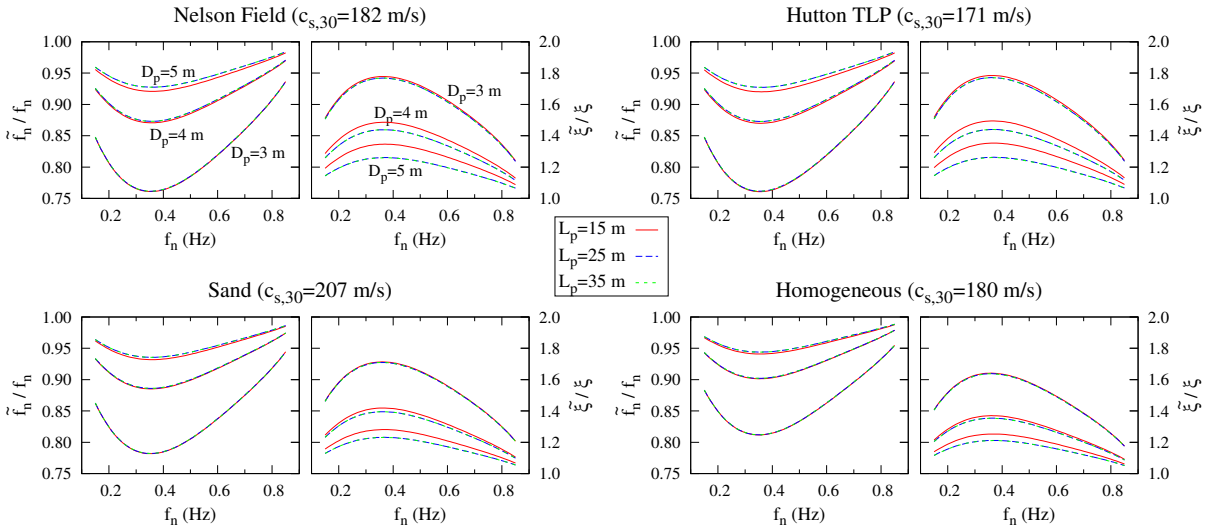


Figure 7: Variations in the system fundamental frequency and damping due to SSI effects. Results grouped by soil profile.

First, Fig. 7 shows the results corresponding to the same soil profile in each graph in order to manifest the effects of the monopile dimensions. Attending to the results, the pile diameter is found to play a significant role as its dimension is strongly related to the foundation stiffness: the higher the diameter, the stiffer the foundation yielding lower variations in the system frequency and damping. On the contrary, the pile length plays a minor role in the magnitude of the SSI effects as the results slightly vary between one value to other, being the damping variations more affected by this factor. The small influence of the pile length is explained assuming that the highest lengths are over the pile critical length.

On the other hand, Fig. 8 shows the results corresponding to all the studied soils in the same graph in order to analyse the differences between each profile. As mention before, the results marginally change depending on the pile length, thus only the curves corresponding to one length value are presented. The results for the Nelson Field, Hutton TLP and Clay profiles are virtually the same, showing that the superficial layers of the profiles are the ones that govern the studied problem. As expected, for harder soils (in terms of the c_s of the superficial layers), the variations in the system fundamental frequency and damping are reduced. It is important to notice that significant differences are produced between the variable profiles and the homogeneous assumption, despite that all soils present close values of the $c_{s,30}$ mean velocity.

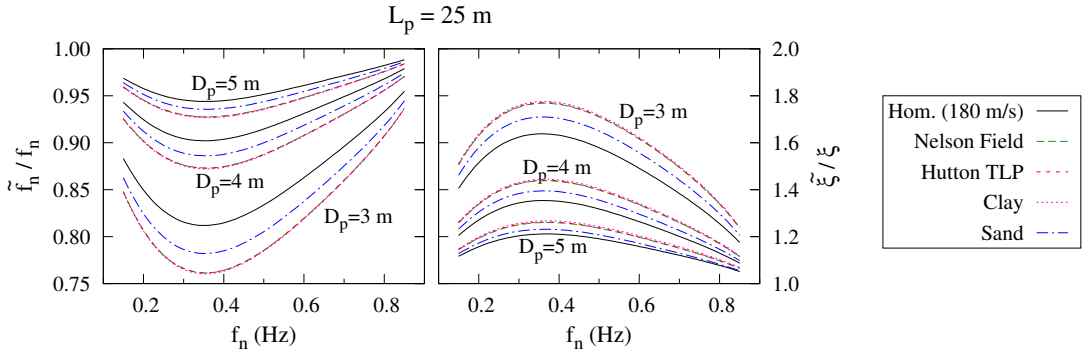


Figure 8: Variations in the system fundamental frequency and damping due to SSI effects. Results grouped by pile length.

5.1 Influence of soil Poisson's ratio

Fig. 9 presents the variations in the fundamental frequency and damping of the OWT system for the Nelson Field profile and considering different values of the soil Poisson's ratio. This analysis is made in order to show the influence of this parameter on the SSI effects, as high Poisson's ratios are commonly used to represent saturated soils. Values of ν_s between 0.35-0.49 are used, while the rest of soil properties have the same values introduced in section 2.3.

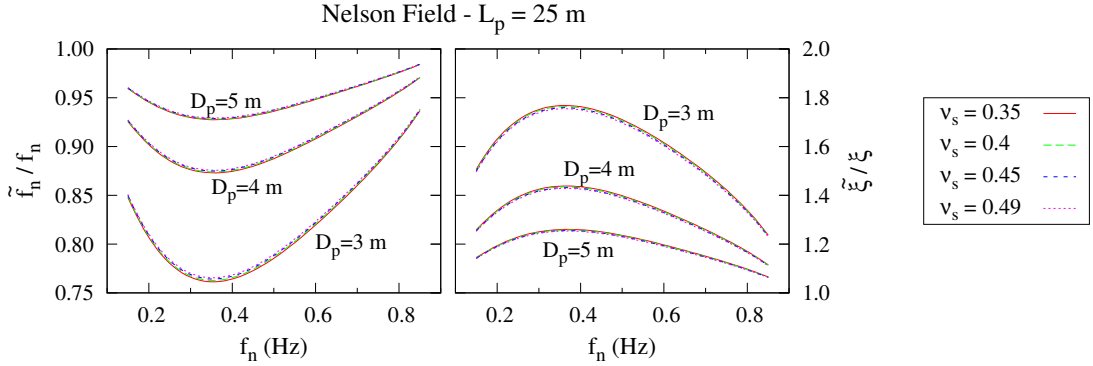


Figure 9: Variations in the system fundamental frequency and damping due to SSI effects. Influence of the soil Poisson's ratio.

The results of Fig. 9 show that increasing the soil Poisson's ratio has almost no influence on the SSI effects, as recently found by Daamgard et al. [12]. As the Poisson's ratio augments, the soil becomes slightly more rigid and, therefore, the variations in the system fundamental frequency and damping are marginally reduced. However, practically the same results are obtained for the extreme cases of $\nu_s = 0.35$ and 0.49. This effect is seen for all the studied profiles.

6 CONCLUSIONS

This work introduces a simplified substructuring model to analyse variations in the fundamental frequency and damping of OWT-monopile systems due to the SSI effects. Only the fundamental mode of the superstructure is taken into account in terms of the base shear effective modal mass and height, while the foundation stiffness modelling is addressed through impedance functions. The proposed methodology agrees well with the results of a more elaborated FEM formulation that takes all modes into account.

The effect of real soil profiles are studied based on typical boreholes of the North Sea, while relations between the modal mass or height and the fixed-base fundamental frequency that are characteristic for the studied type of constructions are obtained based on data from real OWT systems found in the literature. The results showed reductions of 5-20% in the system fundamental frequency, revealing the importance of the consideration of the SSI effects when analysing the dynamic behaviour of this type of structures.

Finally, the soil profile is found to play a fundamental role in the magnitude of the SSI effects, specially the superficial layers. The use of a homogeneous soil with mean properties close to the actual profile is not valid to study the SSI effects as significantly smaller variations in the system fundamental frequency and damping are produced under this assumption.

ACKNOWLEDGEMENTS

This work was supported by the Subdirección General de Proyectos de Investigación of the Ministerio de Economía y Competitividad (MINECO) of Spain and FEDER through research project BIA2014-57640-R. G.M. Álamo is a recipient of the FPU research fellowship FPU14/06115 from the Ministerio de Educación, Cultura y Deporte of Spain. The authors are grateful for this support. The authors also want to thank Esteban García, former Master student at University Institute SIANI for his data compilation.

References

[1] DNV/Risø. *Guidelines for design of Wind Turbines. 2nd ed.* Det Norske Veritas, Copenhagen and Wind Energy Department, Risø National Laboratory, 2002.

[2] M.B. Zaaijer. Foundation modelling to assess dynamic behaviour of offshore wind turbines. *Applied Ocean Research*, **28**(1):45–57, 2006.

[3] S. Adhikari and S. Bhattacharya. Dynamic analysis of wind turbine towers on flexible foundations. *Shock and Vibration*, **19**(1):37–56, 2012.

[4] L. Arany, S. Bhattacharya, S. Adhikari, S.J. Hogan, and J.H.G. Macdonald. An analytical model to predict the natural frequency of offshore wind turbines on three-spring flexible foundations using two different beam models. *Soil Dynamics and Earthquake Engineering*, **74**:40–45, 2015.

[5] V. Zania. Natural vibration frequency and damping of slender structures founded on monopiles. *Soil Dynamics and Earthquake Engineering*, **59**:8–20, 2014.

[6] S. Bisoi and S. Haldar. Dynamic analysis of offshore wind turbine in clay considering soil-monopile-tower interaction. *Soil Dynamics and Earthquake Engineering*, **63**:19–35, 2014.

[7] D. Lombardi, S. Bhattacharya, and D. Muir Wood. Dynamic soil-structure interaction of monopile supported wind turbines in cohesive soil. *Soil Dynamics and Earthquake Engineering*, **49**:165–180, 2013.

[8] W. Carswell, S.R. Arwade, D.J. DeGroot, and M.A. Lackner. Soil-structure reliability of offshore wind turbine monopile foundations. *Wind Energy*, **18**(3):483–498, 2015.

[9] S. Bhattacharya and S. Adhikari. Experimental validation of soil-structure interaction of offshore wind turbines. *Soil Dynamics and Earthquake Engineering*, **31**(5-6):805–816, 2011.

[10] M. Damgaard, L.V. Andersen, and L.B. Ibsen. Computationally efficient modelling of dynamic soil-structure interaction of offshore wind turbines on gravity footings. *Renewable Energy*, **68**:289–303, 2014.

[11] M. Damgaard, V. Zania, L.V Andersen, and L.B. Ibsen. Effects of soil-structure interaction on real time dynamic response of offshore wind turbines on monopiles. *Engineering Structures*, **75**:388–401, 2014.

[12] M. Damgaard, L.V. Andersen, and L.B. Ibsen. Dynamic response sensitivity of an offshore wind turbine for varying subsoil conditions. *Ocean Engineering*, **101**:227–234, 2015.

[13] M. Damgaard, L.B. Ibsen, L.V. Andersen, and J.K.F. Andersen. Cross-wind modal properties of offshore wind turbines identified by full scale testing. *Journal of Wind Engineering and Industrial Aerodynamics*, **116**:94–108, 2013.

[14] S. Bhattacharya, N. Nikitas, J. Garnsey, N.A. Alexander, J. Cox, D. Lombardi, D. Muir Wood, and D.F.T. Nash. Observed dynamic soil-structure interaction in scale testing of offshore wind turbine foundations. *Soil Dynamics and Earthquake Engineering*, **54**:47–60, 2013.

[15] M. Damgaard, M. Bayat, L.V. Andersen, and L.B. Ibsen. Assessment of the dynamic behaviour of saturated soil subjected to cyclic loading from offshore monopile wind turbine foundations. *Computers and Geotechnics*, **61**:116–126, 2014.

[16] H. Yu, X. Zeng, B. Li, and J. Lian. Centrifuge modeling of offshore wind foundations under earthquake loading. *Soil Dynamics and Earthquake Engineering*, **77**:402–415, 2015.

[17] S. Bisoi and S. Haldar. Design of monopile supported offshore wind turbine in clay considering dynamic soil-structure-interaction. *Soil Dynamics and Earthquake Engineering*, **73**:103–117, 2015.

[18] A.T. Myers, S.R. Arwade, V. Valamanesh, S. Hallowell, and W. Carswell. Strength, stiffness, resonance and the design of offshore wind turbine monopiles. *Engineering Structures*, **100**:332–341, 2015.

[19] R.A. Kjørlaug and A.M. Kaynia. Vertical earthquake response of megawatt-sized wind turbine with soil-structure interaction effects. *Earthquake Engineering & Structural Dynamics*, **44**(13):2341–2358, 2015.

[20] C. Tibaldi, T. Kim, T.J. Larsen, F. Rasmussen, R. de Rocca Serra, and F. Sanz. An investigation on wind turbine resonant vibrations. *Wind Energy*, 2015.

[21] W. Hu, S. Thöns, S. Said, and W. Rücker. Resonance phenomenon in a wind turbine system under operational conditions. In *Proceedings of the 9th International Conference on Structural Dynamics, EURODYN 2014, Porto*, pages 3619–3626, 2014.

[22] J. van der Tempel and D.P. Molenaar. Wind turbine structural dynamics - a review of the principles for modern power generation, onshore and offshore. *Wind Engineering*, **26**(4): 211–222, 2002.

[23] M. Novak and T. Nogami. Soil-pile interaction in horizontal vibration. *Earthquake Engineering & Structural Dynamics*, **5**(3):263–281, 1977.

[24] J.P. Wolf. *Foundation vibration analysis using simple physical models*. Prentice-Hall, Englewood Cliffs NJ, 1994.

[25] J. Jonkman, S. Butterfield, W. Musial, and G. Scott. *Definition of a 5-MW reference wind turbine for offshore system development. Technical Report NREL/TP-500-38060*. National Renewable Energy Laboratory, 2009.

[26] D. Lombardi. *Dynamics of Offshore Wind Turbines*. MSc Thesis, University of Bristol, 2010.

[27] *Recommended Practice for Planning, Designing, and Constructing Fixed Offshore Platforms - Working Stress Design, Ver. 21*. American Petroleum Institute, USA, 2000.

[28] Health and Safety Executive, London (UK), A J Bond, D W Hight, and R J Jardine. *Design of Piles in Sand in the UK Sector of the North Sea - Report OTH 94 457*. 1997.

[29] Y. Ohta and N. Goto. Empirical shear wave velocity equations in terms of characteristic soil indexes. *Earthquake Engineering & Structural Dynamics*, **6**(2):167–187, 1978.

[30] *Eurocode 8: Design of structures for earthquake resistance. Part 5: Foundations, Retaining Structures and Geotechnical Aspects*. European Committee for Standardization, Brussels, 2004.

[31] A.K. Chopra. *Dynamic of structures. Theory and applications to earthquake engineering*. NJ: Prentice-Hall, 2001.

[32] G.M. Álamo, A.E. Martínez-Castro, L.A. Padrón, J.J. Aznárez, R. Gallego, and O. Maeso. Efficient numerical model for the computation of impedance functions of inclined pile groups in layered soils. *Submitted to Engineering Structures*.

[33] R.Y.S. Pak and B.B. Guzina. Three-dimensional Green's functions for a multilayered half-space in displacement potentials. *Journal of Engineering Mechanics*, **128**(4):449–461, 2002.

[34] A.S. Veletsos and J.W. Meek. Dynamic behaviour of building-foundation systems. *Earthquake Engineering & Structural Dynamics*, **3**(2):121–138, 1974.

[35] C. Medina, J.J. Aznárez, L.A. Padrón, and O. Maeso. Effects of soil-structure interaction on the dynamic properties and seismic response of piled structures. *Soil Dynamics and Earthquake Engineering*, **53**:160–175, 2013.

[36] *Guideline for the Certification of Offshore Wind Turbines*. Germanischer Lloyd WindEnergie GmbH, Germany, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

DYNAMIC ANALYSIS OF SPECIAL CARS ON UNEVEN ROADS

Szymon Tengler

Department of Mechanics, University of Bielsko-Biala,
Willowa 2, 43-309 Bielsko-Biala, Poland
e-mail: stengler@ath.bielsko.pl

Keywords: dynamics analysis, special cars, uneven road surface; contact point.

Abstract. *The method of the dynamic analysis of special cars with a high gravity centre is presented in this article. The method proposed by the authors includes a comprehensive mathematical model of the car, and this model is supplemented with the models of uneven road surface, algorithms to determine the contact point between the tire and the road surface, and programs to perform computer simulations. In the article the mathematical model of the cooperation between road surface and tires are discussed in details. As it is known the road acts on the tire in a form of the reaction forces which should be applied in the concrete point named "contact point C". Thus, it is very important to determine this point location correctly. For this purpose the special mathematical models of the road surface and the algorithms to define the contact point location are developed. Two mathematical models of the road are discussed – continuous and discrete one. Two algorithms to find the location of the contact point are presented, as well. According to the conception of the authors names of the algorithms refer to the essence of the matter of the procedures assumed. The first of them – named Plane – can be used considering the continuous model of the surface, and the second – named 4Points – in the case of the discrete model of this surface.*

1 INTRODUCTION

While designing special cars with a high gravity centre (Figure 1), which move along the uneven road surface and are prone to a loss of stability (e.g. rollover), an essential issue is to take a preparation phase into account and to investigate a mathematical model of a constructed car.

Figure 1 Technical rescue vehicle of fire service as an example of a special vehicle with a high gravity centre

While considering motion of these type of cars on uneven road surface an important issue is to determine cooperation between road surface and tires. In this case a mathematical road surface model and also special algorithms for determine a location of contact point are needed.

Computer simulations of a motion of this model, performed in a case of different variants of constraints, can predict a lot of potential threats, and the conclusions drawn from them can constitute important guidelines for car designers. It is important for the prepared mathematical model of the car to be a sufficiently accurate representation of the real system. According to the car designers the computer simulations of an appropriately prepared mathematical model enable to shorten time of testing on a real prototype of the car, and also to reduce costs of the designing process.

2 MATHEMATICAL FORMALISM

For needs of the dynamics analysis, vehicles with high gravity centre can be modeled as multibody systems in a form of open kinematic chains. In this case the location (the position and orientation) of particular bodies should be know. Authors of the proposed method use joint coordinates basing on the approach applied in robotics. The proposed method is based on use of the homogenous transformation matrices with dimensions 4x4, which enable to make transformations between the assumed coordinate systems (related with bodies) [1].

If a position of any point A in the given coordinate system $\{j\}$, expressed by position vector ${}^{\{j\}}\mathbf{r}_A$ of dimensions 3x1, is known, then the position of this point in the coordinate system $\{i\}$ can be determined by the position vector ${}^{\{i\}}\mathbf{r}_A$ of dimensions 3x1 (Figure 2) using only one arithmetic operation, namely multiplication:

$$ {}^{\{i\}}\mathbf{r}_A^* = {}^{\{i\}}_{\{j\}}\mathbf{T}\,{}^{\{j\}}\mathbf{r}_A^* \tag{1} $$

where: ${}^{\{i\}}\mathbf{r}_A^* = \begin{bmatrix} {}^{\{i\}}\mathbf{r}_A \\ 1 \end{bmatrix}$, ${}^{\{j\}}\mathbf{r}_A^* = \begin{bmatrix} {}^{\{j\}}\mathbf{r}_A \\ 1 \end{bmatrix}$ - position vectors of dimensions 4x1, named vectors of homogenous coordinates, determining the position of point A in the system $\{i\}$ and $\{j\}$, respectively,

${}^{\{i\}}_{\{j\}}\mathbf{T} = \begin{bmatrix} {}^{\{i\}}_{\{j\}}\mathbf{R} & {}^{\{i\}}\mathbf{r}_j \\ 0 \quad 0 \quad 0 & 1 \end{bmatrix}$ – the transformation matrix of dimensions 4x4 from the coordinate system $\{j\}$ to the system $\{i\}$,

${}^{\{i\}}_{\{j\}}\mathbf{R} = \begin{bmatrix} \hat{X}_j \cdot \hat{X}_i & \hat{Y}_j \cdot \hat{X}_i & \hat{Z}_j \cdot \hat{X}_i \\ \hat{X}_j \cdot \hat{Y}_i & \hat{Y}_j \cdot \hat{Y}_i & \hat{Z}_j \cdot \hat{Y}_i \\ \hat{X}_j \cdot \hat{Z}_i & \hat{Y}_j \cdot \hat{Z}_i & \hat{Z}_j \cdot \hat{Z}_i \end{bmatrix}$ – the rotation matrix of dimensions 3x3 from the coordinate system $\{j\}$ to the system $\{i\}$ (elements of this matrix are dot products of the versors).

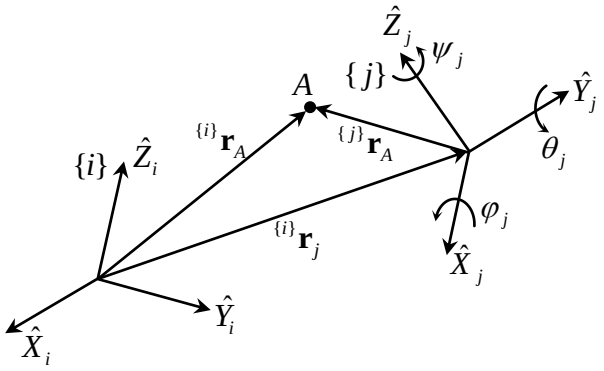


Figure 2. Determining the A point position in the coordinate system $\{i\}$ and $\{j\}$

3 MATHEMATICAL MODEL OF THE CAR

In the vehicle model, assumed in the form of open kinematic chain, 12 subsystems as rigid bodies were distinguished: frame, cabin, engine, car body, front and rear axle, two axle stubs, wheels connected by 6 spring-damper elements, respectively (Figure 3).

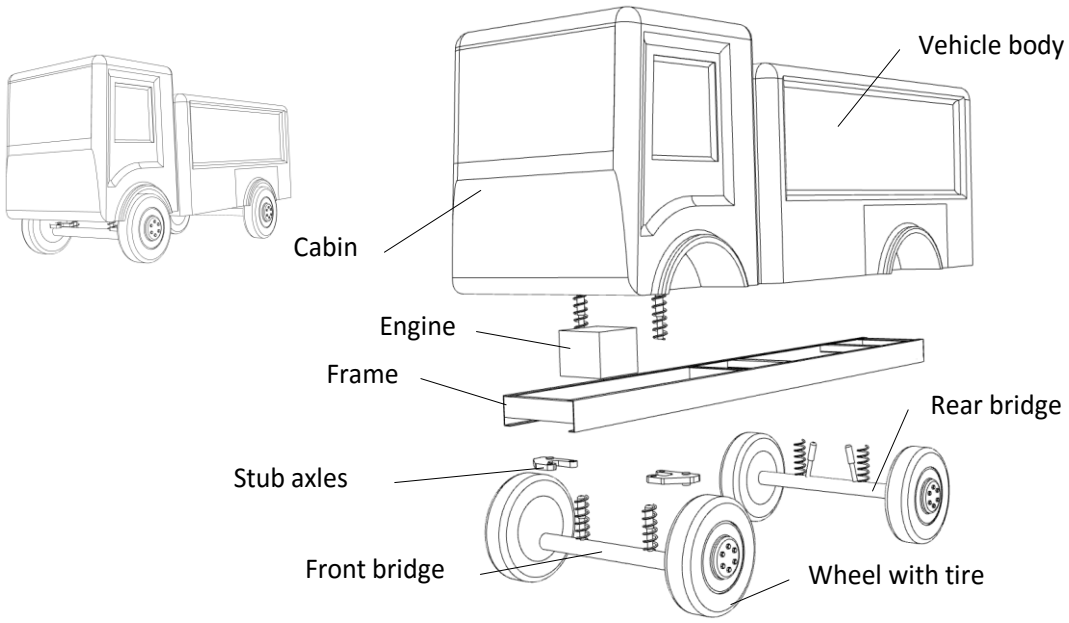


Figure 3 Components of the modelled car

The structure of the discussed chain can be considered as a tree structure, in which each body has a determined number of degrees of freedom (dof) in relation to the preceding body (see Figure 4). The vehicle model in question has 19 degrees of freedom.

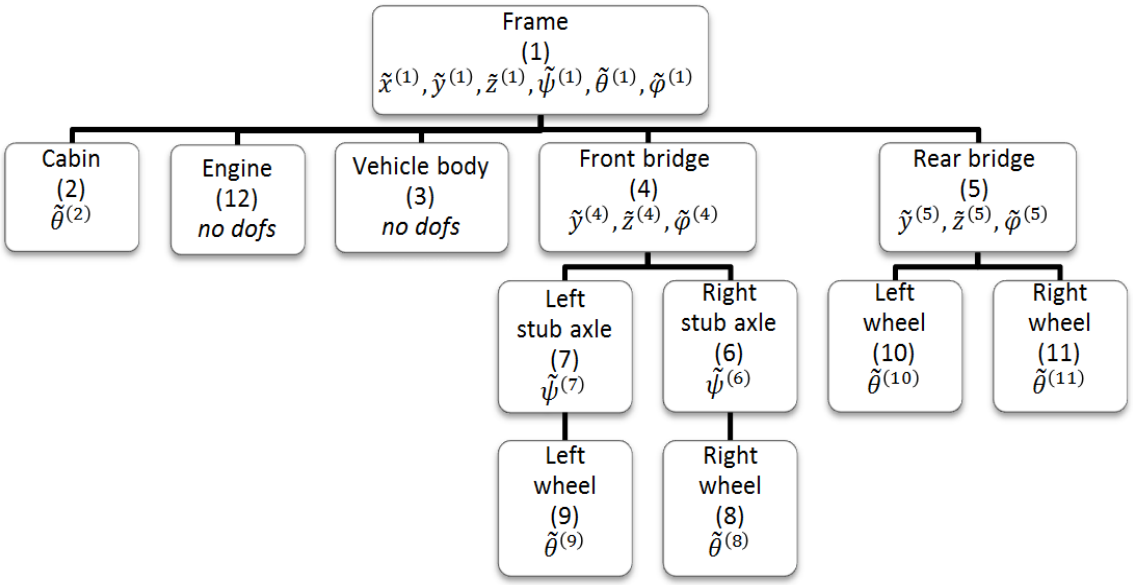


Figure 4 Models of the sub-assemblies forming the tree structure of the considered multibody system, where: (p) – body number in the multibody system; $\tilde{x}^{(p)}, \tilde{y}^{(p)}, z^{(p)}, \tilde{\psi}^{(p)}, \tilde{\theta}^{(p)}, \tilde{\varphi}^{(p)}$ – generalized coordinates of this body $(p = 1, \ldots, 12)$ described in relation to the preceding body

The frame model is a root of the tree, and its motion is described in respect to a global reference coordinate system $\{0\}$. the generalized coordinates of the frame model are displacements of the origin of the coordinate system $\{1\}$ – longitudinal $\tilde{x}^{(1)}$, lateral $\tilde{y}^{(1)}$, and vertical $\tilde{z}^{(1)}$ determined in the coordinate system $\{\square\}$ and the successive angles of rotation of this model around the axis (unit vectors) $\hat{Z}^{(1)}$, $\hat{Y}^{(1)}$, and $\hat{X}^{(1)}$ of the coordinate system $\{1\}$ – yaw $\tilde{\psi}^{(1)}$, pitch $\tilde{\theta}^{(1)}$,

and roll $\tilde{\varphi}^{(1)}$. The other models of sub-assemblies are connected directly or indirectly with the frame respectively.
In the doctoral dissertation [10], the detailed description of the process of mathematical modeling of the analyzed vehicle leading to formulation of the equations of motion is presented, and they can be written as:

$$\begin{cases} \boldsymbol{A}(t,\boldsymbol{q})\ddot{\boldsymbol{q}} - \boldsymbol{D}\boldsymbol{r} = \boldsymbol{f}(t,\boldsymbol{q},\dot{\boldsymbol{q}}) \\ \boldsymbol{D}^T\ddot{\boldsymbol{q}} = \boldsymbol{w} \end{cases} \tag{2}$$

where: t – time, $\boldsymbol{q}, \dot{\boldsymbol{q}}, \ddot{\boldsymbol{q}}$ – vectors of generalized coordinates, velocities and accelerations of considered mathematical model; $\boldsymbol{A}$ – mass matrix; $\boldsymbol{D}$ – matrix of coefficients corresponding to particular reaction forces; $\boldsymbol{r}$ – vector of unknown reaction forces; $\boldsymbol{f}$ – vector of generalized forces and derivatives of kinetic energy, potential energy, Rayleigh's dissipation function; $\boldsymbol{w}$ – vector of right sides of constraint equations.
To determine reaction forces and moments of interacting of the road surface on the vehicle wheel tires Pacejka's tire model was used [8]. Those reactions are applied at a contact point C of the tire model with the road surface. Detailed algorithms for modeling of the road surface, and algorithms which allow determining the position of contact point C in the coordinate system {□} are presented in next section.

4 MATHEMATICAL MODEL OF THE UNEVEN ROADS

For the purpose of the developed method, two ways of road surface modeling including occurrence of possible unevenness of any shape, were proposed.

4.1 The continuous model of the road surface

In the continuous model, the Bicubic interpolation [5] was used for the mathematical representation of the road surface, thus for determining the three dimensional interpolation surface. The input data are here the points in the three dimensional space – so called control points (interpolation nodes). The points located between these nodes are searched.
Let $P'_{i,j}$ and P' mean the xy plane projections of the control point $P_{i,j}(x_i, y_i, z_i)$ and the searched point $P(x_P, y_P, z_P)$ respectively, being on the interpolated surface, of which first two coordinates x_P and y_P are known. Then, the third coordinate z_P of the searched P point is determined by transforming the given grid cell specified by points $P'_{i,j}$, $P'_{i+1,j}$, $P'_{i+1,j+1}$ $P'_{i,j+1}$, into the square of the side length equal to 1 (Figure 5).

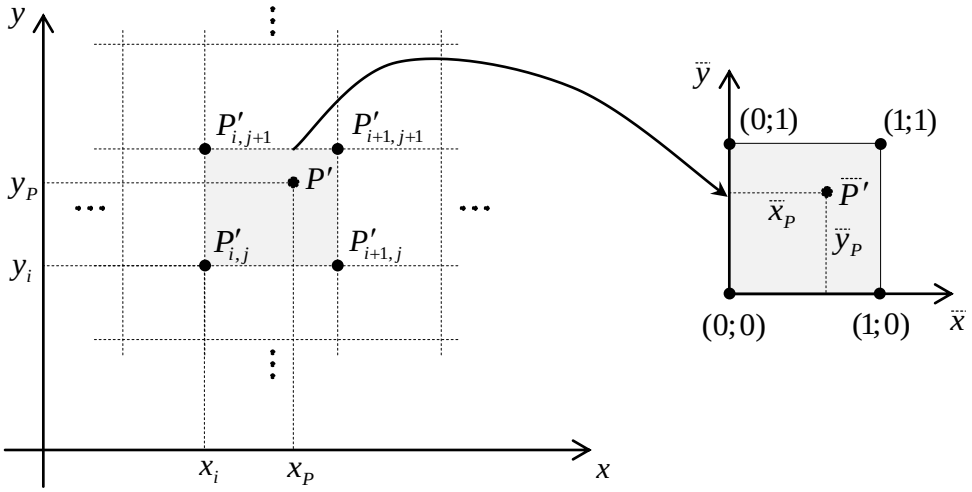


Figure 5 Transforming the selected grid cell

After transformations, coordinate z_P is calculated according to the formula:

$$z_P = f(x_P, y_P) = \bar{z}_P = \bar{f}(\bar{x}_P, \bar{y}_P) = \sum_{i=0}^{3}\sum_{j-0}^{3} a_{i,j}\bar{x}_P^i \bar{y}_P^i \,, \tag{3}$$

where: $\left.\begin{array}{l} \bar{x}_P = \dfrac{x_P - x_i}{x_{i+1} - x_i} \\ \bar{y}_P = \dfrac{y_P - y_j}{y_{i+1} - y_j} \end{array}\right\} \in \langle 0;1 \rangle$ $\left.\begin{array}{l} \bar{x}_P = \dfrac{x_P - x_i}{x_{i+1} - x_i} \\ \bar{y}_P = \dfrac{y_P - y_j}{y_{j+1} - y_j} \end{array}\right\} \in \langle 0;1 \rangle$ are the new coordinates of point P.

Formula (3) contains 16 unknown coefficients $a_{i,j}$, which are determined according to the method described in the doctoral dissertation [10].

From analyses made within the scope of the cited doctoral dissertation, it results that smooth interpolation surfaces are obtained when the continuous model is used, and they have an advantageous influence on efficiency of the calculation process performed in the scope of the analysis of the vehicle dynamics. However, the continuous model does not allow imitating unevenness of the road surface, which fragments are flat in some places (e.g. a vehicle drives through a speed bump on one side, and on the other it drives over the flat surface). In such a case, the discrete model should be taken into account.

4.2 The discrete model of the road surface

In the discrete model of the road surface developed by the author of the cited doctoral dissertation, it was assumed that this surface is modelled by the surface area built out of triangles (Figure 6).

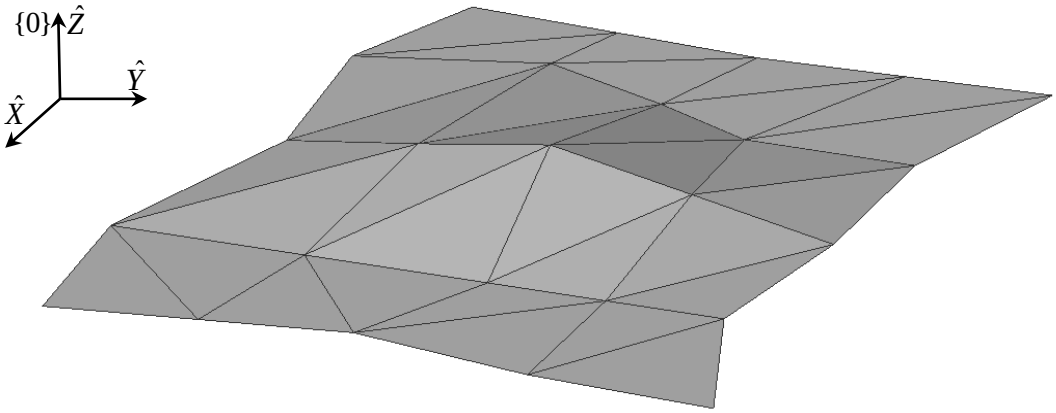


Figure 6 Example of the road surface fragment modelled by triangles

A number of triangles and their sizes are selected to represent the real shape of the road surface as accurately as possible.

The surface area made of the triangles can be defined on the basis of two sets:

1. $S^{(P)} = \{P_1(x_1, y_1, z_1), ..., P_i(x_i, y_i, z_i), ..., P_n(x_n, y_n, z_n)\}$ – a set of all points from which the triangles are built, where $P_i(x_i, y_i, z_i)$ $\mathrm{P_i(x_i, y_i, z_i)}$;
2. $S^{(v)} = \{\mathbf{v}^{(1)}, ..., \mathbf{v}^{(i)}, ..., \mathbf{v}^{(m)}\}$ – a set of 3-element vectors defining all the triangles, that is $\mathbf{v}^{(i)} = \left[v_1^{(i)} \quad v_2^{(i)} \quad v_3^{(i)}\right]^T$, where $v_1^{(i)}, v_2^{(i)}, v_3^{(i)} \in \{1, ..., n\}$ are indexes of the points (from the $\mathrm{S^{(P)}}$ set) being vertexes i-th of this triangle. It means mathematically that for each element of the $S^{(v)}$ set a mapping function, which indicates that the $\mathbf{v}^{(i)}$ element from the $S^{(v)}$ set describes a triangle of vertexes $P_{v_1^{(i)}}, P_{v_2^{(i)}}, P_{v_3^{(i)}}$, was determined in a form of: $\mathbf{v}^{(i)} \mapsto \{P_{v_1^{(i)}}, P_{v_2^{(i)}}, P_{v_3^{(i)}}\}$ $\mathrm{v^{(i)} \mapsto \{P_{v_1^{(i)}}, P_{v_2^{(i)}}, P_{v_3^{(i)}}\}}$.

It should be emphasised that one element (a point) from the $\mathrm{S^{(P)}}$ set can be a common element (a point) for two or more triangles simultaneously.

As in the continuous model, a position of the $P(x_P, y_P, z_P)$ point being in the surface area in question is searched and its coordinates x_P and y_P are known (Figure 7). The z_P coordinate of point P can be determined in a way described further in this article.

Let point P be in the area of the k triangle of vertexes $\mathrm{P_{v_1^{(k)}}, P_{v_2^{(k)}}, P_{v_3^{(k)}}}$ (Figure 7).

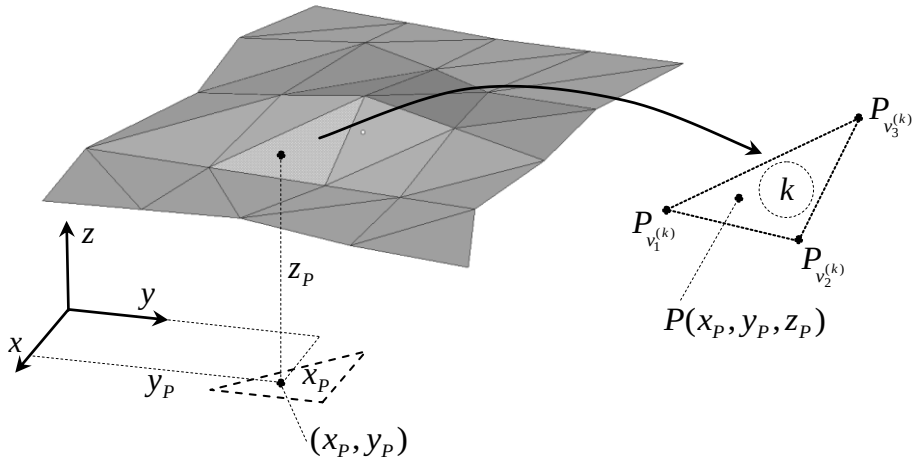


Figure 7 Position of the searched point P on the surface area modelled by the triangles P

The vertexes of this triangle determine the plane of which the normal equation has the following form:

$$e_x^{(k)}x + e_x^{(k)}y + e_x^{(k)}z + \delta^{(k)} = 0, \tag{4}$$

where:

$e_x^{(k)}$, $e_y^{(k)}$, $e_z^{(k)}$ – elements of the $\hat{\mathbf{e}}^{(k)}$ versor normal to the k triangle surface area,

$\delta^{(k)} = -\left(e_x^{(k)}x_l + e_y^{(k)}y_l + e_z^{(k)}z_l\right)$, where l should be taken as any number from set $\left\{v_1^{(k)}, v_2^{(k)}, v_3^{(k)}\right\}$.

Versor $\hat{\mathbf{e}}^{(k)}$ can be determined by the vector product:

$$\hat{\mathbf{e}}^{(k)} = \frac{\mathbf{p}_{1,2}^{(k)} \times \mathbf{p}_{1,3}^{(k)}}{\left|\mathbf{p}_{1,2}^{(k)} \times \mathbf{p}_{1,3}^{(k)}\right|}, \tag{5}$$

where:

$\mathrm{p}_{1,2}^{(\mathrm{k})}$ – the vector with the beginning in point $P_{v_1^{(k)}}$ and the end in point $P_{v_2^{(k)}}$,

$\mathbf{p}_{1,3}^{(k)}$ – the vector with the beginning in point $\mathrm{P}_{\mathrm{v}_1^{(\mathrm{k})}}$ and the end in point $P_{v_3^{(k)}}$.

Having the plane equation determined in (4), the searched $\mathrm{z_P}$ coordinate of the point P can be determined from the formula:

$$z_P = -\frac{e_x^{(k)}x_P + e_y^{(k)}y_P + \delta^{(k)}}{e_z^{(k)}}, \tag{6}$$

for $e_z^{(k)} \neq 0$, excluding the situation when the *k* triangle plane is perpendicular to the *xy* plane – those cases do not concern this work.

In the described procedure, it was assumed that the vertexes of the triangle, on which there is point *P*, are known. However, identification of this triangle is not a trivial task. It becomes especially difficult in a case of computer simulations where short time of calculations is usually significant. Therefore, it is essential to develop an appropriate algorithm of the triangle identification of the surface area in question. The trivial solution of the triangle identification problem consists of searching the whole set of triangles $S^{(v)}$ and checking if the searched element is in the surface area of this triangle. In this case, for each triangle ($k = 1,...,m$) the plane equation (4) should be determined, and it should be checked if the *P* point is in its fragment specified by vertexes $\mathrm{P}_{\mathrm{v}_1^{(\mathrm{k})}}, \mathrm{P}_{\mathrm{v}_2^{(\mathrm{k})}}, \mathrm{P}_{\mathrm{v}_3^{(\mathrm{k})}}$. Such an algorithm does not belong to efficient regarding calculating, and because of three dimensionality it may prove to be problematic. Much better results can be obtained by reducing the problem to a two dimensional issue and narrowing appropriately the set of the searched triangles. In this work, the developed algorithm was divided into two stages: "reducing the problem to a two dimensional issue" and "limiting the search set". These stages are detailed explained in doctoral dissertation [10].

4.3 Normal versor to modeled road surface

Each of the presented road surface models can be characterized by an equation of the assumed mapping surface in the form of:

$$z = z(x, y). \tag{7}$$

In further considerations it is assumed that in any point *P* of this surface of coordinates x_P, y_P, z_P determined in any immovable coordinate system {0} assumed, being a reference system (Figure 8), normal versor $\hat{\mathbf{e}}$ to this surface is known.

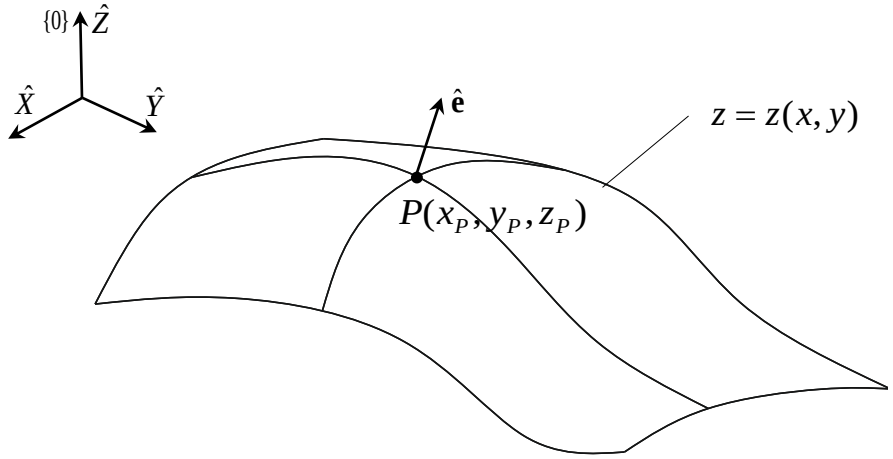


Figure 8. Normal versor $\hat{\mathbf{e}}$ to the mapping surface in point $P(x_P, y_P, z_P)$

According to the suggestions presented in work [3], in the case of the continuous model of the road surface, in the neighborhood of point P (Figure 9), it is assumed that there are four auxiliary points of the coordinates determined in reference system {0} as:

$$\begin{aligned} &P^{(x+)}\big(x_P + \Delta, y_P, z(x_P + \Delta, y_P)\big), \\ &P^{(x+)}\big(x_P - \Delta, y_P, z(x_P - \Delta, y_P)\big), \\ &P^{(y+)}\big(x_P, y_P + \Delta, z(x_P, y_P + \Delta)\big), \\ &P^{(y-)}\big(x_P, y_P - \Delta, z(x_P, y_P - \Delta)\big). \end{aligned} \tag{8}$$

Then, the normal versor can be determined according following formula:

$$\hat{\mathbf{e}} = \frac{\mathbf{r}^{(x)} \times \mathbf{r}^{(y)}}{\left| \mathbf{r}^{(x)} \times \mathbf{r}^{(y)} \right|}, \tag{9}$$

where: $\mathbf{r}^{(x)}$ – vector with the origin in point $P^{(x-)}$ and the end in point $P^{(x+)}$,
$\mathbf{r}^{(y)}$ – vector with the origin in point $P^{(y-)}$ and the end in point $P^{(y+)}$,
Δ [m] – short distance (in the work it was assumed $\Delta = 0{,}01$ m).

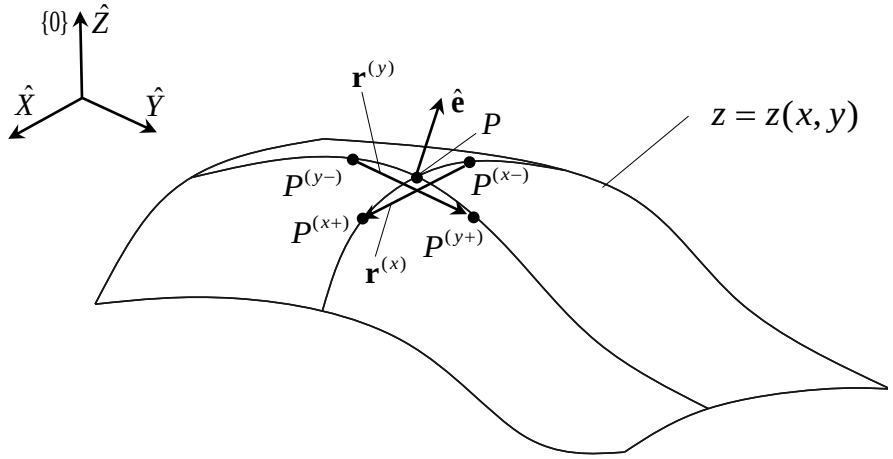


Figure 9. The continuous model of the road surface – determining normal versor $\hat{\mathbf{e}}$ to the mapping surface in point $P(x_P, y_P, z_P)$

In the case of the discrete model of the road surface, based on the triangles or rectangles implemented, the normal versor $\hat{\mathbf{e}}$ can be determined on basis of the known basic geometrical relationships.

5 ALGORITHMS OF ITERATIVE DETERMINING THE CONTACT POINT POSITION

While analyzing vehicle dynamics forces and reaction torques, acting on models of their tires from the road surface, must be considered appropriately. In real conditions the tire contact with this frequently uneven surface takes place within the definite area. When the tire is modeled, the contact surface is usually limited to a point [3], [4], [9]. The authors of this work followed also this procedure, assuming that the modeled tire – considered in a form of a deformable rim – will contact with the mapping surface of the road surface in a definite point

As it is known in any point of the mapping surface a plane tangent to it can be placed. In this method, as it was done by the authors of articles [3], [4], [9], [11], it has been assumed that the tire will be modeled in a form of a deformable rim, obtained as a result of a longitudinal cut of this tire in its symmetry plane. This rim in the deformable part adheres to the mapping surface – even so for needs of the model it is assumed that its contact with this surface takes place in the definite point (it is the C contact point). In a non-deformable form this rim is a circle of the O symmetry center, overlapping with the symmetry center of the non-deformed tire. In the contact point C there is also plane Π – tangent to the mapping surface (Figure 10). On basis of the suggestions included in work [11], in addition to the reference system $\{0\}$ mentioned already, two local coordinate systems – $\{w\}$ and $\{r\}$ were assumed. The movable system $\{w\}$ is connected with the rim. Its origin was placed in the O symmetry center of the non-deformed rim, versor $\hat{Y}_w$ overlaps with its axis of rotation (and therefore, also with the axis of rotation of the modeled tire), and versor $\hat{X}_w$ remains parallel to plane Π during the whole time of vehicle motion. The origin of immovable system $\{r\}$ overlaps with the contact point C, its versor $\hat{Z}_r$ is normal to plane Π – and also to the mapping surface (angle γ between it and versor $\hat{Z}_w$ is an inclination angle of the tire), whereas versor $\hat{X}_r$ lying in this plane remains pararlel to versor $\hat{X}_w$ of the $\{w\}$ system during the vehicle motion. While modeling an interaction of the road surface on the tire, it is assumed that in the contact point C the following forces and reaction torques are applied: $\mathbf{F}_x$ – the reaction force longitudinal, $\mathbf{F}_y$ – the reaction force lateral, $\mathbf{F}_z$ – the reaction force normal to the mapping surface (plane Π), $\mathbf{M}_x$ – the overturning torque, $\mathbf{M}_y$ – the rolling resistance torque, $\mathbf{M}_z$ – the aligning torque. Their directions are consistent with the directions of versors of the $\{r\}$ system. Values of these forces and torques are calculated by use of formulas offered by so called the Pacejka tire model [6], [7], [8] taken into account in the method proposed.

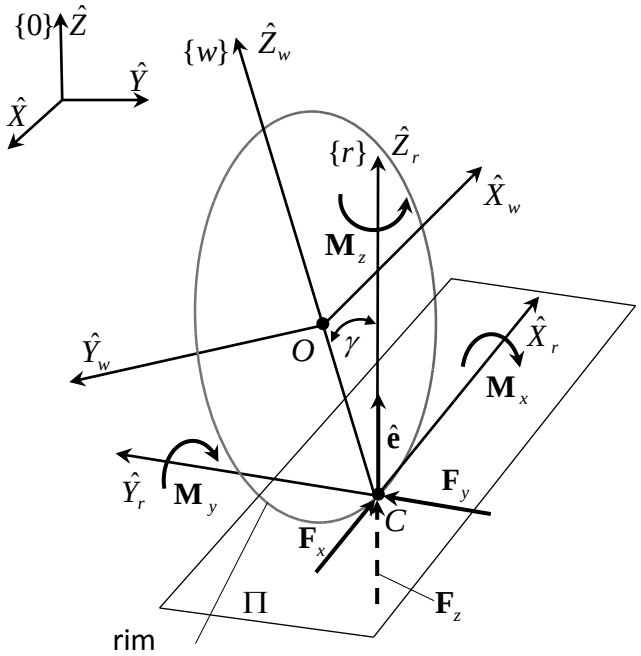


Figure 10. Location of local systems $\{w\}$ i $\{r\}$

While performing an analysis of the vehicle dynamics, it is assumed that a position of the $\{w\}$ system origin is known at any time of its motion (as known, identical with the O symmetry center of the non-deformed rim) and orientation of its versor $\hat{Y}_w$ in the $\{0\}$ reference system. The authors of the article also made the same assumption [4]. Additionally, the position of the contact point C, being the beginning of the $\{r\}$ system in this system and orientation of versors of this system in the $\{0\}$ reference system must be known. Iterative determination of this position and orientation is a subject of the algorithms presented in this work. When a distance of origins of systems $\{w\}$ and $\{r\}$ is known, values of forces and reaction torques acting on the tire from the road surface can be determined by use of the Pacejka tire model. Knowledge about orientation of versors of the $\{r\}$ system in the $\{0\}$ reference system will allow to find directions of acting of these forces and torques – this information is needed to make an analysis of the dynamics of the vehicle in question while using the Pacejka tire model mentioned or other tire models, relying on the similar assumptions regarding a way of applying forces and torques.

Two algorithms, intended for determining the position of the contact point C and orientation of the vectors of the $\{r\}$ coordinate system were proposed. In accordance with intention of the authors the names of the algorithms are to refer to the essence of a procedure assumed in each case. The Plane algorithm is designed for the continuous model of the road surface, whereas the 4Points algorithm –for the discrete model of this surface.

6 ALGORITHM PLANE

Determining the position of the contact point C by the Plane algorithm refers to performing a definite number of iterations. Execution of the first of them is presented in Figure 11.

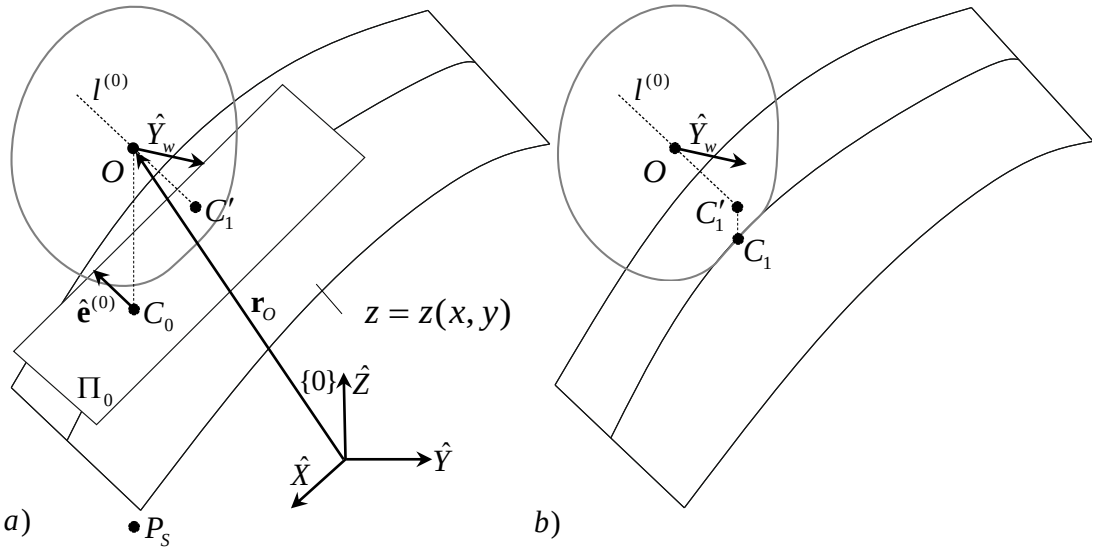


Figure 11. The Plane algorithm – the first approximation of the position of the contact point C

In the neighborhood of the O symmetry center of the non-deformed rim (Figure 11(a)) a start point P_S of coordinates x_S, y_S, z_S, defined in the {0} reference system is selected. In this work it has been assumed that it is the O point. Then, coordinates of the C_0 point, being an orthogonal projection to the mapping surface, are determined in this system:

$$C_0 = \left(x_{C_0}, y_{C_0}, z_{C_0}\right) = \left(x_O, y_O, z(x_O, y_O)\right). \tag{10}$$

A next step is to determine plane Π_0 tangent to the mapping surface in the C_0 point. A point-normal equation of this plane can be presented in the following form (similar to eq. 4):

$$e_x^{(0)} x + \quad e_y^{(0)} y + \quad e_z^{(0)} z + \delta^{(0)} = 0, \tag{11}$$

where: $e_x^{(0)}$, $e_y^{(0)}$, $e_z^{(0)}$ are components of the $\hat{\mathbf{e}}^{(0)}$ versor normal to the mapping surface in the C_0 point determined in the {0} reference system,

$$\delta^{(0)} = -\left(\quad e_x^{(0)} x_{C_0} + \quad e_y^{(0)} y_{C_0} + \quad e_z^{(0)} z_{C_0} \right).$$

Point $C_1'\left(x_{C_1'}, y_{C_1'}, z_{C_1'}\right)$, in which straight line $l^{(0)}$ going through the O points pierces plane Π_0 perpendicularly, is determined next . Its coordinates in the {0} reference system can be determined by the position vector (Figure 11(b)) as:

$$\mathbf{r}_{C_1'} = \quad \mathbf{r}_O - d_0 \quad \hat{\mathbf{e}}^{(0)} \quad , \tag{12}$$

where: $d_0 = \left| e_x^{(0)} x_O + e_y^{(0)} y_O + e_z^{(0)} z_O + \delta^{(0)} \right|$ – a distance of points O and C_1'.

As a result, these coordinates can be presented as:

$$C_1'\left(x_{C_1'}, y_{C_1'}, z_{C_1'}\right) = C_1'\left(x_O - e_x^{(0)} d_0, y_O - e_y^{(0)} d_0, z_O - e_z^{(0)} d_0\right). \tag{13}$$

Then, the coordinates of point C_1, being the first approximation of the contact point C, are determined:

$$C_1\left(x_{C_1}, y_{C_1}, z_{C_1}\right) = C_1\left(x_{C_1'}, y_{C_1'}, z\left(x_{C_1'}, y_{C_1'}\right)\right). \tag{14}$$

To determine n-th approximation of the position of the contact point C the algorithm can be generalized to $i = 1, \ldots, n$ iterations writing formulas (13) and (14) as:

$$C_i'\left(x_{C_i'}, y_{C_i'}, z_{C_i'}\right) = C_1'\left(x_O - e_x^{(i-1)} d_{i-1}, y_O - e_y^{(i-1)} d_{i-1}, z_O - e_z^{(i-1)} d_{i-1}\right) \quad (15)$$

and:

$$C_i\left(x_{C_i}, y_{C_i}, z_{C_i}\right) = C_i\left(x_{C_i'}, y_{C_i'}, z\left(x_{C_i'}, y_{C_i'}\right)\right). \quad (16)$$

The n value is determined by the criterion:

$$\sqrt{\left(x_{C_{n-1}} - x_{C_n'}\right)^2 + \left(y_{C_{n-1}} - y_{C_n'}\right)^2 + \left(z_{C_{n-1}} - z_{C_n'}\right)^2} \leq \varepsilon, \quad (17)$$

where: ε – an assumed acceptable absolute error of calculations.

The versor $\hat{X}_r^{(n)}$ of the $\{r\}$ system in the reference system $\{0\}$ can be determined by the formula (Figure 10):

$$\hat{X}_r^{(n)} = \frac{\hat{Y}_w \times \hat{\mathbf{e}}^{(n)}}{|\hat{Y}_w \times \hat{\mathbf{e}}^{(n)}|}, \quad (18)$$

and then other versors as:

$$\begin{aligned} \hat{Y}_r^{(n)} &= \hat{X}_r \times \hat{\mathbf{e}}^{(n)}, \\ \mathrm{Z}_r^{(n)} &= \hat{\mathbf{e}}^{(n)}. \end{aligned} \quad (19)$$

7 ALGORITHM 4POINTS

As it has been found out the Plane algorithm is used in the case of the continuous model of the road surface. However, in the case of the discrete model when some fragments of the surface are flat, determining the position of the contact point C by its use may not be accurate enough. Such a situation is presented in Figure 12.

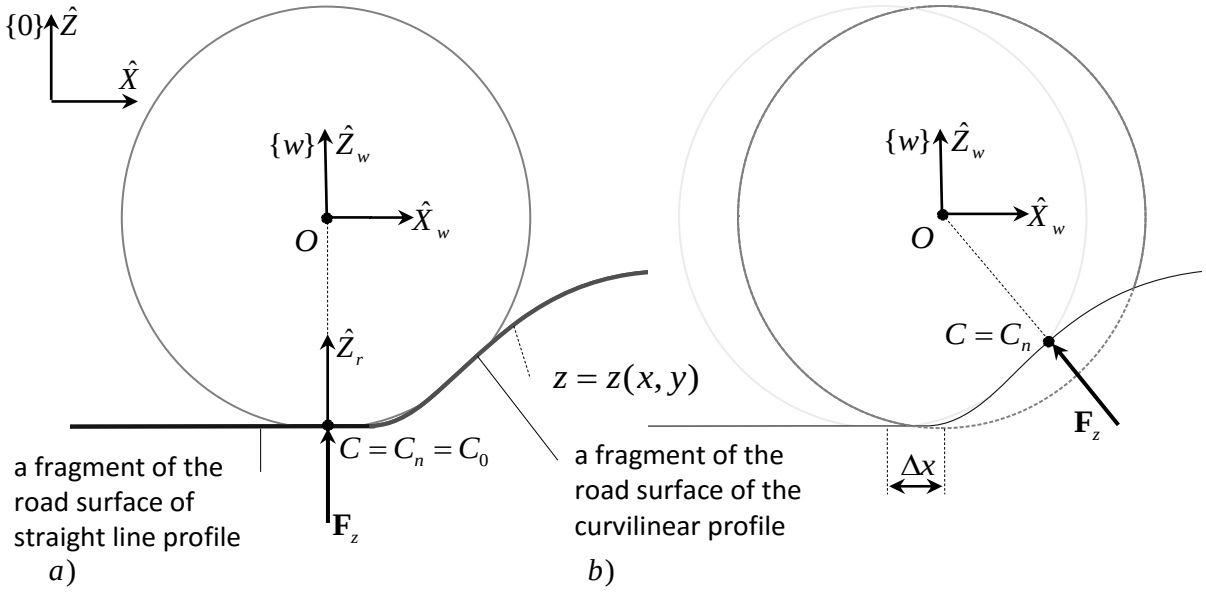


Figure 12. The Plane algorithm – the determined positions of the contact point C: a) at beginning of running the modeled tire over unevenness, b) after covering distance Δx

While considering the position of the modeled tire presented in Figure 12(a), it can be noticed that the contact point C is the orthogonal projection of the symmetry center O of the non-deformed rim on the flat fragment of the road surface. It can be stated that condition (17),

determining completion of calculations, is met after the first iteration, because already then it is $\sqrt{\left(x_{C_{n-1}} - x_{C_n'}\right)^2 + \left(y_{C_{n-1}} - y_{C_n'}\right)^2 + \left(z_{C_{n-1}} - z_{C_n'}\right)^2} = 0 < \varepsilon$. Here, an undesirable effect is too late reaction of the rim to changeable surface profile. Since the Plane algorithm does not allow to take the rim contact with the road fragment of the curvilinear profile (marked in the figure) into account early enough, so a direction of the normal reaction force $\mathbf{F}_z$ (acting on the rim in accordance with the $\hat{Z}_r$ versor direction) turns out to be incorrect. It can be stated that this direction "does not keep up with a new situation on the road". The normal reaction force changes its direction after the modeled tire has covered the Δx distance (Figure 12(b)) – so too late – and this change is rather rapid. Therefore, it has been required to develop an algorithm, which would enable to determine an appropriate position of the contact point C, ensuring that shape changes of the mapping surface are considered early enough, what would provide more accurate direction of the $\mathbf{F}_z$ normal reaction force, because closer to real one.

At the beginning of realization of the new algorithm named 4Points in planes $\hat{X}_w, \hat{Z}_w$ and $\hat{Y}_w, \hat{Z}_w$ of the $\{w\}$ system there are four auxiliary points $O^{(x+)}, O^{(x-)}, O^{(y+)}, O^{(y-)}$ assumed, respectively (Figure 13).

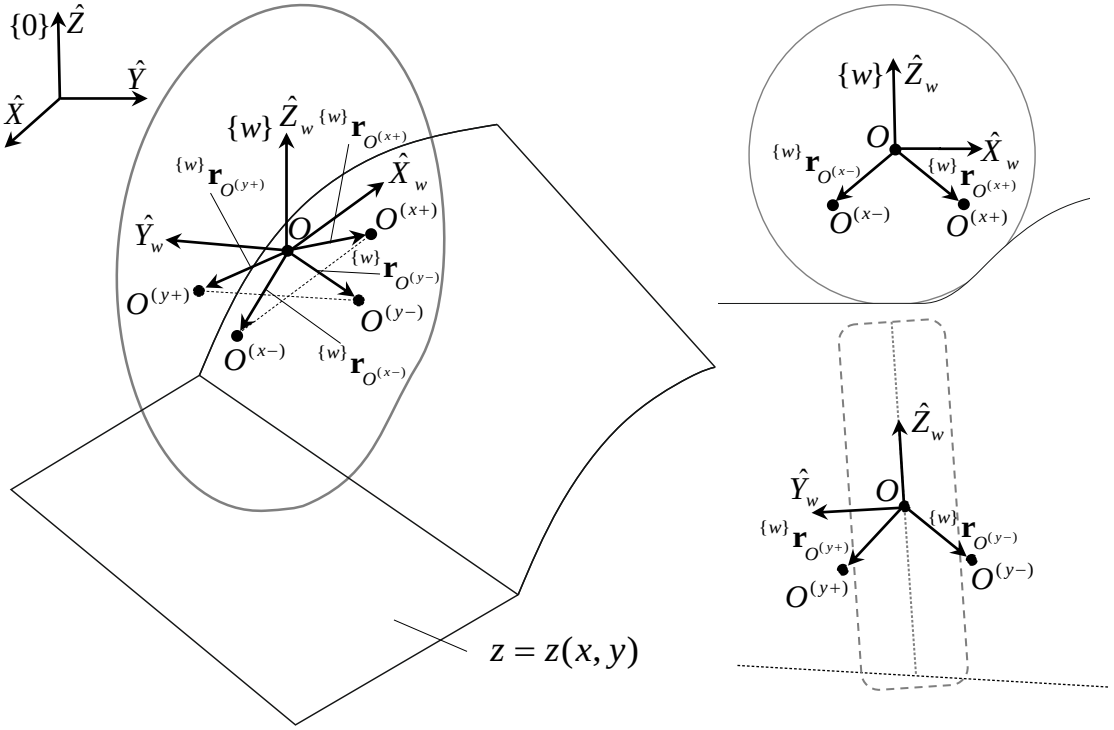


Figure 13. Algorithm 4Points – positions of the auxiliary points

Coordinates of them in the $\{w\}$ system can be determined by the position vectors of the homogenous coordinates:

$$
{}^{\{w\}}\mathbf{r}^{*}_{O^{(x+)}} = \begin{bmatrix} {}^{\{w\}}\mathbf{r}_{O^{(x+)}} \\ 1 \end{bmatrix}, \qquad {}^{\{w\}}\mathbf{r}_{O^{(x+)}} = \begin{bmatrix} \Delta x \\ 0 \\ -\Delta z \end{bmatrix},
$$
$$
{}^{\{w\}}\mathbf{r}^{*}_{O^{(x-)}} = \begin{bmatrix} {}^{\{w\}}\mathbf{r}_{O^{(x-)}} \\ 1 \end{bmatrix}, \qquad {}^{\{w\}}\mathbf{r}_{O^{(x-)}} = \begin{bmatrix} -\Delta x \\ 0 \\ -\Delta z \end{bmatrix}, \tag{20}
$$

$$
{}^{\{w\}}\mathbf{r}^{*}_{O^{(y+)}} = \begin{bmatrix} {}^{\{w\}}\mathbf{r}_{O^{(y+)}} \\ 1 \end{bmatrix}, \qquad {}^{\{w\}}\mathbf{r}_{O^{(y+)}} = \begin{bmatrix} 0 \\ \Delta y \\ -\Delta z \end{bmatrix},
$$

$$
{}^{\{w\}}\mathbf{r}^{*}_{O^{(y-)}} = \begin{bmatrix} {}^{\{w\}}\mathbf{r}_{O^{(y-)}} \\ 1 \end{bmatrix}, \qquad {}^{\{w\}}\mathbf{r}_{O^{(y-)}} = \begin{bmatrix} 0 \\ -\Delta y \\ -\Delta z \end{bmatrix},
$$

where: $\Delta x, \Delta y, \Delta \mathrm{z}$ [m] – distances resulting from tire size (in the work the following values were assumed: $\Delta x = 0{,}17, \Delta y = 0{,}07, \Delta \mathrm{z} = 0{,}1$).

Then, the vectors of the homogenous coordinates determining the position of the auxiliary points in the {0} reference system, can be determined as:

$$
\begin{aligned}
&\mathbf{r}^{*}_{O^{(x+)}} = \mathbf{T}_w \, {}^{\{w\}}\mathbf{r}^{*}_{O^{(x+)}}, \mathbf{r}^{*}_{O^{(x+)}} = \begin{bmatrix} \mathbf{r}_{O^{(x+)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(x+)}} = \begin{bmatrix} x_{O^{(x+)}} \\ y_{O^{(x+)}} \\ z_{O^{(x+)}} \end{bmatrix}, \\
&\mathbf{r}^{*}_{O^{(x-)}} = \mathbf{T}_w \, {}^{\{w\}}\mathbf{r}^{*}_{O^{(x-)}}, \mathbf{r}^{*}_{O^{(x-)}} = \begin{bmatrix} \mathbf{r}_{O^{(x+)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(x-)}} = \begin{bmatrix} x_{O^{(x-)}} \\ y_{O^{(x-)}} \\ z_{O^{(x-)}} \end{bmatrix}, \\
&\mathbf{r}^{*}_{O^{(y+)}} = \mathbf{T}_w \, {}^{\{w\}}\mathbf{r}^{*}_{O^{(y+)}}, \mathbf{r}^{*}_{O^{(y+)}} = \begin{bmatrix} \mathbf{r}_{O^{(y+)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(x+)}} = \begin{bmatrix} x_{O^{(y+)}} \\ y_{O^{(y+)}} \\ z_{O^{(y+)}} \end{bmatrix}, \\
&\mathbf{r}^{*}_{O^{(y-)}} = \mathbf{T}_w \, {}^{\{w\}}\mathbf{r}^{*}_{O^{(y-)}}, \mathbf{r}^{*}_{O^{(y-)}} = \begin{bmatrix} \mathbf{r}_{O^{(y-)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(y-)}} = \begin{bmatrix} x_{O^{(y-)}} \\ y_{O^{(y-)}} \\ z_{O^{(y-)}} \end{bmatrix},
\end{aligned}
\tag{21}
$$

where: $\mathbf{T}_w$ – a known transformation matrix from system $\{w\}$ to reference system {0}.

In further procedure auxiliary points $O^{(x+)}, O^{(x-)}, O^{(y+)}, O^{(y-)}$ are projected on the mapping surface (Figure 14(a)), and next the homogenous vectors, determining coordinates of their projections $O'^{(x+)}, O'^{(x-)}, O'^{(y+)}, O'^{(y-)}$ in the {0} reference system are determined as:

$$
\begin{aligned}
&\mathbf{r}^{*}_{O'^{(x+)}} == \begin{bmatrix} \mathbf{r}_{O'^{(x+)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(x+)}} = \begin{bmatrix} x_{O'^{(x+)}} \\ y_{O'^{(x+)}} \\ z_{O'^{(x+)}} \end{bmatrix} = \begin{bmatrix} x_{O^{(x+)}} \\ y_{O^{(x+)}} \\ z\left(x_{O^{(x+)}}, y_{O^{(x+)}}\right) \end{bmatrix}, \\
&\mathbf{r}^{*}_{O'^{(x-)}} == \begin{bmatrix} \mathbf{r}_{O'^{(x-)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(x-)}} = \begin{bmatrix} x_{O'^{(x-)}} \\ y_{O'^{(x-)}} \\ z_{O'^{(x-)}} \end{bmatrix} = \begin{bmatrix} x_{O^{(x-)}} \\ y_{O^{(x-)}} \\ z(x_{O^{(x-)}}, y_{O^{(x-)}}) \end{bmatrix}, \\
&\mathbf{r}^{*}_{O'^{(y+)}} == \begin{bmatrix} \mathbf{r}_{O'^{(y+)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(y+)}} = \begin{bmatrix} x_{O'^{(y+)}} \\ y_{O'^{(y+)}} \\ z_{O'^{(y+)}} \end{bmatrix} = \begin{bmatrix} x_{O^{(y+)}} \\ y_{O^{(y+)}} \\ z(x_{O^{(y+)}}, y_{O^{(y+)}}) \end{bmatrix}, \\
&\mathbf{r}^{*}_{O'^{(y-)}} == \begin{bmatrix} \mathbf{r}_{O'^{(y-)}} \\ 1 \end{bmatrix}, \qquad \mathbf{r}_{O^{(y-)}} = \begin{bmatrix} x_{O'^{(y-)}} \\ y_{O'^{(y-)}} \\ z_{O'^{(y-)}} \end{bmatrix} = \begin{bmatrix} x_{O^{(y-)}} \\ y_{O^{(y-)}} \\ z(x_{O^{(y-)}}, y_{O^{(y-)}}) \end{bmatrix},
\end{aligned}
\tag{22}
$$

On basis of formula (9) the versor normal to plane Π, including the projections of the auxiliary points can be determined as:

$$
\hat{\mathbf{e}} = \frac{\mathbf{r}^{(x)}_{O'} \times \mathbf{r}^{(y)}_{O'}}{\left| \mathbf{r}^{(x)}_{O'} \times \mathbf{r}^{(y)}_{O'} \right|}, \tag{23}
$$

where: $\mathbf{r}^{(x)}_{O'} = \mathbf{r}_{O'^{(x+)}} - \mathbf{r}_{O'^{(x-)}}$ – vector of origin in point $O'^{(x+)}$ and end in point $O'^{(x-)}$,

$\mathbf{r}^{(y)}_{O'} = \mathbf{r}_{O'^{(y+)}} - \mathbf{r}_{O'^{(y-)}}$ – vector of origin in point $O'^{(y+)}$ and end in point $O'^{(y-)}$.

On basis of the determined normal versor **ê** and one of any selected auxiliary points $O'^{(x+)}, O'^{(x-)}, O'^{(y+)}, O'^{(y-)}$ the Π plane mentioned is sought for, of which the point-normal equation has a from:

$$e_x x + e_y y + e_z z + \delta = 0, \tag{24}$$

where: e_x, e_y, e_z are the components of the versor **ê** normal to plane Π determined in the {0} reference system,

$\delta = -\left(e_x x_{O'^{(x+)}} + e_y y_{O'^{(x+)}} + e_z z_{O'^{(x+)}}\right)$ if the selected point is $O'^{(x+)}$.

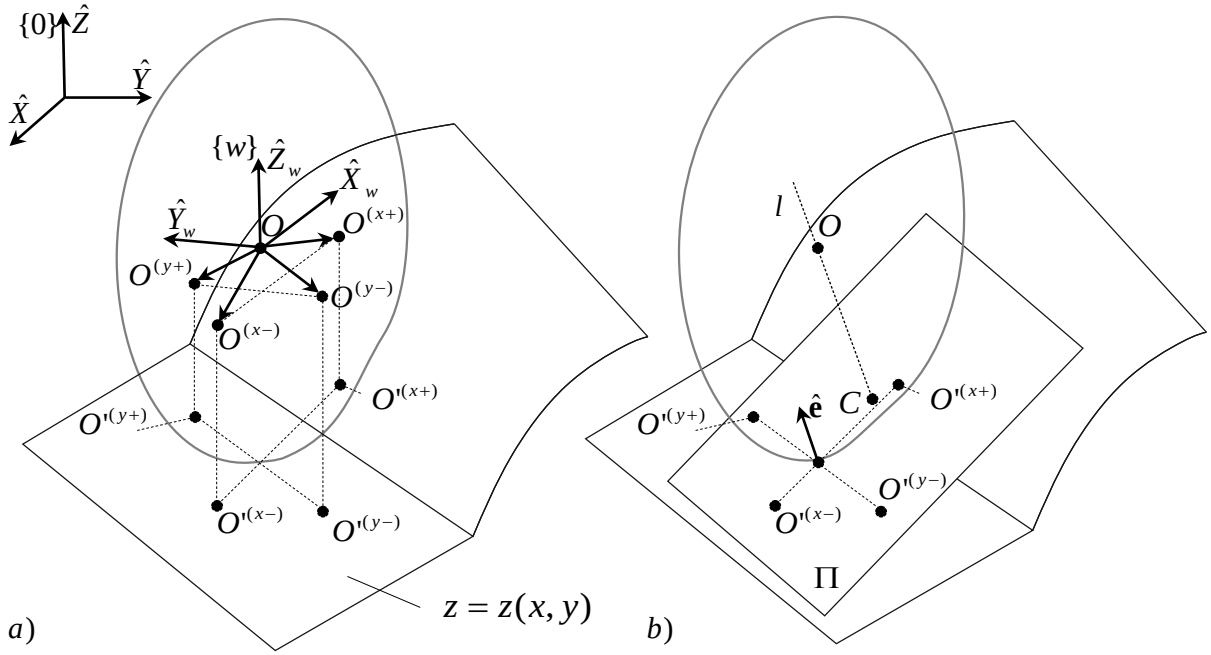


Figure 14. Algorithm 4Points – determining the orthogonal projections of the auxiliary points on the mapping surface

The sought contact point C is determined as a point in which straight line l going through the point O pierces the plane Π perpendicularly (Figure 14(b)). Its coordinates in the {0} reference system are determined as the components of the position vector:

$$\mathbf{r}_C = \mathbf{r}_O - d_0 \hat{\mathbf{e}}, \tag{25}$$

where: $d_0 = \left| e_x\, x_O + e_y\, y_O + e_z\, z_O + \delta \right|$ – a distance of points O and C.
As a result the coordinates of the contact point C can be determined as:

$$C(x_C, y_C, z_C) = C\left(x_O - e_x\, d_0, y_O - e_y\, d_0, z_O - e_z\, d_0\right). \tag{26}$$

Using this algorithm for the case illustrated in Figure 12(a), the position of the contact point C can be determined as presented in Figure 15. The corrected direction of reaction force $\mathbf{F}_z$ is presented in this figure.

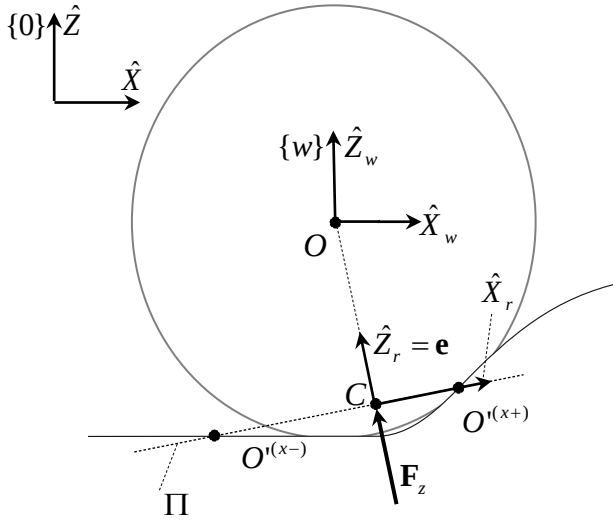


Figure 15. Algorithm 4Points – determining the position of the contact point C and the corrected direction of normal reaction force $\mathbf{F}_z$

The versor directions of the $\{r\}$ system are determined in a similar way as in the case of the Plane algorithm, that is according to relationships (18) and (19).

8 COMPUTER SIMULATIONS

A technical rescue vehicle, which can drive in the terrain conditions, was used for the analysis. Its physical model in a form of multibody system of open structure, built by use of joint coordinates defining a relative position of the modeled components, and a mathematical model corresponding to it, developed on basis of Lagrange equations formalism by use of homogenous transformation matrices [2], is presented in the doctoral dissertation [10]. Program Blender [12] was used to model the uneven road surface and to develop a model of the vehicle used in the computer animations.

In each case considered the modeled vehicle moved in the direction consistent with versor $\hat{X}$ of the {0} reference system. The vehicle initial speed was 5 km/h, and simulation duration time was 6s.

Example I

The assumed continuous model of the road surface in a form of a grid of control points is presented in figure 16. The boundary values of the coordinates of those points in the {0} reference system were following: $x_{min} = -2$; $x_{max} = 49$; $y_{min} = -8$; $y_{max} = 8$; $z_{min} = -0{,}8$; $z_{max} = 1{,}2$.

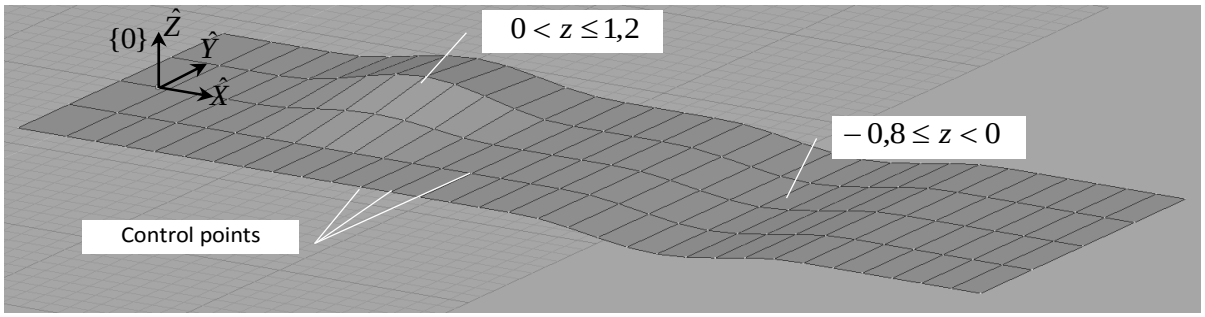
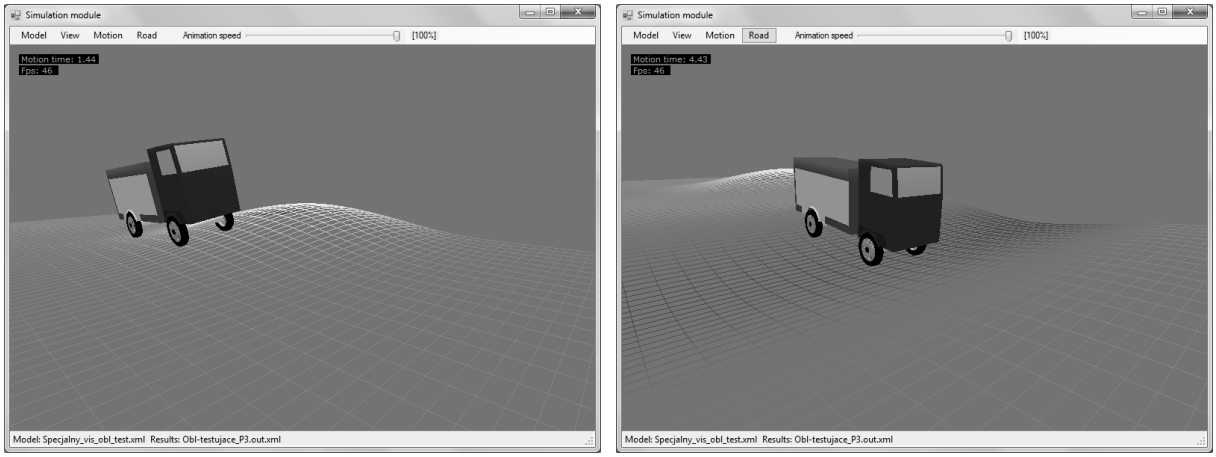


Figure 16. The control point grid in the case of the continuous model of the road surface

Within the computer animation performed vehicle drives over the uneven road surface of the shape as shown in Figure 17 were simulated.



a) *b*)

Figure 17. Some screen shots made during the computer animation:
a) a phase of the vehicle going up to unevenness, b) a phase of going down from unevenness

Some examples of the calculation results which concern determining a vertical course of the gravity center displacement of the vehicle model (towards versor $\hat{Z}$ of the {0} reference system) – taking the Plane algorithm into account – are presented in Figure 18.

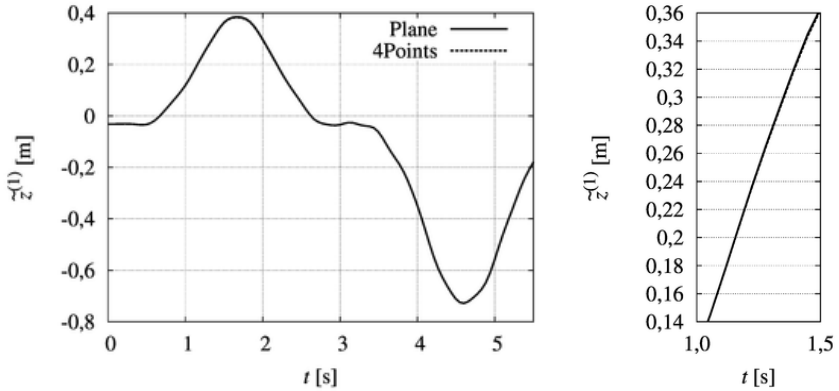


Figure 18. A course of the vertical displacement of the vehicle gravity center in the case of considering the continuous model of the road surface

This diagram was compared with the vertical course of this center, determined by using the 4Points algorithm. The results obtained are almost identical. Therefore, it may be concluded that in the case of smooth unevenness a selection of the algorithm has a slight influence on the computer simulation results obtained.

Example II

The assumed discrete model of the road surface in the form of the grid of the control points is presented in Figure 19. It consists of two flat fragments adjacent to a bump. Since there are no inclination of the surface in the direction consistent with versor $\hat{Y}$ of the {0} reference system, its model was made by use of rectangles placed as shown in the figure. The assumed boundary values of the grid points coordinates were following: $x_{min} = -2$; $x_{max} = 11{,}5$; $y_{min} = -2$; $y_{max} = 2$; $z_{min} = 0$; $z_{max} = 0{,}2$.

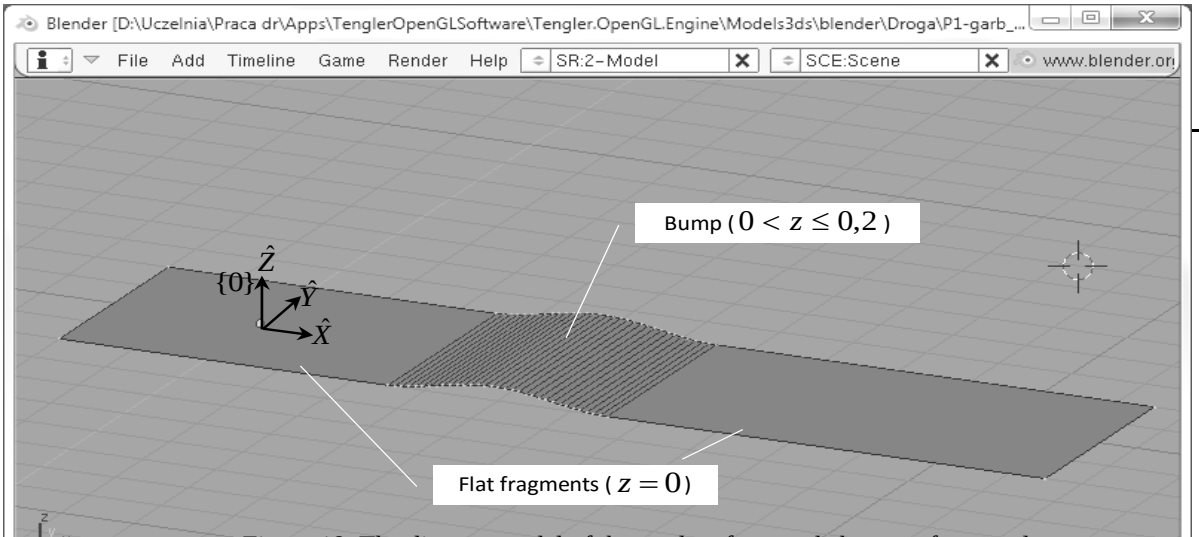
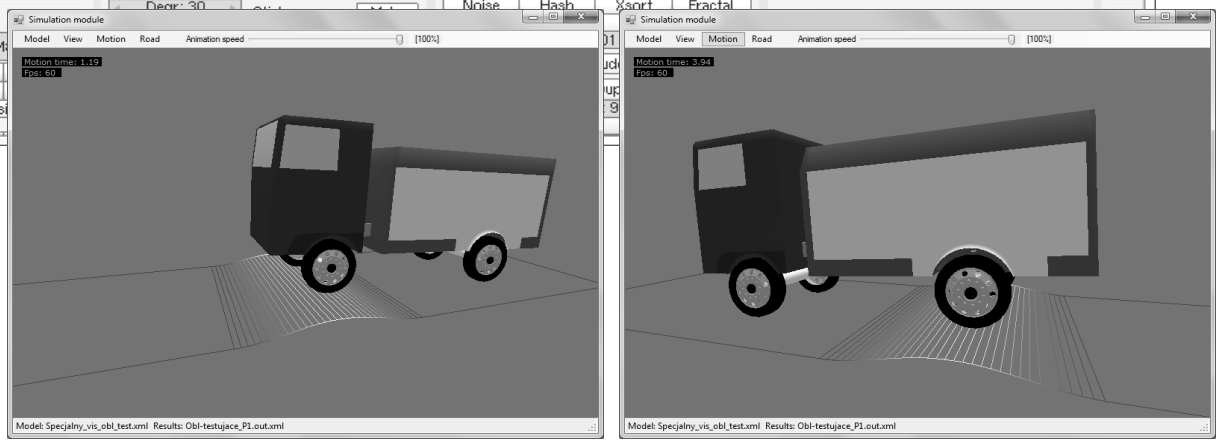


Figure 19. The discrete model of the road surface made by use of rectangles

Within the computer animation performed a drive of the vehicle over the bump was simulated (Figure 20).



a) *b*)

Figure 20. Some screen shots made during the computer animation:
a) a phase of the vehicle going up the bump, b) a phase of the vehicle going down from the bump

Some examples of the calculations results which concern determining the vertical displacement course of the gravity center of the vehicle model – considering algorithm 4Points – are presented in Figure 21(a). Two phases of motion can be differentiated here when first front wheels and then rear wheels of the vehicle drive over the bump.

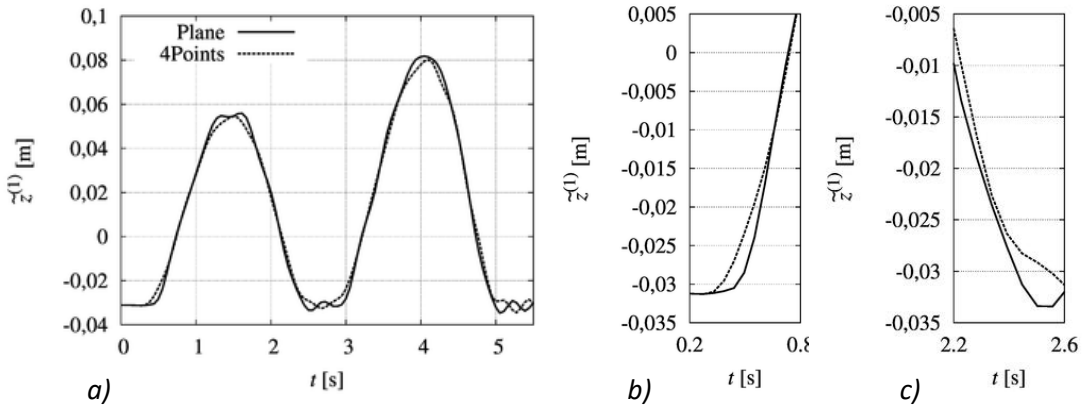


Figure 21. A course of the vertical displacement of the vehicle gravity center when the discrete model of the road surface made by use of rectangles is considered

By analyzing the results in Figure 21(a), it can be noticed that when the 4Points algorithm is used, the displacement of the gravity center of the vehicle while its going up to the bump, takes place earlier than in the case of the Plane algorithm – the dashed line is visible before the solid line (Figure 21(b)). The analogical situation can be observed during the vehicle going down from the bump. In this case the 4Points algorithm is more „sensitive" to the unevenness profile change behind the wheels – the dashed line is visible behind the solid line (Figure 21(c)). Therefore, the thesis is confirmed that in the case of overcoming unevenness, where the road fragments are flat, better results are obtained when the 4Points algorithm is used.

Within realization of the doctoral dissertation [10] a number of test computer simulations, verifying effectiveness of the presented algorithms with reference to the assumed models of the road surface, were made. The computer results were verified positively after comparing them to the experimental test results made on the test track.

9 CONCLUSIONS

The presented algorithms have a general significance and that is why they can be used in future in the case of considering more advanced tire models. In order to sum up this article it should be emphasized that the development of the presented algorithms was only a part of the task undertaken by the authors. These algorithms with the tire models were included into the advanced mathematical model of the selected terrain vehicle, developed with view to performing an analysis of its dynamics. This model with the prepared models of the road surface and the developed computer programs constitute a prototype of a technical rescue vehicle. According to the authors, the observations made during the computer simulations of its motion can aid a process of designing of this type of vehicles in the future. In addition to the statements presented, the authors would like to point out to wide use – in the case of the proposed method – of possibilities of the Blender program, especially while developing the road surface models, and also a vehicle model used in the computer animations.

REFERENCES

[1] Craig J.J., Introduction to Robotics. Mechanics and Control, Addison-Wesley Publishing Company, Inc, 1989, 1986

[2] Grzegożek W., Adamiec-Wójcik I. and Wojciech S. Computer Modelling of dynamics of motor vehicles (in Polish), Publisher of Technical University of Cracow, Cracow 2003.

[3] Hirschberg W., Rill G., Weinfurter H., Tire model TMeasy, Vehicle System Dynamics, Vol.45, 2007, pp.101-119

[4] Hirschberg W., Rill G., Weinfurter H., User-Appropriate Tyre-Modeling for Vehicle Dynamics in Standard and Limit Situation, Vehicle System Dynamics, Vol.38, No.2. 2002

[5] Keys R., Cubic convolution interpolation for digital image processing, Acoustics, Speech and Signal Processing, Vol. 29, pp. 1153-1160, 1981

[6] Pacejka H.B, Bakker E., Lidner L., New Tire Model with an Application in Vehicle Dynamics Studies, SAE Technical Paper 890087, 1989

[7] Pacejka H.B., Bakker E., The Magic Formula Tyre Model, Vehicle System Dynamics, Vol.21, 1993, pp. 1-18

[8] Pacejka H.B., Spin: camber and turning, Vehicle System Dynamics, Vol.43, 2005, pp.3-17

[9] Rill G., TMeasy – The Handling Tire Model for all Driving Situations, in: Savi M.A. (ed.), Proc. of the XV International Symposium on Dynamic Problems of Mechanics DINAME 2013, Buzios, Brazil, 2013

[10] Tengler S., Analysis of dynamics of special vehicles with high gravity center (in Polish), PhD Thesis, Faculty of Mechanical Engineerim and Computer Science, University of Bielsko-Biala, Poland, 2012
[11] Unrau H.-J., Zamow J., TYDEX-Format. Description and Reference Manual. Release 1.3, 1997
[12] www.blender.org

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

STRUCTURED SPATIAL DISCRETIZATION OF DYNAMICAL SYSTEMS

Marko Jokic[1], **Andrej Jokic**[1], **and Bruno Dogancic**[1]

[1]University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture
10000 Zagreb, Ivana Lucica 5, Croatia
e-mail: mjokic@fsb.hr, ajokic@fsb.hr, bruno.dogancic@stud.fsb.hr

Keywords: Spatial Discretization, Structured Dynamical System, Uncertainty Modeling.

Abstract. *This paper deals with spatial discretization of dynamical systems, such that the resulting discretized model is structured as a set of interconnected subsystems. We start by discretizing a dynamical system using an appropriate method (e.q. the finite element method) and discuss a general framework for partitioning a single discretized system into two interconnected systems. As it turns out, partitioning the discretized system into a set of interconnected systems appears to be a straightforward extension of the proposed procedure — we simply recursively apply the procedure. When doing so, we derive a general condition for preserving the interconnection structure, i.e. a necessary condition on the discretized system matrices such that the interconnection graph between the resulting subsystems matches the interconnection graph between the respective physical subdomains of the original system. This framework allows us to address a local discretization error (i.e. an error in a subsystem) influence on the overall system response. Furthermore, by considering a series of successively finer discretizations of an individual subsystem, we are able to construct a series of successively finer discretized system models with respective uncertainties.*

1 INTRODUCTION

Spatial discretization of partial differential equations [1] describing a dynamical system has been one of the cornerstones of structural dynamics, vibration systems modeling, wave propagation studies, and so on for several decades. As the result, a plethora of techniques for modeling and simulation of dynamical systems, such as the finite element method [2, 3], have been extensively researched and established. Virtually all of the analysis techniques based on discretization rely on the assumption that a series of successively finer spatial discretizations of the system converge to the "accurate" mathematical model of the system. In this setting, choosing a spatial discretization which is accurate enough for a problem at hand usually boils down to a-posteriori analysis, i.e. addressing the error between a chosen discretization and an "accurate" model, usually a very fine discretization of the entire system.

On the other hand, robust control theory of dynamical systems adopts inherently different mathematical modeling paradigm — a dynamical system is approximated by a *nominal system* (e.q. linear time invariant system) and an *uncertainty model* [4]. The uncertainty model is typically a system with dynamic behaviour which confines a difference in behaviour between the nominal and the original system due to nonlinearities, neglected dynamics, discretization errors and so on. Roughly speaking, we are designing a simple model which captures *all* system behaviours, as opposed to structural dynamics approach where a (hopefully accurate) detailed and complex mathematical model is constructed. Structured control schemes, such as distributed and decentralized control [5], introduce additional mathematical modeling requirements. In structured control applications, a mathematical model of a system to be controlled comprises of a number of (possibly uncertain) interconnected dynamical systems, as shown in Figure 1. Thus, the mathematical model exhibits a certain structure, usually represented by a graph — graph vertices are subsystems and graph edges represent subsystems connections.

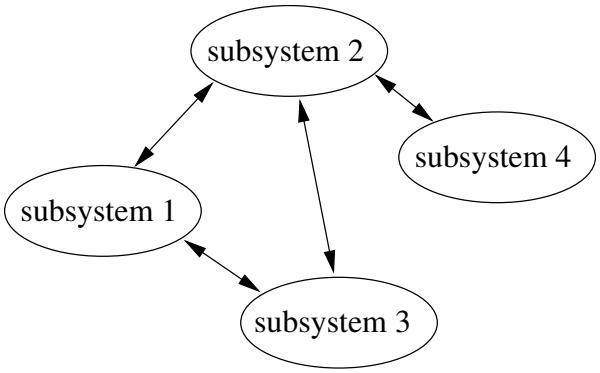


Figure 1: Structured representation of a dynamical system.

This paper addresses the following issue: *How to construct a spatial discretization of a dynamical system, such that its discretized model has a structure corresponding to a given graph?* We begin by providing a general framework for partitioning a discretized mechanical system into a series of interconnected mechanical systems. Then, we derive a necessary condition which, if met, allows us to use the partitioning procedure for constructing structured discretized model of a system.

2 PARTITIONING A SPATIALLY DISCRETIZED SYSTEM INTO A SET OF SUBSYSTEMS

Without loss of generality, we discuss a case of partitioning a single discretized mechanical system into two interconnected mechanical systems. As it turns out, partitioning a system into a series of interconnected systems may be viewed as a straightforward extension of the procedure.

Let us consider a mechanical system spatially discretized by means of the finite element method. In particular, we consider a linear time invariant (LTI) second order system

$$\begin{aligned} M\ddot{q} + C_v\dot{q} + Kq &= B_d d, \\ z &= Cx + Dd, \end{aligned} \tag{1}$$

where M, C_v and K are mass, viscous damping and stiffness matrices, respectively, B, C and D are input, output and feed-through matrices, respectively, and $x = \begin{pmatrix} \dot{q}^T & q^T \end{pmatrix}^T$ are state variables. The matrices B, C and D are input, output and feed-through matrices, and d and z are input and output vectors, respectively.

The system (1) may be rewritten as the equivalent first order descriptor system

$$\begin{aligned} E\dot{x} &= Ax + Bw, \\ z &= Cx + Dw, \end{aligned} \tag{2}$$

with

$$E = \begin{pmatrix} M & 0 \\ 0 & F \end{pmatrix}, \qquad A = \begin{pmatrix} -C_v & -K \\ I & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} B_w \\ 0 \end{pmatrix}, \tag{3}$$

where F may be any nonsingular matrix (usually chosen to be the identity matrix I).

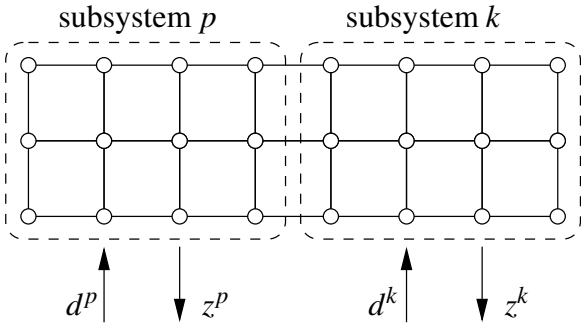


Figure 2: Partitioning the discretized system.

To partition the system (1) into two interconnected systems, we begin by identifying the boundary that separates the degrees of freedom (DOFs) of the two subsystems, which we denote by subsystem p and subsystem k, as shown in Figure 2. Typically, when dealing with spatially discretized system by means of FEM, this is equivalent to identifying the nodes (with respective DOFs) belonging to either subsystem p or subsystem k. More formally, we reorder the state variables $\tilde{x}$ such that $\tilde{x} = \begin{pmatrix} (x^p)^T & (x^k)^T \end{pmatrix}^T$, where x^p and x^k denote the state variables for the subsystem p and subsystem k, respectively. In the similar fashion, we reorder the inputs w and the outputs z such that $\tilde{w} = \begin{pmatrix} (w^p)^T & (w^k)^T \end{pmatrix}^T$ and $\tilde{z} = \begin{pmatrix} (z^p)^T & (z^k)^T \end{pmatrix}^T$, where p and k denote the variables for the subsystem p and subsystem k, respectively.

Furthermore, we assume that a single component d_i of the input vector $d = \begin{pmatrix} d_1 & d_2 & \cdots & d_n \end{pmatrix}^T$ acts either on subsystem p or the subsystem k, but not on the both subsystems. If this is not the case, i.e. if the signal d_i acts on both subsystems, it can be split into two signals $(d^p)_i$ and

$(d^k)_i$ acting on subsystems p and k, respectively. Equivalently, if the output signal z_i originates form both subsystems, it can be split into two output signals $(z^p)_i$ and $(z^k)_i$ originating from subsystems p and k, respectively. We also assume that the feed-through matrix D maps the inputs that act on a single subsystem only to the outputs that correspond to the same subsystem – if this is not the case, the matrix D may be modified in the similar fashion as described above.

The above described procedure may be accomplished by constructing a permutation matrix P^{pk} such that

$$\tilde{x} = \begin{pmatrix} x^p \\ x^k \end{pmatrix} = P^{pk}x, \quad \tilde{d} = \begin{pmatrix} d^p \\ d^k \end{pmatrix} = P^{pk}d, \quad \tilde{z} = \begin{pmatrix} z^p \\ z_k \end{pmatrix} = P^{pk}z. \tag{4}$$

By multiplying the equation (2) by the matrix P^{pk} from the left and taking into the account (4), we obtain

$$\begin{aligned} \tilde{E}\ddot{\tilde{x}} &= \tilde{A}\tilde{x} + \tilde{B}\tilde{d}, \\ \tilde{z} &= \tilde{C}\tilde{x} + \tilde{D}\tilde{d}, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \tilde{E} = P^{pk}E(P^{pk})^T, \quad \tilde{A} = P^{pk}A(P^{pk})^T, \quad \tilde{B} = P^{pk}B(P^{pk})^T, \\ \tilde{C} = P^{pk}C(P^{pk})^T, \quad \tilde{D} = P^{pk}D(P^{pk})^T, \end{aligned} \tag{6}$$

where

$$\tilde{E} = \begin{pmatrix} E^{pp} & E^{pk} \\ E^{kp} & E^{kk} \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} A^{pp} & A^{pk} \\ A^{kp} & A^{kk} \end{pmatrix}. \tag{7}$$

Due to our previous assumptions (i.e. a single input acts only on the subsystem p or the subsystem k, a single output originates only from the subsystem p or the subsystem k and the feed-through matrix maps the inputs that act on a single subsystem only to the outputs that correspond to the same subsystem), the matrices $\tilde{B}$, $\tilde{C}$ and $\tilde{D}$ are block-diagonal

$$\tilde{B} = \begin{pmatrix} B^p & 0 \\ 0 & B^k \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} C^p & 0 \\ 0 & C^k \end{pmatrix}, \quad \tilde{D} = \begin{pmatrix} D^p & 0 \\ 0 & D^k \end{pmatrix}. \tag{8}$$

Consequently, we have two subsystems that are coupled through both matrices $\tilde{E}$ and $\tilde{A}$. The subsystem p is defined by

$$\begin{aligned} E^{pp}\dot{x}^p + E^{pk}\dot{x}^k &= A^{pp}x^p + A^{pk}x^k + B^p d^p, \\ z^p &= C^p x^p + D^p d^p, \end{aligned} \tag{9}$$

and the subsystem k is defined by

$$\begin{aligned} E^{kp}\dot{x}^p + E^{kk}\dot{x}^k &= A^{kp}x^p + A^{kk}x^k + B^k d^k, \\ z^k &= C^k x^k + D^k d^k. \end{aligned} \tag{10}$$

We eliminate $\dot{x}^k$ from (9) and $\dot{x}^p$ from (10), and after some manipulation we finally obtain two interconnected subsystems. The subsystem p is defined by

$$\begin{aligned} E^p\ddot{x}^p &= A^p x^p + v^{pk} + B^p d^p, \\ z^p &= C^p x^p + D^p d^p, \\ w^{pk} &= C^{pk}x^p + D^{pk}d^p, \end{aligned} \tag{11}$$

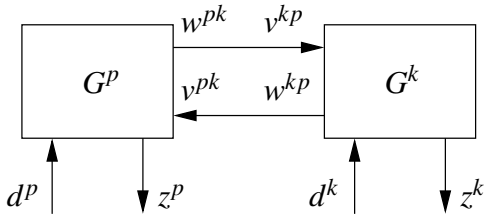


Figure 3: Interconnected systems.

where

$$\begin{aligned} E^p &= E^{pp} - E^{pk}(E^{kk})^{-1}E^{kp}, \quad A^p = A^{pp} - E^{pk}(E^{kk})^{-1}A^{kp}, \\ C^{pk} &= A^{kp} - E^{kp}(E^{pp})^{-1}A^{pp}, \quad D^{pk} = -E^{kp}(E^{pp})^{-1}B^p. \end{aligned} \tag{12}$$

In the similar fashion, the subsystem k is defined by

$$\begin{aligned} E^k\ddot{x}^k &= A^k x^k + v^{kp} + B^k d^k, \\ z^k &= C^k x^k + D^k d^k, \\ w^{kp} &= C^{kp} x^k + D^{kp} d^k, \end{aligned} \tag{13}$$

where

$$\begin{aligned} E^k &= E^{kk} - E^{kp}(E^{pp})^{-1}E^{pk}, \quad A^k = A^{kk} - E^{kp}(E^{pp})^{-1}A^{pk}, \\ C^{kp} &= A^{pk} - E^{pk}(E^{kk})^{-1}A^{kk}, \quad D^{kp} = -E^{pk}(E^{kk})^{-1}B^k. \end{aligned} \tag{14}$$

By setting $v^{kp} = w^{pk}$ and $v^{pk} = w^{kp}$, the subsystems p and k are interconnected LTI systems with corresponding transfer functions G^p and G^k, as shown in Figure 3.

Note that the terms E^{pk} and E^{kp} in the above equations are due to the coupling terms in the inertia matrices of the subsystems p and k. To be more specific, the terms E^{pk} and E^{kp} are due to the nondiagonal terms in the inertia matrices of the finite elements that contain DOFs from both subsystems p and k. Consequently, if the inertia matrices of the finite elements that contain both x^p and x^k are lumped (i.e. diagonal), the terms E^{pk} and E^{kp} vanish and the interconnected subsystems are defined by

$$\begin{aligned} E^{pp}\dot{x}^p &= A^{pp} x^p + v^{pk} + B^p d^p, \\ z^p &= C^p x^p + D^p d^p, \\ w^{pk} &= A^{kp} d^p, \end{aligned} \tag{15}$$

and

$$\begin{aligned} E^{kk}\dot{x}^k &= A^{kk} x^k + v^{kp} + B^k d^k, \\ z^k &= C^k x^k + D^k d^k, \\ w^{kp} &= A^{pk} d^k. \end{aligned} \tag{16}$$

Partitioning an arbitrary discretized system into a set of interconnected systems appears to be a straightforward extension of the above described procedure — we simply recursively apply the procedure to divide the resulting subsystems into a desired set of (new) subsystems. However, if we also require that the interconnection graph between the resulting subsystems matches the interconnection graph of the respective physical domains, the inertia matrices must meet a certain condition. We illustrate this in the following example.

3 EULER BEAM AS A STRUCTURED DISCRETIZED SYSTEM

Consider an Euler beam subject to a series of concentrated loads, as shown in Figure 4. If the distance between each two adjacent loads equals l, and the beam properties (flexural rigidity EI, cross section area A and density ρ) are constant, we may view the beam as an assembly of four identical segments, which we denote by 1 through 4. The remaining parts of the beam on the left and the right side are denoted by L and R, respectively. By specifying the desired outputs (e.q. beam displacements at the positions of the loads) and after the discretization, as described in the previous sections, we end up with the LTI second order system (1) which represents the entire beam.

For the purpose of our discussion, we wish to partition the resulting LTI system into a set of subsystems, such that each subsystem represents individual beam segments (L, R and 1 thorough 4). Furthermore, we also require that the interconnection graph for the resulting subsystems matches the interconnection graph for the actual beam segments, as shown in Figure 5. Without making any assumptions regarding the beam discretization, we proceed by recursively partitioning the LTI system that represents the entire beam.

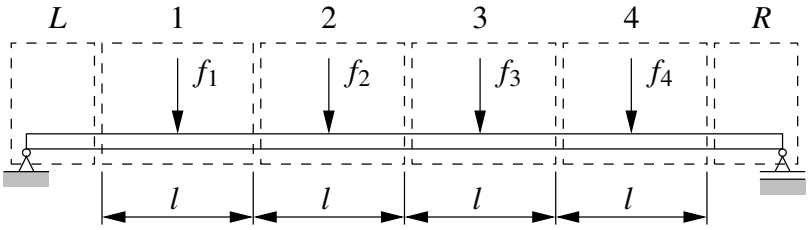


Figure 4: Euler beam as an assembly of segments.

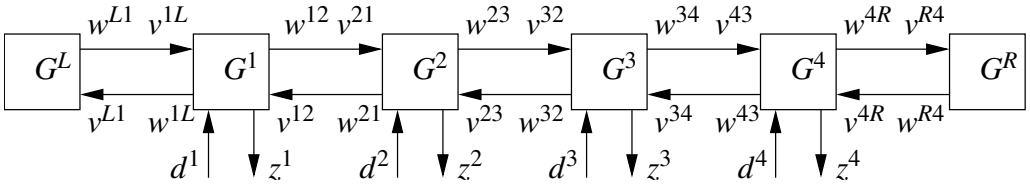


Figure 5: Euler beam as a series of interconnected systems.

First, we partition the system into the subsystem G^L that represents the segment L, and the subsystem G^K that represents the remaining segments (segments 1 through 4 and R). The resulting interconnected systems are shown in Figure 6, and the subsystems are defined as

- subsystem G^L:

$$\begin{aligned} E^L\ddot{x}^L &= A^L x^L + v^{LK}, \\ w^{LK} &= C^{LK} x^L, \end{aligned} \tag{17}$$

- subsystem G^K:

$$\begin{aligned} E^K\ddot{x}^K &= A^K x^K + v^{KL} + B^K d^K, \\ z^K &= C^K x^K + D^K d^K, \\ w^{KL} &= C^{KL} x^K + D^{KL} d^K. \end{aligned} \tag{18}$$

Now, let us take a closer look at the signal $v^{KL} = w^{LK} = C^{LK}x^L$. Observe that the matrix

$$C^{LK} = A^{KL} - E^{KL}(E^{LL})^{-1}A^{LL} \tag{19}$$

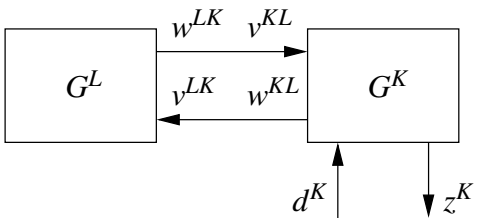


Figure 6: First partition of the beam.

contains the term $-E^{KL}(E^{LL})^{-1}A^{LL}$, which has no special structure (i.e. it is full) due of the presence of $(E^{LL})^{-1}$. The fact that E^{LL} represents a part of the inertia matrix which is sparse and banded does not help here — its inverse is full matrix, as we illustrate on the Figure 7. Consequently, the matrix C^{LK} is full and all state variables x^L are fed via v^{KL} to all DOFs of the subsystem K.

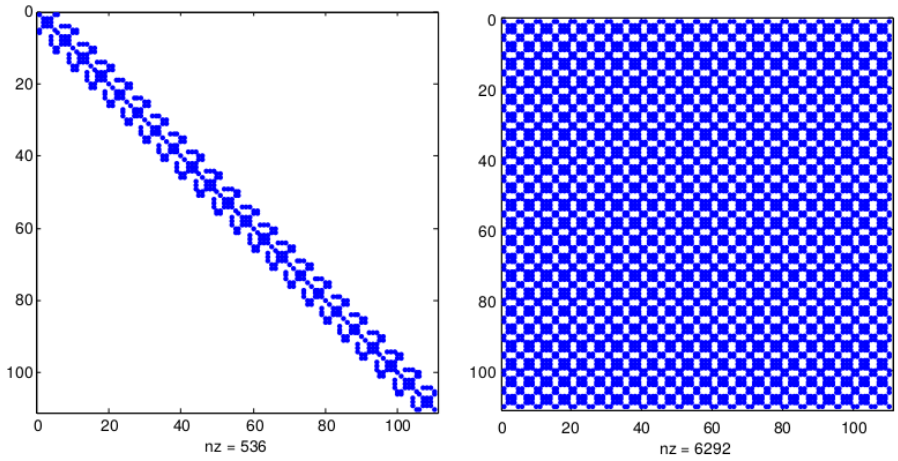


Figure 7: Nonzero entries of the inertia matrix (left) and its inverse (right).

Next, let us further partition the subsystem G^K into a subsystem G^1 that represents the beam segment 1 and a subsystem G^M that represents the beam segments 2 through 4 and R. Due to the fact that x^L are fed to all DOFs of the subsystems G^1 and G^M, i.e. the subsystems that make up the subsystem G^K, we have the interconnection graph shown on Figure 8. Further partitioning of the system would result in a set of subsystems where each subsystem is connected to *all* remaining subsystems.

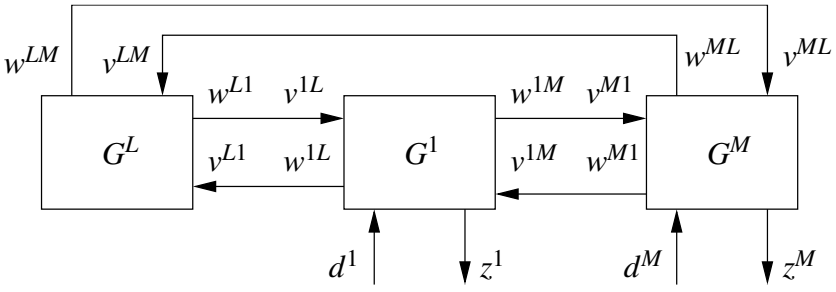


Figure 8: Second partition of the beam.

To remedy this, let us go back to the equation (19). Recall that when the inertia matrices of

the finite elements that contain both x^L and x^K are lumped, the terms E^{KL} vanishes and we have sparse and structured

$$C^{LK} = A^{KL}. \tag{20}$$

In this case, the state variables x^L are fed via v^{KL} only to the DOFs of the subsystem K which belong to the elements that contain DOFs from both subsystems L and K. In other words, only the states exchanged between the systems are the states of the DOFs at the boundaries between the systems. By taking this requirement into the account, recursive partitioning of the system results in a set of subsystems with the desired interconnection structure shown in Figure 5.

Thus, we conclude our discussion with a general condition for preserving the interconnection graph for a partitioned discretized system: an interconnection graph for a set of subsystems obtained by partitioning a discretized system of will have the same interconnection graph as the subdomains of the discretized system if the elements that represent the boundaries between the subsystems have lumped inertia matrices.

4 CONCLUSIONS

In the discussion above, we have presented a general framework for partitioning a linear time-invariant system system (e.q. a system obtained by means of the finite element method) into a set of interconnected systems. In doing so, we have derived a general condition for preserving the interconnection structure, i.e. a necessary condition on the discretized system matrices such that the interconnection graph between the resulting subsystems matches the interconnection graph between the respective physical subdomains. This framework allows us to address a local discretization error (i.e. an error in a subsystem) influence on the overall system response — for a possible approach, please refer to [6]. Furthermore, by considering a series of successively finer discretizations of an individual subsystem, we are able to construct a series of successively finer discretized system models with respective uncertainties.

5 ACKNOWLEDGMENTS

This work has been fully supported by Croatian Science Foundation under the project number 9354, Control of Spatially Distributed Systems.

REFERENCES

[1] C. Grossmann, H.-G. Roos, *Numerical Treatment of Partial Differential Equations*. Springer-Verlag, 2007.

[2] K.-J. Bathe, *Finite Element Procedures*. Prentice Hall, 1996.

[3] R.D. Cook, D.S. Malkus, M.E. Plesha, R.J. Witt, *Concepts and Applications of Finite Element Analysis, 4th Edition*. John Wiley & Sons, 2002.

[4] G.E. Dullerud, F. Paganini, *A Course in Robust Control Theory – A Convex Approach*. Springer-Verlag, 2000.

[5] D. Siljak, *Decentralized Control of Complex Systems*. Academic Press, 1991.

[6] B.L. Jones, E.C. Kerrigan, When is the discretization of a spatially distributed system good enough for control?. *Automatica*, **46**(9), 1462–1468, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

WAVE PROPAGATION IN THIN PRETWISTED ANISOTROPIC STRIPS

Maloth Thirupathi[1], **Mira Mitra**[2] **and P. J. Guruprasad**[2]

[1] Department of Aerospace Engineering,
Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
e-mail: maloththirupathi@gmail.com

[2] Department of Aerospace Engineering,
Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
e-mail: {mira,pjguru}@aero.iitb.ac.in

Keywords: Composite Materials, Vibration, Variational Asymptotic Method, Spectral Finite Element Method

Abstract. *This paper presents an asymptotically exact cross-sectional framework coupled with spectrally formulated one-dimensional (1-D) finite element model for the study of wave propagation in thin pre-twisted anisotropic strips. The cross-sectional model used here is based on the dimensional reduction of laminated shell theory to nonlinear 1-D theory using the variational asymptotic method (VAM). A linearized version of this model (zeroth-order asymptotic analysis) is used in the development of a 1-D spectral finite element (SFE) for a pre-twisted strip. Within this modeling framework, an exact dynamic stiffness matrix is derived, as the governing equations are solved using the exact interpolating functions in frequency-wavenumber domain. Compared to regular three-dimensional (3-D) finite element (FE) analysis, this model requires less computation cost since single element is sufficient to capture the frequency response of the strips. For numerical validation of the model, the natural frequencies of the strip without delamination are compared to the data available in literature. Finally, the developed framework is used to simulate wave propagation due to modulated sinusoidal pulse input.*

1 INTRODUCTION

Thin pre-twisted beams, made of laminated composites, are increasingly being used in helicopter rotor blades, wind turbine blades and propellers. In the case of helicopter rotor blades they are used to accommodate the centrifugal force along with flapping, lead-lag, and torsional motion using a flexible structural component known as flexbeams. This allows for the design of a rotor system that is independent of bearings making the design bearing-less [1]. However, the design of flexbeams and flexbeam like structures found in other applications is challenging owing to the nonlinearity that arises because of the large displacements and moderate rotations. This geometrical nonlinearity problem coupled with the anisotropy at the material scale adds to the complexity of the design. This challenge is further compounded when the design has to account for damages typical of laminated composites, e.g., delaminations [2]. Generally, it is desirable to develop a structural health monitoring (SHM) strategy, preferably online, which can aid in the early diagnosis of damages. Over the last few decades, SHM techniques based on dynamic identification techniques (see [3]) have gained prominence. In particular, wave propagation based technique is being extensively used owing to its ability to identify smaller defects, and potential to be used for online SHM [4]. The primary objective of the present work is towards the development of a computationally efficient model, which can not only accurately describe the overall deformation characteristics of the pre-twisted strip but also capture its dynamic behavior. It is envisaged that such a model will be an ideal framework based on which SHM system for damage detection in pre-twisted anisotropic strips can be developed.

For structures with one of it's dimensions much larger than the other two, simple beam models are often used in the preliminary design stage because they can provide valuable insight into the behavior of the structures with much less effort. These models, in general, are developed on $adhoc$ assumptions. These traditional 'ordering scheme' based methodologies can be circumvented when the exact kinematics of the beam reference line displacement and cross-sectional rotation is accounted for in the development of the geometrically-exact equations of motion. $Adhoc$ methodologies have been replaced by rigorous beam modeling (easily extendable to plates and shells) technique, incorporating the kinematics as described above, through the variational asymptotic method (VAM) [5, 6]. It has also been used to model pre-twisted anisotropic strips both in the absence and presence of delaminations [2, 7]. The present work extends the scope of these studies to investigate vibration and wave propagation in pre-twisted strips.

Owing to the importance of pre-twisted anisotropic strips, considerable work has been done on studying the vibration and dynamic characteristics of pre-twisted strips. For brevity, only few important works are reported here. Rosard [8], Anliker and Troesch [9], Gupta and Rao [10] investigated the free vibration characteristics of pre-twisted beams. Carnegie and Thomas extended these works to investigate the coupled

bending vibration of pre-twisted beam [11]. More recently, Filiz et al. [12] developed a 3-D model to investigate the dynamics of pre-twisted beam using the spectral-Tchebychev solution. Even though much work has been done on vibration and dynamic behavior of pre-twisted strips, not much work has been done on wave propagation in these structures, in particular.

Conventional finite element (FE) analysis of pre-twisted anisotropic strips is difficult due to lack of computational resources. This takes us to the realm of frequency domain analysis and in particular, spectral analysis. This spectral analysis is used to construct a frequency domain based matrix methodology proposed by Doyle [13], which is called spectral finite element (SFE) method. The first steps in the development of SFE involves the transformation of the governing wave equations from time domain to frequency-wavenumber domain using discrete Fourier transform (DFT). This process converts the governing coupled partial differential equations (PDEs) into a set of ordinary differential equations (ODEs) in the spatial domain which can be solved by standard methods. The resulting complex shape function matrix is a linear superposition of all the wave modes. Following the conventional FE method, the complex dynamic stiffness matrix is then formed which is exact. As shown in [13], this process leads to an efficient model suitable for use within the framework of automated FE method. The primary difference between conventional FE and SFE methods is that at the elemental nodes spectral amplitudes of the nodal variables are determined at each frequency steps in SFE method. On the contrary, pseudo-static variables are evaluated at each time steps or eigen frequencies in standard FE technique.

In the current paper, SFE has been developed to find the natural frequencies of pre-twisted anisotropic strips based on dimensional reduction of laminated shell theory to a one dimensional non-linear beam theory, and validated with the data available in literature. Later, it is used to simulate wave propagation using narrow banded modulated sinusoidal pulse as input. As mentioned earlier, though SFE has been used extensively for vibration and wave propagation study of composite beams and laminates [14], it has not been developed for pre-twisted beams.

The manuscript is organized as follows. Section 2 provides the mathematical formulation of the governing equations using VAM. In Section 3, The SFE formulation is described briefly. The numerical results are presented in Section 4. The paper ends with a section on important conclusions.

2 MATHEMATICAL FORMULATION

The development of the kinematics for this problem follows the procedure enumerated in [7]. Here, the details are omitted and only the general procedure is enumerated. The overall deformation of a cross-section contains all forms of warping and location moderate rotations. Following the classical laminated shell theory, the 3-D strain measures are related to 2-D strain measures by $\Gamma_{\alpha\beta} = \epsilon_{\alpha\beta} + x_3\rho_{\alpha\beta}$, where, $\epsilon_{\alpha\beta}$ are membrane

strains and $\rho_{\alpha\beta}$ are middle surface bending curvatures. Starting from the kinematics, then applying VAM in an iterative manner, relations between membrane strains and beam 1-D strains are obtained. The strip, being considered a 2-D elastic body, it's strain energy density can be written as $U_{2D} = \frac{1}{2}\epsilon^T[\mathbf{AB}|\mathbf{BD}]\epsilon$. Here, $\mathbf{A}$, $\mathbf{D}$, $\mathbf{B}$ are the familiar membrane, bending, and coupling stiffness matrices, respectively. The 1-D strain energy density can be obtained by integrating the 2-D strain energy density along the width direction. However, the unknown warping terms, which are functions of x_2 have to be predetermined to carry out the integration which can be done at any given order employing variational minimization principles. In doing so, a 9×9 stiffness matrix is obtained, where the first 4×4 terms represent the linear stiffness matrix (obtained from zeroth order approximation). The above described procedure can also be followed to determine the cross-sectional stiffness terms in the presence of delamination [2].

In the present work, focus is limited to linear analysis. Hence, the kinetic energy of the beam, including three translation degrees of freedom (DOF) and rotation about x_1, excluding the other variables for contributing to higher order terms, can be determined as

$$K = \frac{1}{2}\begin{Bmatrix} \dot{u}_1 \\ \dot{\theta}_1 \\ \dot{u}_3 \\ \dot{u}_2 \end{Bmatrix}^T \begin{bmatrix} \rho A & 0 & 0 & 0 \\ 0 & \rho I_p & 0 & 0 \\ 0 & 0 & \rho A & 0 \\ 0 & 0 & 0 & \rho A \end{bmatrix} \begin{Bmatrix} \dot{u}_1 \\ \dot{\theta}_1 \\ \dot{u}_3 \\ \dot{u}_2 \end{Bmatrix} \tag{1}$$

Applying Hamilton's principles, the governing equations can be written as,

$$S_{11}\frac{\partial^2 u_1}{\partial x_1^2} + S_{12}\frac{\partial^2 \theta_1}{\partial x_1^2} - S_{13}\frac{\partial^3 u_3}{\partial x_1^3} + S_{14}\frac{\partial^3 u_2}{\partial x_1^3} = \rho A\frac{\partial^2 u_1}{\partial t^2} \tag{2}$$

$$S_{12}\frac{\partial^2 u_1}{\partial x_1^2} + S_{22}\frac{\partial^2 \theta_1}{\partial x_1^2} - S_{23}\frac{\partial^3 u_3}{\partial x_1^3} + S_{24}\frac{\partial^3 u_2}{\partial x_1^3} = \rho I_p\frac{\partial^2 \theta_1}{\partial t^2} \tag{3}$$

$$S_{14}\frac{\partial^3 u_1}{\partial x_1^3} + S_{24}\frac{\partial^3 \theta_1}{\partial x_1^3} - S_{34}\frac{\partial^4 u_3}{\partial x_1^4} + S_{44}\frac{\partial^4 u_2}{\partial x_1^4} = -\rho A\frac{\partial^2 u_2}{\partial t^2} \tag{4}$$

$$-S_{13}\frac{\partial^3 u_1}{\partial x_1^3} - S_{23}\frac{\partial^3 \theta_1}{\partial x_1^3} + S_{33}\frac{\partial^4 u_3}{\partial x_1^4} - S_{34}\frac{\partial^4 u_2}{\partial x_1^4} = -\rho A\frac{\partial^2 u_3}{\partial t^2} \tag{5}$$

where, S_{ij} are the stiffness terms, which can be readily determined for healthy and delaminated pre-twisted strips [2, 7], ρ is the material mass density, A cross-sectional area and I_p is the polar moment of inertia. Since the sectional resultants are conjugates to the 1-D beam strains in terms of the 1-D strain energy density, the stiffness constitutive law is obtained, which serve as boundary conditions. For this problem, they can be

written in compact form as,

$$\begin{Bmatrix} F_1 \\ M_1 \\ M_2 \\ M_3 \end{Bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{12} & S_{22} & S_{23} & S_{24} \\ S_{13} & S_{23} & S_{33} & S_{34} \\ S_{14} & S_{24} & S_{34} & S_{44} \end{bmatrix} \begin{Bmatrix} \gamma_{11} \\ \kappa_1 \\ \kappa_2 \\ \kappa_3 \end{Bmatrix} \tag{6}$$

where, F_1 is the resultant axial force, M_1 being the twist, M_2, M_3 are the bending moments about 2 and 3 axes respectively.

2.1 SPECTRAL FINITE ELEMENT (SFE) FORMULATION

The SFE formulation begins by assuming a solution for the displacement field. In particular, using discrete Fourier transformation (DFT), the displacement vectors can be represented as,

$$\mathbf{u}(x_1, t) = \sum_{n=1}^{N} \hat{\mathbf{u}}(x_1, \omega_n) e^{-i\omega_n t} \tag{7}$$

where, $i = \sqrt{-1}$, ω_n is the frequency at n^{th} sampling point and N is the Nyquist point in the fast Fourier transform (FFT) used for numerical implementation. Substituting Eq. 7 in Eq. 2-5, a set of ODEs are obtained for the Fourier coefficients, $\hat{u_1}(x_1)$, $\hat{\theta}_1(x_1)$, $\hat{u_3}(x_1)$ and $\hat{u_2}(x_1)$, and can be solved, since it contains constant coefficients, by writing $\tilde{u_1}e^{-ikx_1}$, $\tilde{\theta_1}e^{-ikx_1}$, $\tilde{u_3}e^{-ikx_1}$ and $\tilde{u_2}e^{-ikx_1}$, respectively. Here, k is the wave number in x_1 direction, yet to be determined and $\tilde{u_1}$, $\tilde{\theta}_1$, $\tilde{u_3}$ and $\tilde{u_2}$ are unknown constants. Substituting these forms in the ODEs, a polynomial eigenvalue problem (PEP) is posed to find $(k, \boldsymbol{v})$ as below

$$\boldsymbol{\Psi}(k)\boldsymbol{v} = \left(k^4\boldsymbol{A}^4 + k^3\boldsymbol{A}^3 + k^2\boldsymbol{A}^2 + k\boldsymbol{A} + \boldsymbol{A}_0\right)\boldsymbol{v} = 0, \qquad \boldsymbol{v} \neq 0, \tag{8}$$

where, $\boldsymbol{A}^i \epsilon\ \mathbb{C}^{4\times4}$, k is an eigenvalue and $\boldsymbol{v}$ is the right eigenvector. For brevity, the explicit expressions are not given in the paper. Since $\boldsymbol{A}^4$ is singular [16], $\boldsymbol{\Psi}(k)$ is not regular and admits infinite eigenvalues [15]. This PEP can be solved to find the wavenumbers by several methods available. The one adopted in this paper is linearization of PEP method.

Following the solution of the PEP problem given by Eq. 8, the displacements presented in Eq. 7 can also be written in a compact form in frequency-wavenumber domain, as,

$$\boldsymbol{u}(x, t) = \sum_{n=1}^{N} \hat{\boldsymbol{u}}(x_1, \omega_n) e^{-i\omega_n t} = \sum_{n=1}^{N} \left(\sum_{j=1}^{12} \tilde{\boldsymbol{u}}_j e^{-ik_j x} \right) e^{-i\omega_n t} \tag{9}$$

where, k_j is the wavenumber associated with the jth mode of wave at ω_n. The generic

displacement vector at any x_1 location can be given explicitly as,

$$\hat{\boldsymbol{u}}(x_1,\omega_n) = \left\{ \begin{array}{c} \hat{u_1}(x_1,\omega_n) \\ \hat{\theta_1}(x_1,\omega_n) \\ \hat{u_3}(x_1,\omega_n) \\ \hat{u_2}(x_1,\omega_n) \end{array} \right\} = \left[\begin{array}{ccccc} R_{11} & R_{12} & \dots & R_{1\,11} & R_{1\,12} \\ R_{21} & R_{22} & \dots & R_{2\,11} & R_{2\,12} \\ R_{31} & R_{32} & \dots & R_{3\,11} & R_{3\,12} \\ R_{41} & R_{42} & \dots & R_{4\,11} & R_{4\,12} \end{array} \right] [\boldsymbol{D}] \left\{ \begin{array}{c} c_1 \\ c_2 \\ \vdots \\ c_{11} \\ c_{12} \end{array} \right\} \tag{10}$$

where, $[\boldsymbol{D}]$ is diagonal matrix of order 12×12 and can be expressed as,

$$[D] = \left\{ \begin{array}{ll} e^{-ik_j x} & j \text{ odd} \\ e^{-ik_j (L-x)} & \text{otherwise} \end{array} \right. \tag{11}$$

where, L is length of the element. The entries associated with $j = 1, 3, ...9, 11$ and $j = 2, 4, ...10, 12$ contribute respectively to forward and backward propagating waves. It is to be noted that all these expressions are evaluated at a particular value of ω_n. The $\boldsymbol{R}_{4\times 12}$ matrix is called amplitude ratio or eigenvector matrix, obtained as the solution of the PEP given by Eq. 8. $\boldsymbol{c} = \{c_1, c_2, \dots, c_{12}\}$ are unknowns constants, to be found out by applying boundary conditions. These unknowns are written in terms of nodal displacements, namely, $x = 0$ and $x = L$.

Eq. 10 can be written as,

$$\hat{\boldsymbol{u}}(x_1,\omega_n) = [\boldsymbol{R}]\,[\boldsymbol{D}]\boldsymbol{c} = [\boldsymbol{T}_1(x_1,\omega_n)]\boldsymbol{c} \tag{12}$$

Eq. 12 at the two nodes, *i.e.*, $x_1 = 0$ and $x_1 = L$ gives

$$\hat{\boldsymbol{u}}_e = \left[\begin{array}{c} \boldsymbol{T}_1(0,\omega_n) \\ \boldsymbol{T}_1(L,\omega_n) \end{array} \right] \boldsymbol{c} = [\boldsymbol{T_2}]\boldsymbol{c} \tag{13}$$

where, $[\boldsymbol{T}_2]$ is a non-singular 12×12 complex matrix and represents the local wave characteristics of the displacement field. Further, Eq. 7 can be used to eliminate unknown constants $\boldsymbol{c}$ from Eq. 13 as,

$$\hat{\boldsymbol{u}}(x_1,\omega_n) = [\boldsymbol{T}_1(x_1,\omega_n)][\boldsymbol{T}_2]^{-1}\hat{\boldsymbol{u}}_e = [\boldsymbol{\zeta}(x_1,\omega_n)]_e\hat{\boldsymbol{u}}_e, \tag{14}$$

where, $[\boldsymbol{\zeta}(x_1,\omega_n)]_e$ is the exact element shape function matrix that connects the nodal displacement vector to the generic displacement vector at arbitrary location along the length of the beam.

Next, using the forced boundary conditions given in Eq. 6, and adding the additional shear or transverse forces $F_2 = \frac{dM_3}{dx_1}$ and $F_3 = \frac{dM_2}{dx_1}$ for the additional DOF, the force vector can be computed in the terms of unknown constants $\boldsymbol{c}$ as,

$$\hat{\boldsymbol{f}} = [\boldsymbol{P}(x_1,\omega_n)]\,\boldsymbol{c} \tag{15}$$

Eq. 15 is evaluated at two nodes, and the nodal force vectors are obtained as,

$$\hat{\boldsymbol{f}}_e = \left\{ \begin{array}{c} \hat{\boldsymbol{f}}_1 \\ \hat{\boldsymbol{f}}_2 \end{array} \right\} = \left[\begin{array}{c} \boldsymbol{P}(0,\omega_n) \\ \boldsymbol{P}(L,\omega_n) \end{array} \right] \boldsymbol{c} = \left[\begin{array}{c} \boldsymbol{P}(0,\omega_n) \\ \boldsymbol{P}(L,\omega_n) \end{array} \right] [\boldsymbol{T}_2]^{-1} \hat{\boldsymbol{u}}_e = [\boldsymbol{K}_e]\, \hat{\boldsymbol{u}}_e \qquad (16)$$

where, $[\boldsymbol{K}_e]$ is the exact dynamic stiffness matrix at ω_n relating the nodal displacements and forces in frequency domain.

3 NUMERICAL RESULTS AND DISCUSSION

First, the natural frequencies of the pre-twisted anisotropic strips are extracted from the the exact dynamic stiffness matrix given in the formulation. A structurally coupled $AS4/3501-6$ graphite-epoxy strip is considered for obtaining the natural frequencies of the beam. The beam material and geometrical properties are as given below in Table 1 and 2, respectively.

Property	Value
E_{11}(GPa)	142
E_{22}/E_{33}(GPa)	9.8
G_{12}(GPa)	6.0
G_{13}/G_{23}(GPa)	4.8
ν_{12}	0.3
ν_{13}/ν_{23}	0.34
ρ (kg/m^3)	1580

Table 1: Material properties of AS4/3501-6 graphite epoxy composite.

Dimension	Value
Length (m)	0.56
Breadth (m)	0.03
Thickness (m)	0.0147

Table 2: Geometrical properties of AS4/3501-6 graphite epoxy composite test specimen.

The natural frequencies of the beam are plotted with different pretwist values and are compared with the data from [17] and are shown in the Figure 1. The first nine modes shown are in the order of first, second, and third flapwise bending (F1, F2, F3), first

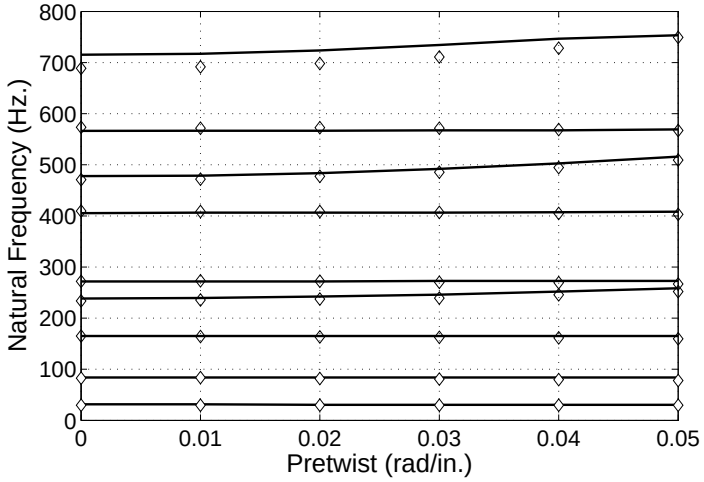


Figure 1: Natural frequencies of the beam marked by the sharp peaks. Data points show are from Ref. [17].

torsion (T1), fourth flapwise bending (F4), second torsion (T2), fifth and sixth flapwise bending (F5, F6), and third torsion (T3). It is observed that pretwist affects the free vibration characteristics of the strip and the current model predictions agree well with those presented in [17].

Next, $AS4/3501-6$ graphite epoxy beam is considered with a pre-twist of $k_1 = 0.05\,rad/in;$ for wave propagation analysis. To get rid of the reflected wave forms from the boundaries a semi-infinite beam is considered for the wave propagation analysis. This configuration will result in bending-twist coupling. A torsional loading of sinusoidal pulse modulated at $5\,kHz$ is applied at the free end. Figure 2 shows the torsional response at a distance from the free end under the torsional loading. The bigger blob represents the torsional mode and the smaller blob represents the flap-wise flexural mode.

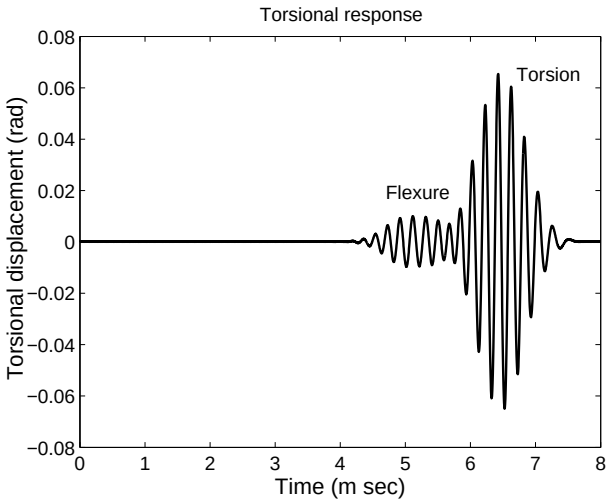


Figure 2: Torsion response due to modulated sinusoidal input

4 CONCLUSIONS

Wave propagation in pre-twisted anisotropic strip is modeled using spectral finite element method. The framework developed combines the advantages of beam modeling using variational asymptotic method and capturing the dynamics from spectral formulation. The model developed is validated by determining the natural frequency of the strip as a function of pretwist and comparing with results available in the literature. Model predictions are in good agreement with literature. Wave response simulated using the framework is presented to highlight the capability of the technique to investigate wave characteristics with application to structural health monitoring.

REFERENCES

[1] P. P. Friedmann, Rotary-wing aeroelasticity: current status and future trends. *AIAA Journal*, **42(10)**, 1953–1972, 2004.

[2] P. J. Guruprasad, M. Thejasvi, D. Harursampath, Nonlinear analysis of a thin pretwisted and delaminated anisotropic strip. *Acta Mechanica*, **225(10)**, 2815–2832, 2014.

[3] O. S. Salawu, Detection of structural damage through changes in frequency: a review. *Engineering Structures*, **19(9)**, 718–723, 1997.

[4] M. Mitra, S. Gopalakrishnan, Wavelet based spectral finite element modelling and detection of de-lamination in composite beams. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **462(2070)**, 1721–1740, 2006.

[5] C. E. S. Cesnik, D. H. Hodges, VABS: A new concept for composite rotor blade cross-sectional modeling. *Journal of the American Helicopter Society*, **42(1)**, 27–38, 1997.

[6] D. H. Hodges, Geometrically exact, intrinsic theory for dynamics of curved and twisted anisotropic beams. *AIAA Journal*, **41(6)**, 1131–1137, 2003.

[7] D. H. Hodges, D. Harursampath, V. V. Volovoi, C. E. S. Cesnik, Non-classical effects in non-linear analysis of pretwisted anisotropic strips. *International Journal of Non-linear Mechanics*, **34(2)**, 259-277, 1999.

[8] D. D. Rosard, Natural frequencies of twisted cantilever beams. *Journal of Applied Mechanics*, 241–244, 1953.

[9] M. Anliker, A. Troesch, Lateral vibrations of pretwisted rods, *Journal of Applied Mathematics and Physics*, **14(3)**, 218–236, 1963.

[10] R. S. Gupta, S. S. Rao, Finite element eigenvalue analysis of tapered and twisted Timoshenko beams *Journal of Sound and Vibration*, **56(2)**, 187–200, 1978.

[11] W. Carnegie, J. Thomas, Coupled bending. Bending vibration of pre- twisted tapered blading *Journal of Engineering for industry Transactions A.S.M.E. Series B*, **94(1)**, 255–266, 1972.

[12] S. Filiz, B. Bediz, L. A. Romero, O. B. Ozdoganlar, Three dimensional dynamics of pretwisted beams: A spectral-Tchebychev solution *Journal of Sound and Vibration*, **333(10)**, 2823-2839, 2014.

[13] J. F. Doyle, *Wave propagation in structures, Mechanical Engineering Series, Second Edition*. Springer, 1997.

[14] S. Gopalakrishnan, A. Chakraborty, D. R. Mahapatra, *Spectral Finite Element Method*. Springer, 2008.

[15] F. Tisseur, N. J. Higham, Structured pseudospectra for polynomial eigenvalue problems, with applications *SIAM Journal on Matrix Analysis and Applications*, **23(1)**, 187–208, 2002.

[16] P. Lancaster, M. Tismenetsky, *Theory of Matrices with Applications, Second Edition*. Academic Press, 1985.

[17] R. K. Kovvali, D. H. Hodges, Verification of variational-asymptotic sectional analysis for initially curved and twisted beams, *Journal of Aircraft*. **49(3)**, 861–869, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

NUMERICAL ROUTINE FOR DYNAMIC ANALYSIS OF TRANSMISSION LINES GUYED TOWERS SUBMITTED TO BROKEN CABLE

Thiago Brazeiro Carlos[1], **João Kaminski Jr.**[2]

[1]University of Coimbra
Department of Civil Engineering
Rua Luís Reis Santos, Polo II da Universidade, 3030-378, Coimbra, Portugal
thiago_brazeiro@hotmail.com

[2] Federal University of Santa Maria
Department of Structures and Civil Construction
Av. Roraima, nº 1000 - Cidade Universitária, 97105-900, Santa Maria, Rio Grande do Sul, Brasil
jkj@ufsm.br

Keywords: Guyed Lattice Metallic Towers, Transmission Lines Tower, Cable Break, Direct Integration Method, Central Finite Differences.

Abstract. *In the design and sizing of guyed lattice metallic structures used in transmission lines (TL) as a way of simplification, static analyzes are usually carried, where the dynamic actions are represented by "equivalent static loads". However, the static response to this type of structure is not the most appropriate, since such structures are lightweight, slender and that are always subject to dynamic nature action, except the own weight. Therefore, a dynamic analysis is critical to get a more precise result in terms of efforts on the bars and nodal displacements. In this context, this paper deals with the development of a computational routine in Fortran, for obtaining the dynamic response of guyed metallic towers of TL, submitted to dynamic action of broken cable. For the solution of this problem is used the Direct Integration Method of the equations of motion, explicitly, whit central finite differences. In the output result files, the axial forces in all the bars are displayed along the time and the nodal displacements are recorded at certain intervals of time during the analysis, in order to facilitate the structural behavior viewing with animation in post-processing programs. The accuracy of the dynamic response is evaluated by analyzing the isolated guyed tower, subjected to an impulsive load, comparing the fundamental frequency, calculated on the top of tower displacements obtained in the time domain, at the same frequency determined by an modal analysis. Knowing that the method is conditionally stable, the stability is evaluated by analyzing two responses of the same model, obtained with two different integration time intervals and smaller than the critical interval, which must be identical*

1 INTRODUCTION

The main actions operating over transmission lines structures (TL) are dynamic in nature, such as wind and cable breakage. The wind is the main action and must be carefully considered in the design phase. The cable breakage, though occurring less frequently, is one of the actions that can take the tower to collapse, including causing the phenomenon known as "ripple effect" in which many towers collapse in sequence.

The rupture of a cable or bundle of cables, generates forces in the longitudinal towers TL, while in the wind action, the forces on the towers can be longitudinal or transverse. In the design of a tower, these actions should be considered as dynamic actions, however, for simplicity, are usually regarded as "equivalent static". Currently, this is no longer justified but, due to the great advances in numerical methods and in the computational area, making it possible to obtain the response of the structure through a dynamic analysis, which would provide the obtainment of more realistic results and consequently the design of a more efficient and economical structure.

The mechanical model usually adopted in TL metal lattice towers project is quite simple, using space frame elements on main legs and spatial lattice in the diagonal bracings, rigid joints (support or bezels) and pinned or fixed connections. The answer is usually obtained through a static and linear analysis. It should be noted that in some types of towers, for example, guyed towers, or freestanding towers with great height the nonlinear geometric analysis is common.

Determining the effects of the dynamic nature actions on TL towers is a complex task, given the range of variables involved and its randomness. The use of methods that result in more proximity to reality is essential to design the towers safely, seeking to maintain the efficiency and economy in their design.

Therefore, a program for performing a dynamic analysis of TL structures, in a simplified and precise manner, becomes necessary. The explicit direct integration method (DIM) of the equations of motion, with central finite differences, which presents a relatively simple formulation, can be used for the response over time in the towers bars, the cable elements and insulators of a TL, also allowing to treat physical and geometrical non-linearities with relative ease, besides the advantage of not requiring the assembly of the global stiffness matrix, since the integration is carried out at element level [1]. It is worth mentioning that the analysis is restricted to structures that can be discretized with the bar elements, in this case, spatial lattice elements.

2 OBJECTIVE

The main objective of this work is to develop a computational routine in FORTRAN, for the analysis of a complete excerpt of TL, including all the components in the model (towers, guy-wires, lightning rod cables, conductor cables and insulators chains) to obtain the dynamic response, in the time domain, using the direct integration method (DIM) of equations of motion explicitly applied to transmission lines with guyed lattice towers and subjected to cable breakage (conductor or lightning rod).

To this end, an adjustment in the routine developed by [2] was made, which allows the analysis of freestanding towers.

Then, the routine is tested in a numerical model of a full TL excerpt, considering all of its components with three guyed towers and four spans, simulating a dynamic loading from the rupture of a cable (conductor or lightning rod) in the model. The response is obtained in terms of displacements of the nodes and effort on the bars of the towers along the analysis time, besides the visualization of the frame moving in the time, from the set of nodal coordinates of

the entire TL portion generated at certain time intervals and visualized in post-processing programs.

3 EXPLICIT DIRECT INTEGRATION METHOD

In the solution of dynamic problems the explicit direct integration methods (DIM) have been frequently used, however, in TL structures analysis their application is not very common. The DIM theory is presented next, in a nutshell, explicitly, using central finite differences.

For the solution of the nodal coordinate vector of the structureIn which: $\boldsymbol{q}_{(t)}$ is the vector of nodal coordinates at time "t", $\dot{\boldsymbol{q}}_{(t)}$ is the vector of nodal velocities at time "t", $\ddot{\boldsymbol{q}}_{(t)}$ is the vector of nodal acceleration at time "t", $\boldsymbol{F}_{(t)}$ is the vector of nodal external forces at time "t" and $\boldsymbol{M}$, $\boldsymbol{D}$ and $\boldsymbol{S}$ are the matrices of mass, damping and stiffness of the structure, respectively.

Using the central finite differences to solve the equations of dynamic equilibrium, the equation for obtaining the nodal coordinates of the structure at each instant of time, in the directions x, y and z is summarized as follows:

$$\left[\frac{1}{\Delta t^2}\boldsymbol{M}+\frac{1}{2\Delta t}\boldsymbol{C}\right]\boldsymbol{q}_{(t+\Delta t)}=\boldsymbol{F}\left[\boldsymbol{K}-\frac{2}{\Delta t^2}\boldsymbol{M}\right]\boldsymbol{q}_{(t)}-\left[\frac{1}{\Delta t^2}\boldsymbol{M}-\frac{1}{2\Delta t}\boldsymbol{C}\right]\boldsymbol{q}_{(t-\Delta t)} \tag{2}$$

in which: t is the time (in seconds) and Δt is the integration time interval (in seconds).

After the system state is known at time "(t)" and "$(t-\Delta t)$", it is possible to calculate the second member of the equation (2), and then determine the system state in the next time interval "$(t+\Delta t)$". To begin the process, the initial conditions "$\boldsymbol{q}_{(t)}$" and "$\dot{\boldsymbol{q}}_{(t)}$" must be specified, since:

$$\boldsymbol{q}_{(0-t)} = \boldsymbol{q}_{(0-t)} - \Delta t\,\dot{\boldsymbol{q}}_{(0)} + \frac{\Delta t^2}{2}\,\ddot{\boldsymbol{q}}_{(0)} \tag{3}$$

The acceleration vector "$\ddot{\boldsymbol{q}}_{(0)}$" in the above equation can be calculated from Equation (1), considering the time $t=0$:

$$\ddot{\boldsymbol{q}}_{(0)}=\boldsymbol{M}^{-1}\left(\boldsymbol{F}_{(0)}-\boldsymbol{C}\cdot\dot{\boldsymbol{q}}_{(0)}-\boldsymbol{K}\cdot\boldsymbol{q}_{(0)}\right) \tag{4}$$

When a discrete mass matrix "*M*" and damping matrix "*C*" proportional to the mass matrix are used, therefore, "*M*" and "*C*" diagonal, the resulting matrix which multiplies "$\boldsymbol{q}_{(t+\Delta t)}$"in Eq. (2) will also result diagonal, and the problem is reduced to *n* algebraic equations whose solution is trivial. This is when the *explicit method* happens. Consequently, as only space truss elements will be used in the model, there will be no need to form the global stiffness matrix *K*, for the integration can be performed in the element level without the need to use a solution process of equations systems to determine the vector "$\boldsymbol{q}_{(t+\Delta t)}$", resulting in a significant reduction of the computational effort compared to implicit methods. Thus, the equation of central finite differences to calculate the displacement at any node of the structure in the directions *x*, *y* and *z* at time $(t+\Delta t)$, is given by:

$$q_{(t+\Delta t)} = \frac{1}{c_1}\left[\frac{\Delta t^2}{m}\,f_{(t)} + 2\,q_{(t)} - c_2\,q_{(t-\Delta t)}\right] \tag{5}$$

in which: $f_{(t)}$ is the component of the resulting nodal force in the corresponding direction (in Newton) at time t, $q_{(t)}$ is the coordinated of linear displacement of the structure node in the direction *x*, *y* or *z* (meters) at time t and *m* is the nodal mass (in kg). Being:

$$c_1 = \left(1 + \frac{c_m\,\Delta t}{2}\right) \tag{6}$$

$$c_2 = \left(1 - \frac{c_m\,\Delta t}{2}\right) \tag{7}$$

where: c_m is the damping coefficient proportional to the mass or proportionality constant ($c_m = c/m$) and c is the structural damping coefficient (in N.s/m). Substituting Equations (6) and (7) in Equation (5), finally the central finite difference equation results in:

$$q_{(t+\Delta t)} = \frac{1}{\left(1 + \frac{c_m\,\Delta t}{2}\right)}\left[\frac{\Delta t^2}{m}\,f_{(t)} + 2\,q_{(t)} - \left(1 - \frac{c_m\,\Delta t}{2}\right)q_{(t-\Delta t)}\right] \tag{8}$$

The resulting nodal force component $f_{(t)}$ is composed by the gravitational forces $f_{g(t)}$ (Own weight and external nodal forces) and the axial forces $f_{a(t)}$ in the lattice elements, due to axial deformation. The proportionality constant c_m must be adjusted so that the ratio of critical damping (ζ) is the same as the structure analyzed. For each integration step, that is, in the evaluation of Eq. (8) for all nodes in all directions, the updated nodal coordinates lead to axial deformation of the elements, which react with axial forces which oppose to the displacement. For an axial stiffness element "$E\,A$" and length "L" the axial force at a time t is given by:

$$f_{a(t)} = E\,A\left[\frac{L_{(t)} - L_{(0)}}{L_{(0)}}\right] \tag{9}$$

where: E is the longitudinal elastic modulus of the lattice material, A is the transverse section of the element, $L_{(0)}$ is the lattice element length at time $t = 0$ and $L_{(t)}$ is the lattice element length in a given time t.

Therefore, for obtaining the resulting nodal force components $f_{(t)}$, the axial force $f_{a(t)}$ must be multiplied by director cosines of the element axis in the deformed state, and such components added with the respective gravitational force $f_{g(t)}$ in the directions *x*, *y* and *z*, acting on the node under consideration at time t. It should be noted that, since the nodal coordinates are updated at each integration step, the geometric nonlinearity is always considered.

According to [3], the convergence and accuracy of the solution depend primarily on the integration time interval Δt adopted, therefore the method is conditionally stable. Then, according to [4], for this stability to be ensured, it is necessary that:

$$\Delta t \leq \Delta t_{crit} = \frac{2}{\omega_n} = \frac{2}{2\,\pi\,f_n} = \frac{T_n}{\pi} \tag{10}$$

where: Δt_{crit} is the critical integration time interval, f_n is the corresponding structure natural frequency of vibration with n degrees of freedom (in Hz), ω_n is the BIGGEST angular frequency of vibration of the structure with n degrees of freedom (in rad/s) and T_n It is SMALLEST structure natural period of vibration with n degrees of freedom (in seconds). In determining "Δt" through Eq. (10) the difficulty consists in calculating "T_n", which corresponds to the way of vibration associated with the largest eigenvalue of the structure "λ_n" (Or higher natural vibration frequency "f_n"). However, according to [5], for the case of structures formed by lattice components, the critical time interval "Δt_{crit}" can be given, approximately, by the following equation:

$$\Delta t_{crit} = \frac{L_{min(0)}}{\sqrt{\frac{E}{\rho}}} \tag{11}$$

in which: $L_{\min(0)}$ is the initial length of the shortest lattice bar (meters) at time $t = 0$, E is the longitudinal elastic modulus of the material (in N/m²), ρ is the specific mass of the material (in kg/m³).

According to Eq. (11), the integration time interval "Δt" that should be adopted for a structure of the lattice type, basically depends on the initial length of the shortest lattice bar and the speed of sound in the material used in the bar. It is important to note that Eq. (11) serves only to give an approximate idea of the value of "Δt". In order that the accuracy of the results is proved, without calculating the largest eigenvalue of the structure, the obtainment of at least two identical responses with two different values of "$\Delta t \leq \Delta t_{crit}$" is necessary. In the analysis of metal lattice towers according to [6] in order that the method stability is ensured, the time interval between integration steps (Δt) must be of the order of 10^{-5} to 10^{-6} seconds, which requires between 100,000 and 1,000,000 integration steps for each second of analysis of the structure.

4 METHODOLOGY

The dynamic response of the structure in the time domain is obtained through a numerical routine developed in FORTRAN and adapted from [3] to allow analysis of guyed TL metal lattice towers. The routine uses the direct integration method (DIM) of the equations of motion, explicit, using central finite differences. The application routine is performed in three stages, called: pre-processing, processing and post-processing.

In the pre-processing stage, the input data of the complete excerpt of TL are provided, including the towers, the insulator chains, conductor cables, lightning rods and guy-wires, with their discretizations informing the geometric properties of every element, the mechanical properties of the materials and the cable element selected to break.

In the second step, called processing, from the input data informed, the structure is analyzed, and the dynamic response in the time domain is obtained.

In the final step called post-processing, the results can be viewed, from the output file generated in the previous step, in terms of displacements at the top of the towers and effort in elements, in addition to the structure movement visualization over time of analysis with the aid of a post-processing program.

4.1 Cable breakage simulation

The total time set for the analysis of the complete excerpt of TL is 40 seconds. The own weight loads of the towers and the cable elements are applied gradually from 0 to 100% of their value for a 5 second time interval (t = 0 a 5 s). At this same time interval, the initial deformation is applied to give the desired prestressing to the guy-wires, also gradually. The time interval of 5 to 20 seconds is used to dampen any vibration induced in the numerical model. The rupture of the selected cable element occurs at time 20 seconds (t = 20 s) and the remaining 20 seconds of analyzes are spent in obtaining the structure dynamic response over the time. In Fig. 1 the conductor cable elements are shown (breakage hypothesis 1) and of lightning rod (breakage hypothesis 2) selected to break in the model

The cable elements which must break, both the conductor cable and the lightning rod, are situated near tower 2, the line span 2, between towers 1 and 2, specifically the element 2422 for the conductor cable 1 and the 2022 for the lightning rod 1, as shown in Fig. 1.

The definition of span 2 for the cable breakage, near tower 2 was made with the objective of minimizing the return of the conductor cables and lightning rods vibrations that can occur close to the extremity nodes of the model, when the cable breakage happens, since these nodes are defined as immovable. The tower 2 is the one that is farthest away from the extremities, and therefore, selected for the analysis of efforts and displacements.

The breakage of a cable in the TL full numerical model is introduced into routine nullifying the axial force ($f_a(t)$) acting on the selected element to break in a moment of time, specified, in the case of this analysis at t = 20 s.

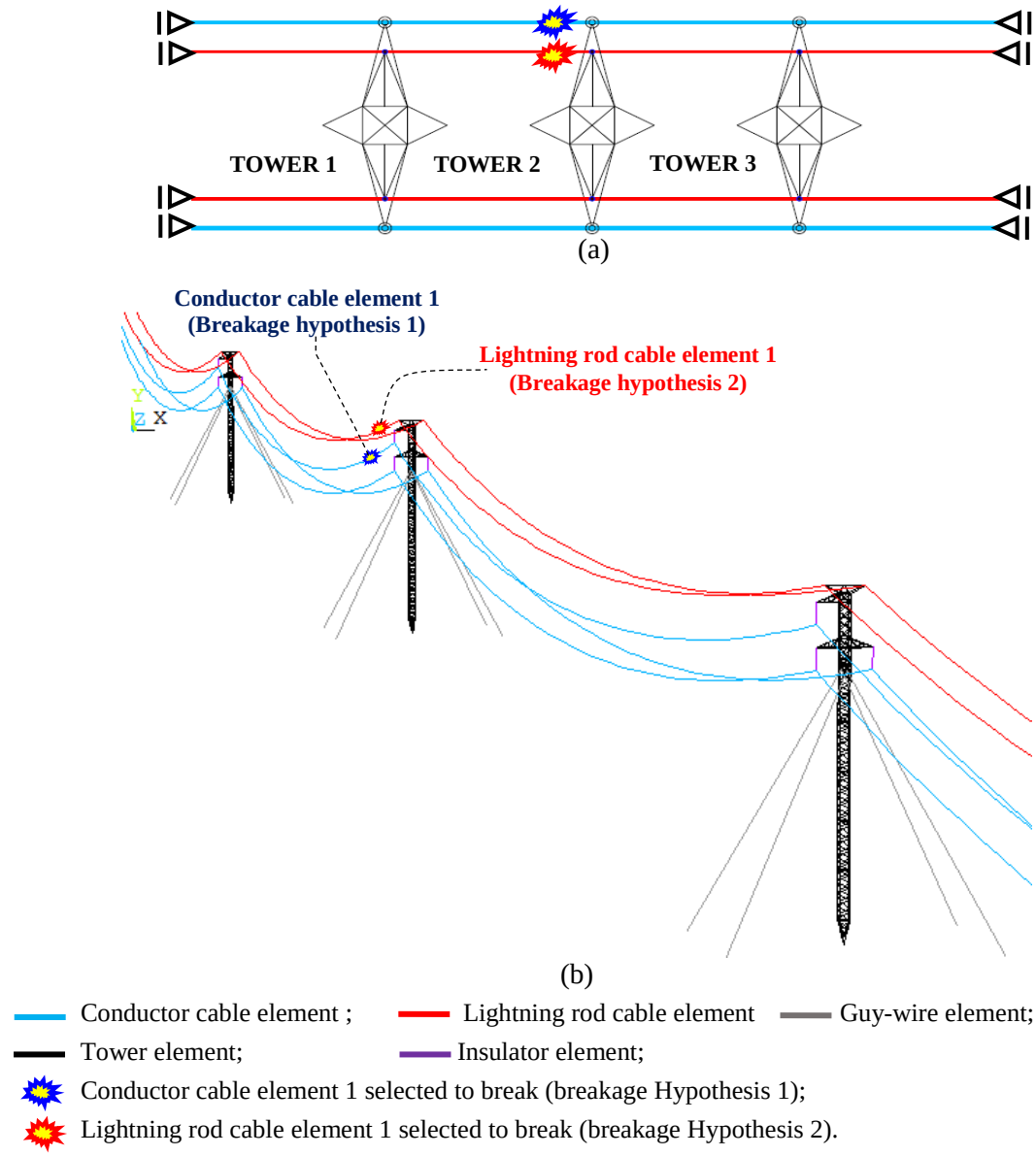


Figure 1: Selected items to break in the dynamic analysis: (a) top view of the full TL model (b) model view in 3D perspective.

For the viscous damping (proportional to the nodal mass) defined by the constant cm (constant that relates the damping with the nodal mass), present in the equations (6) and (7), the values $\mathbf{c_m = 3}$ and $\mathbf{c_m = 2}$ were adopted for the elements of the towers and insulators, respectively. For cable elements, where the damping has a more significant role in the dynamic response, the value of $\mathbf{c_m = 1}$ was adopted. The values of the constant $\mathbf{c_m}$, for each node of the structure, have been defined based on a parametric study by [7]. It should be noted that the constant $\mathbf{c_m}$ must be adjusted so that the critical damping ratio (ζ) is close to the one of the actual structure, in the case of metal lattice towers it varies from 3% to over 10%. Table 1 shows the main input data of the numerical routine for the excerpt in question.

Gravitational acceleration	9,81 m/s²
Constant $\mathbf{c_m}$ applied to the tower nodes	3,0
Constant $\mathbf{c_m}$ applied to the insulators nodes	2,0
Constant $\mathbf{c_m}$ applied to the cables nodes	1,0
Total analysis time	40 s
Time interval for application of the own weight and prestressing of stays	0 to 5 s
Time interval for the damping of induced vibration	5 to 20 s
Instant of time set for the cable breakage	20 s
Time interval for the structure response analyses	20 to 40 s
Total number of nodes in the model	1836
Total number of elements in the model	4012
Integration time intervals used in the analyses	2,5E-06 and 4,0E-06 s
Initial deformation in the stays 1 and 3 (Pair 1)	1,452840E-03 m/m
Initial deformation in the stays 2 and 4 (Pair 2)	1,452083E-03 m/m

Table 1: Basic input data in the routine in FORTRAN for the dynamic analysis by explicit direct integration in the full TL excerpt.

For the application of DIM, the developed program makes it possible to discretize the structure only with space truss elements, which is considered as a limitation, since lattice towers when modeled with only space truss elements may have an interior hypostatization. This problem can be circumvented with the use of dummy rods with an axial stiffness such as to avoid instability of the structure and to not significantly alter the results. The calibration and evaluation process of the dummy bar inserted in the numerical model is described in [8].

The evaluation of the response accuracy and stability of the numerical method is done in two ways, namely: comparison of the fundamental frequency of a model with only one tower, without cables or insulators, with the modal analysis of the same model and of the model dynamic response verification using two different time intervals of integration (Δt_1 and Δt_2), smaller than Δt_{crit}, which should result identical.

4.2 Evaluation of the method accuracy

To evaluate the accuracy of the numerical method, a model with only one tower, without cables or insulators, is subjected to an impulsive horizontal load on top, in order to obtain the horizontal displacement of the structure in the direction of the force applied over time analysis. Thus, it is possible to determine the frequency of vibration associated with this mode and compare it with that obtained in a modal analysis with the same model of isolated tower. This evaluation is done in two horizontal directions, X and Z.

The application of the impulsive load is performed as follows: the bearing of the structure own weight is applied gradually from, 0 to 100%, in a time interval from 0 to 5 seconds. At the same time, the initial deformation is imposed for the stays prestressing, also gradually. After that, at time t = 5 seconds, the application of horizontal load in the direction X starts, increasingly, during 5 seconds, reaching its maximum value in t = 10 s. Then, the load value is kept constant for 5 more seconds and stopped abruptly in t = 15 s, putting the structure in free vibration for 5 more seconds, until it reaches t = 20 s, that is the final time of this analysis. With the displacements at the top of the tower in the direction X, over time in the last 5 seconds of analysis it is possible to obtain the vibration frequency of the isolated tower associated with this mode. In order to obtain the frequency of vibration of the tower in the other horizontal direction the same procedure is performed, however, with the impulsive load applied in the direction Z.

Table 2 shows the results of vibration frequencies found in the modal analysis and the values found in the dynamic analysis by the DIM with impulsive loads.

Modal Form	Frequency (Hz)		Variation (%)
	Modal An. *	DIM**	
Bending in the direction X	2,739	2,611	4,57
Bending in the direction Z	3,662	3,649	0,19

* Frequencies obtained with modal analysis for the isolated tower model;

** Frequencies obtained with dynamic analysis by the DIM for the model with the isolated tower and impulsive forces.

Table 2: Comparison of vibration frequencies of guyed tower S1E2 in the directions X and Z, obtained in modal analysis and by the DIM with impulsive forces

From the results shown in Table 2 for the model with the isolated guyed tower, it can be seen that the DIM is quite accurate.

4.3 Evaluation of the method stability

To evaluate the direct integration method stability, two different integration time intervals were used (Δt_1 e Δt_2) for the same model of isolated tower, smaller than the critical time integration (Δt_{crit}) of the structure.

The values of the integration time interval (Δt), adopted in the analyses of the isolated tower, are shown in Table 3.

Δt_1	Δt_2
2,50E-06 s	4,00E-06 s

Table 3. integration time intervals adopted for the dynamic analysis of the isolated tower

The results of the model analyzed were identical for both Δt values used, both for the vibration frequencies in the direction X and in the direction Z, confirming the method stability.

5 NUMERICAL EXAMPLE

The model used as an example for application of the DIM consists of a complete excerpt of TL, consisting of three guyed towers, called S1E2, with their respective chains of insulators and stay cable, four spans of conductor cables and lightning rods with 500 m long, with a total excerpt of 2000 m. The towers are on the same level, that is, the cables suspension points are positioned at the same height. At the extremities of the TL excerpts (in x = 0 m and x = 2000 m), cables are fixed on fixed nodes, that is, with all degrees of freedom to the translation (X, Y, Z) restricted. The same occurs in the anchor points of the stays and at the feet of the towers. The numerical model consists of 1902 actual tower bar elements, 9 insulator elements, 1012 cable elements, (conductors, lightning rods and stays) and 1089 elements of fictitious tower bars. Thus, the complete excerpt results in a total of 1836 nodes and 4012 elements.

Fig. 2 shows the complete TL model excerpt analyzed, identifying the numbering of the three guyed towers and of the four cables spans.

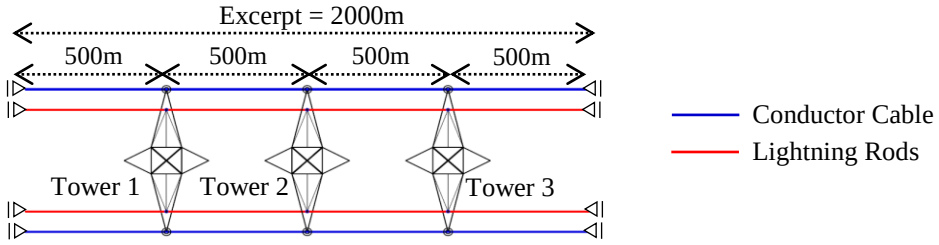


Figure 2: Top view of the TL excerpt analyzed, consists of three towers and four cables spans

The model is subjected to the breaking action of a conductor cable in a first analysis and in a second analysis to the breaking action of a lightning rod.

5.1 Towers and insulators chains

The guyed towers S1E2 are of suspension (tangent), single circuit 230 kV, with arrangement of the conductor cables in an asymmetric triangular shape and two lightning rods. They are metal lattice structures with a total height of 43.5 meters. The central mast has a square cross-section, with 130 x 130 cm. Each tower has four stay cables which ensure the stability of the structure, divided into two pairs, where each pair is fixed on the tower in symmetric points positioned between the trunk and the head of the tower. Each stay cable has a total length of 39.65 m. The silhouette and the spacing between the cable anchor points (foundations of the stay cables) and their connection points in the body of the tower are detailed in Figures 3 and 4.

The stays anchoring happen at 17.0 m from the central mast axis, in the direction X and 15.0 m in the direction Z (model global horizontal axis) as can be seen in Fig. 4.

The insulator chains in the numerical models adopted were considered with a single element, able to withstand only tensile stresses. They are suspended at the extremities of the arms of the conductor cables and are shaped with a length of 2.70 meters.

5.2 Conductor cables, lightning rods and stay cables

Regarding the mechanical performance of the conductor cables and lightning rods used in TL, linear and nonlinear stress-strain relationships can be used. For non-linear relationships, the expressions are given by a polynomial of 4th degree, similar to those employed by [9, 10]. In this work, the behavior of the cables was considered to be linear. Thus, the stress-strain diagram of the conductor cables and lightning rods for a constant temperature is a straight line. For the stay cables the behavior was also considered to be linear.

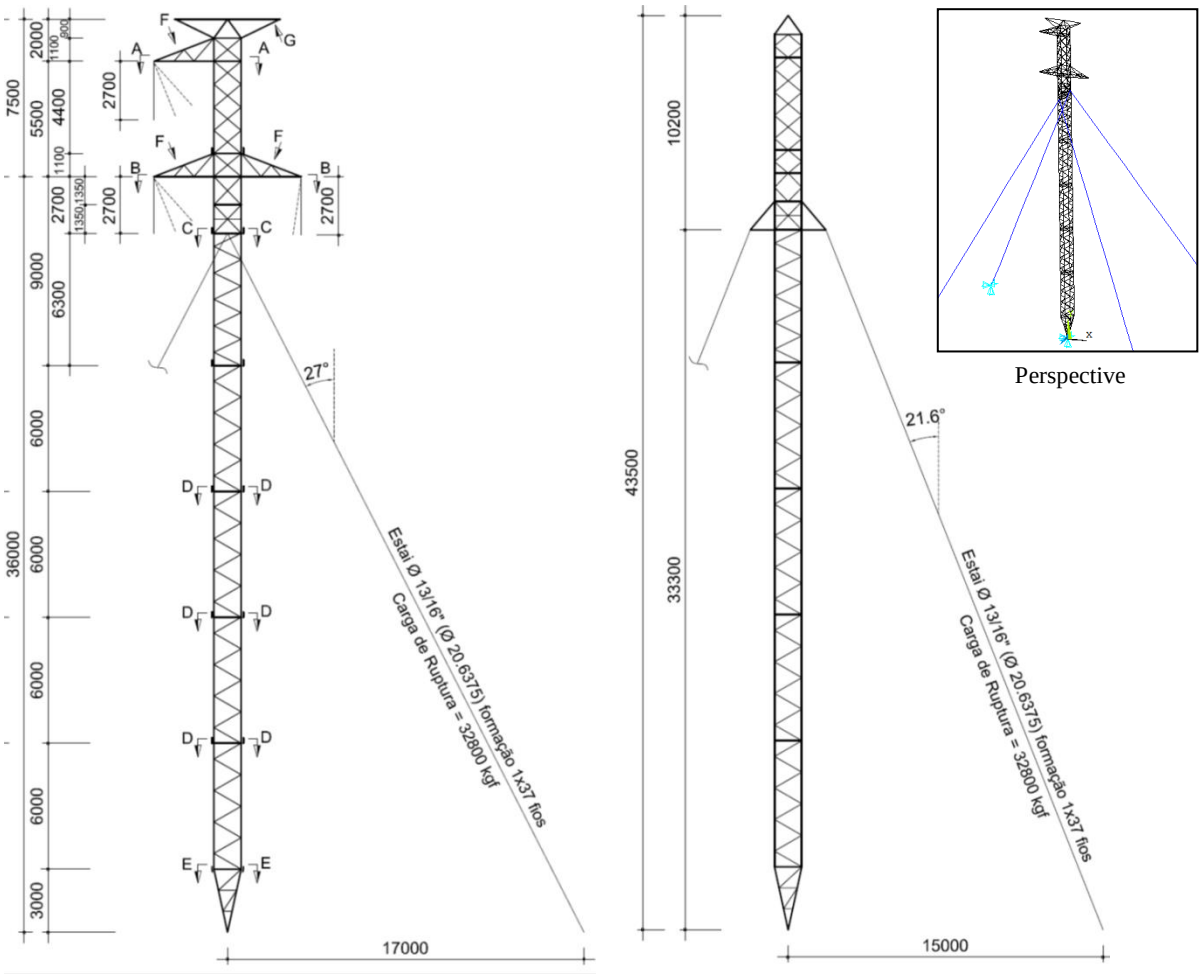


Figure 3: Front, side and perspective view of the guyed tower S1E2 with the structural configuration of the elements and other geometric details (dimensions in millimeters)

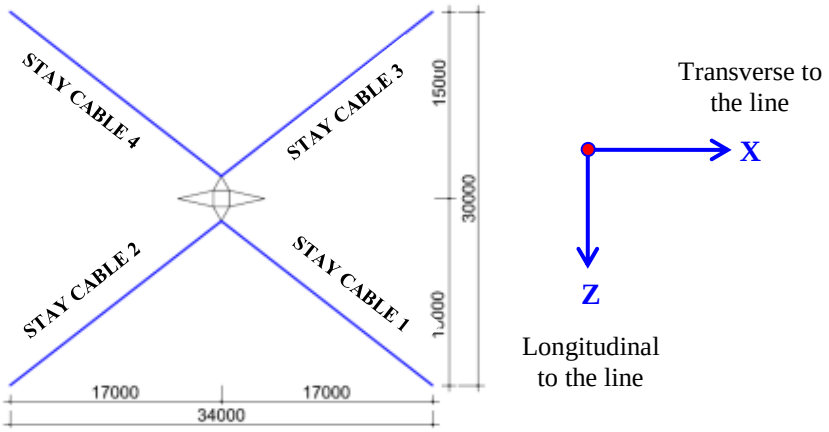


Figure 4: Top view of the guyed tower S1E2, with the distances between the anchoring points of the four stays (dimensions in millimeters).

The conductor cables used in the model are made of aluminum steel-reinforced IBIS type (*CAA/ACSR – Aluminum Conductor Steel Reinforced*) consisting of 26 aluminum wires in the outer layer and beam with 7 steel wires (26/7 cables). The lightning rods used are of steel type Extra High Strength (*EHS - Extra High Strength*) with seven 3/8 inch steel cables. For the stays EHS wire ropes of 37 wires13/16" are used. The mechanical properties of the cables used are detailed in Table 4.

Properties	Conductor cables	Lightning rods	Stay cables
External diameter	19,89 mm	9,14 mm	20,20 mm
Cross-sectional area	234,00 mm²	51,08 mm²	320,00 mm²
Linear weight	7,98 N/m	3,98 N/m	19,42 N/m
Modulus of elasticity	74515 N/mm²	172369 N/mm²	120000 N/mm²
Ultimate tensile strength (T_{rup}) - UTS	72506,0 N	68502,6 N	328000,0 N

Table 4: Properties of IBIS type conductor cables, AAC/ACSR (aluminum cable with steel core), 26/7 wires, lightning rod of EHS type (7-wire) 3/8" and stay cables EHS wire rope type (37 wires) 13/16"

In order that the conductor cables and suspended lightning rods are properly modeled within a TL span, the distorted characteristic of the cables should be taken into account, which in turn has the format of a catenary. The maximum deflection that the cables will reach when subjected to own weight loads and the initial tension should also be considered. The configuration of the catenary, in relation to the center of the wire span (central axis) is given in two ways: symmetrical or asymmetrical. The difference depends on the height of the cable suspension points. The catenary is considered symmetrical when the suspension points are given in the same height, where the vertex is located (center of the span), which is the point where the maximum deflection (f_e) occurs. When the suspension supports are in unequal heights, then the catenary is asymmetrical and the maximum deflection does not occur at the center of

the span, as illustrated in Fig. 5. The deflection is due to the span length, the temperature and the tensile level applied to the cable when installed in the TL suspension points.

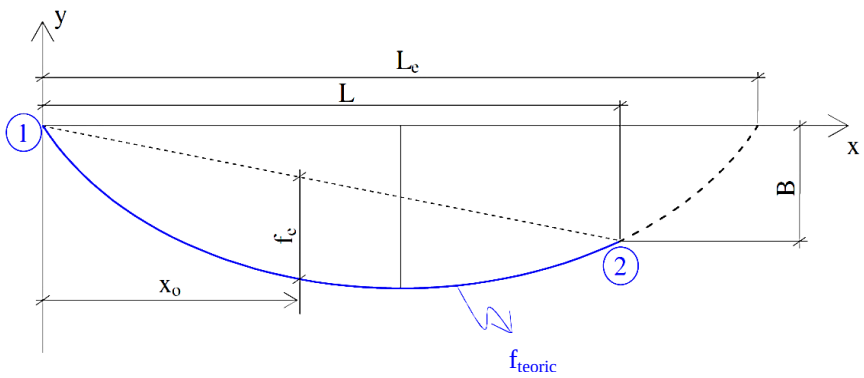


Figure 5: Cable suspended between the supports "1" and "2" with different heights [2]

The suspension points were considered all of the same height (B = 0), that is, all catenaries are symmetrical.

In the case of the suspended conductor cables in transmission lines, their catenary, in the EDS condition (Every Day Stress), is designed for an initial design tension around of 20% of its ultimate capacity of tension (UTS - Ultimate Tension Stress). This value is recommended by [11]. For the lightning rods, the value adopted for the design tensile strength, in the theoretical position, must be calculated for a theoretical catenary deflection "f_{ep}" equal to 90% of the maximum deflection "f_e" of the conductor cables in the span under consideration. It is noteworthy that, according to the [11], the recommended value for the design initial tension in the lightning rods is of the order of 14% of UTS condition.

At the initial time of the analysis ($t = 0$), the weight force is not taken into account in the initial coordinates of the catenary points. Thus, there are two catenaries, one without considering the weight force (own weight), called initial catenary, and another after the application of own weight, called theoretical catenary. In this condition, after the application of the weight force in the structure, the cable must be positioned so that it is subjected to a design tensile strength (T_p), equivalent to a percentage of breaking strength in the cable traction (T_{rup}), with the theoretical catenary ($f_{teórica}$) and maximum deflection (f_e). The value for the design tensile strength of the conductor cables and lightning rods used in the study was calculated by Equations (12) and (13), respectively:

$$T_p = 0.20 \cdot T_{rup} \quad \text{(20\% for the conductor cables)} \tag{12}$$

$$T_{pp} = 0.1175 \cdot T_{rup} \quad \text{(11.75\% for the lightning rods)} \tag{13}$$

The catenary of the conductor cables IBIS AAC/ACSR 26/7 and lightning rods EHS 3/8" adopted in the project (Table 4) in the initial condition (before application of weight force) and theoretical (after the application of weight force) considering a span L = 500 meters and with B = 0, were determined considering the formulations presented in [8]. Additional details can be found in [12]. The results of the catenaries in the theoretical and initial conditions of the conductor cables and lightning rods are shown in Figures 6 and 7, respectively.

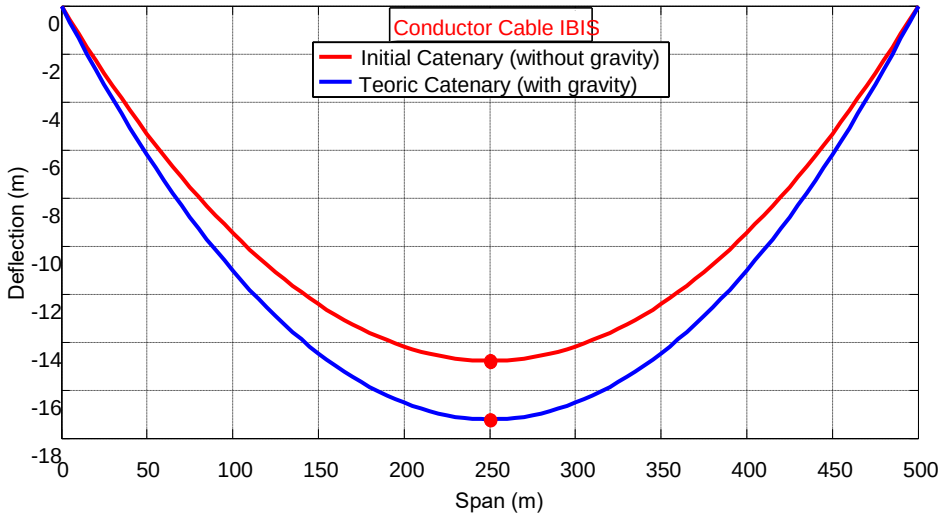


Figure 6: Catenary in initial and theoretical condition for a conductor cable IBIS AAC/ACSR 26/7 considering a span of 500 m between towers and gap between the cables suspension points equal to zero (B = 0)

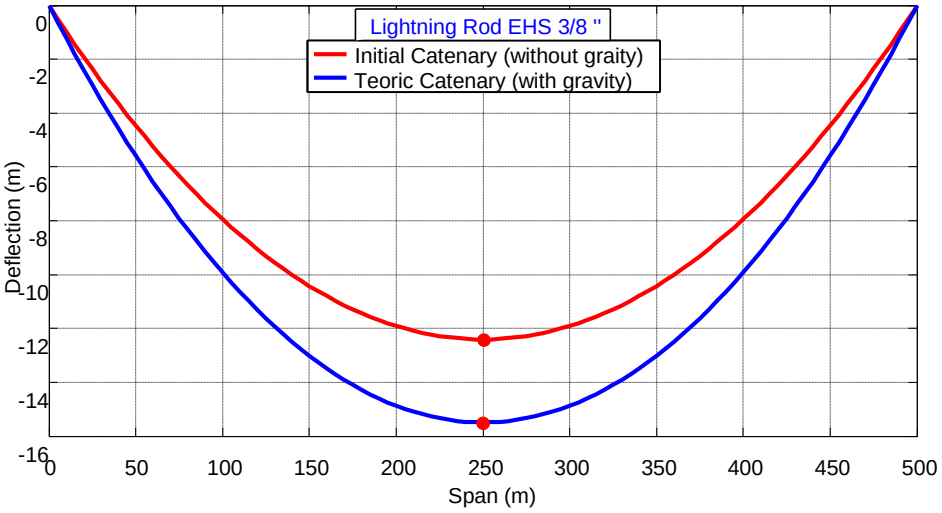


Figure 7: Catenary in initial and theoretical condition for a lightning rod EHS 3/8 ", considering a span of 500 m between towers and gap between the cables suspension points equal to zero (B = 0)

5.3 Stay cables prestressing

The initial prestressing that must be applied to the stay cables of the S1E2 tower, responsible for ensuring the stability of the structure central mast, produces a tensile strength in these cables of approximately 13% of its breaking load, according to recommendations of the [13], which suggests values between 8 and 15% of the cable rated load capacity. In fact, the initial prestressing in the stay cables, applied through a deformation in the cable elements is of the order of 17% of its breaking load, and this value is reduced to 13% after the application of the tower own weight, of the conductor cable and lightning rods and their own prestressing, which causes the central mast to shorten, thereby reducing the initial prestressing.

In the numerical routine, the initial prestressing in stay cable is applied through a constant initial deformation which is added to the deformation of the element when calculating the axial force in the cable element. As the tower has an asymmetrical silhouette, its structure naturally has the center of gravity displaced from the central axis of the mast. Thus, the appli-

cation of a different prestressing on each side, or each pair of stays is necessary, in order to keep the structure in its upright position.

To determine the amount of prestressing to be applied, a calibration process was done, where the first initial deformation value adopted for the stays was that capable of producing prestressing of 17% of the breaking load on the cables, applied equally on the four stays of the tower. The calibration was made by changing the values of the initial deformation, and consequently of the prestressing in the stay pairs, and by monitoring the horizontal displacement in the direction X of the nodes of the stays arms tip, which should be close to zero, in order to maintain the structure of the central mast vertical. This process was carried out considering only the central mast own weight, so the structure construction process could be represented accurately. The own weight of conductor cables, insulators and lightning rods were applied after calibration of stays. So, as expected, when the own weight of the conductor cables and lightning rods were applied a relief occurred in the prestressing of the stay cables. The relief is generated because of the central mast shortening caused by the own weight of the structure, conductor cables and lightning rods and by the application of prestressing. The values obtained for the initial deformation applied to each pair of stays, so as to ensure that the central mast remains vertical, are shown in Table 5.

Element	Initial Deformation
Pair 1 (stays 1 and 3)	1,452840E-03 m/m
Pair 2 (stays 2 and 4)	1,452083E-03 m/m

Table 5: Initial deformation values applied to the pairs of stays

5.4 Analysis Results

The results obtained for tower 2 analyzed in the numerical routine and subjected to conductor cable and lightning rod breakage in terms of displacement and normal stresses in the bars, are presented below. The results of a static analysis of the isolated model of the same tower, which uses the "equivalent static loads" in the cable breakage simulation, is used as a comparator in terms of maximum displacements at the top and normal stresses on the bars.

5.4.1 Displacements

The results for the displacements at the top of the tower 2 resulting from the dynamics of the conductor cable and the lightning rod breakage in the longitudinal direction of the line (Z) are depicted in Fig. 8 along with the respective displacements obtained in static analysis.

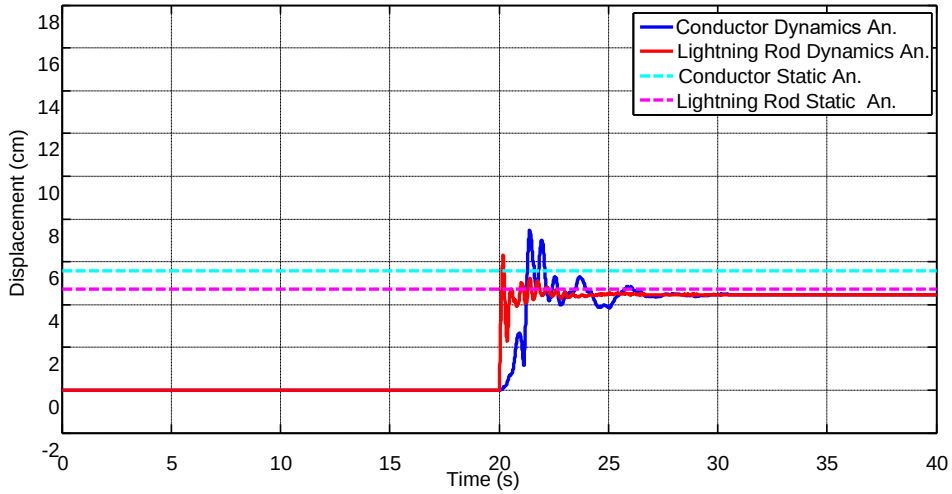


Figure 8: Longitudinal displacements in the node 251 of Tower 2 for the two cases of breakage (conductor cable 1 and lightning rod 1)

5.4.2 Stress in the elements

The normal stresses in stays 1 and 3, for the pair 1 of Tower 2 (adjacent to the breaking point) of the TL excerpt, for the conductor cable breakage hypothesis, are shown in Fig. 9.

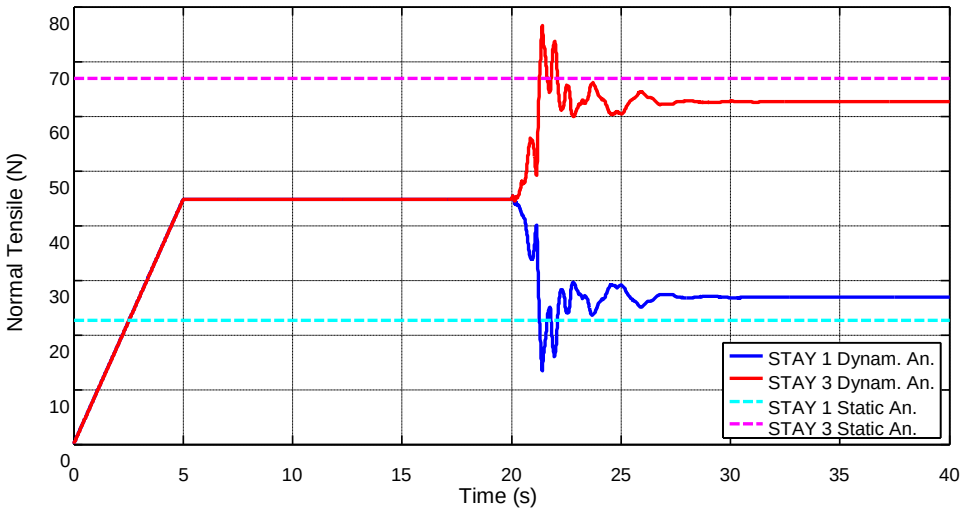


Figure 9: Normal tensile stress in stays 1 and 3 (pair 1) of tower 2 for the conductor cable 1 breakage hypothesis

For the stays 2 and 4, relating to the *pair* 2 of the tower 2 in the TL excerpt, for the conductor cable breakage hypothesis, the values of normal stresses are shown in Fig. 10.

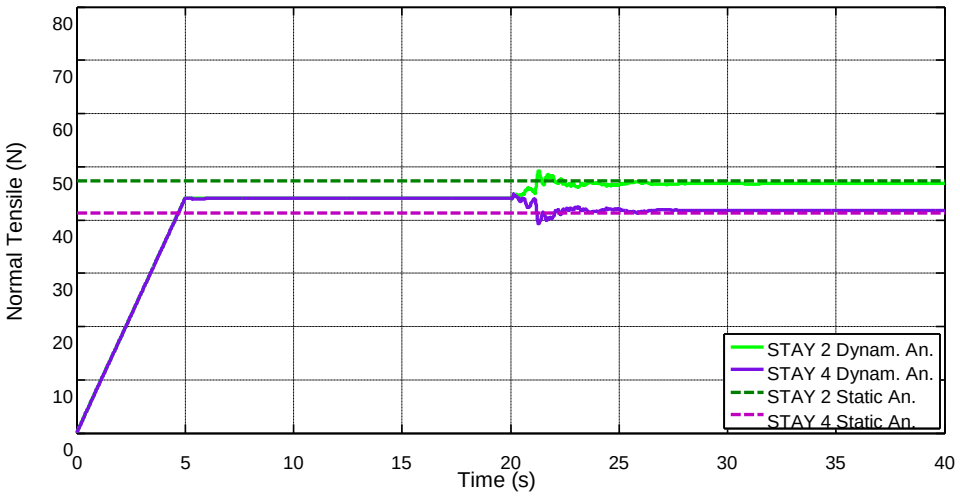


Figure 10: Normal tensile stress in stays 2 and 4 (pair 2) of Tower 2 for the conductor cable 1 breakage hypothesis

Fig. 11 shows the normal stresses on the main leg 173 of the tower 2 positioned below the fixing supports of stays, for the conductor cable breakage hypothesis.

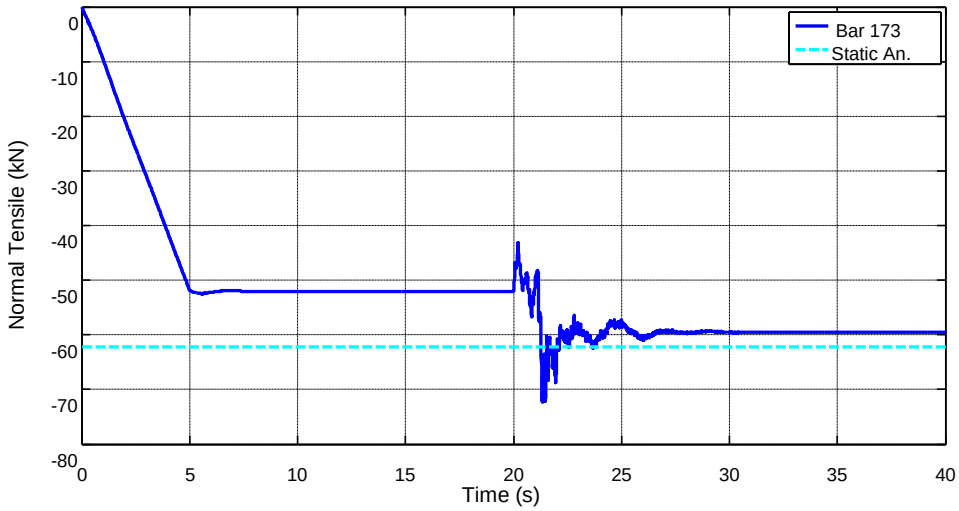


Figure 11: Normal stress on the main leg 173 of tower 2 for the conductor cable 1 breakage hypothesis

In Fig. 12 the stresses of the diagonal bracing 483 of the tower 2 are given, which is situated in the region of the most significant stresses for this type of bar (above the stays fixing corbels) for the conductor cable breakage hypothesis.

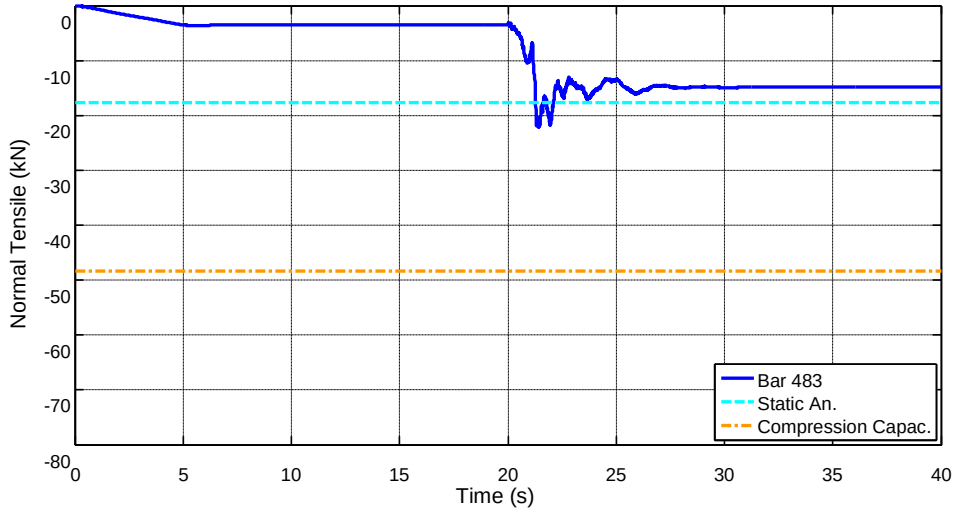


Figure 12: Normal stress in the diagonal bracing 483 of tower 2 for the conductor cable 1 breakage hypothesis

In Figures 13 and 14, the stresses on the bars located in the upper arm that supports the conductor cable which breaks, in tower 2 are illustrated.

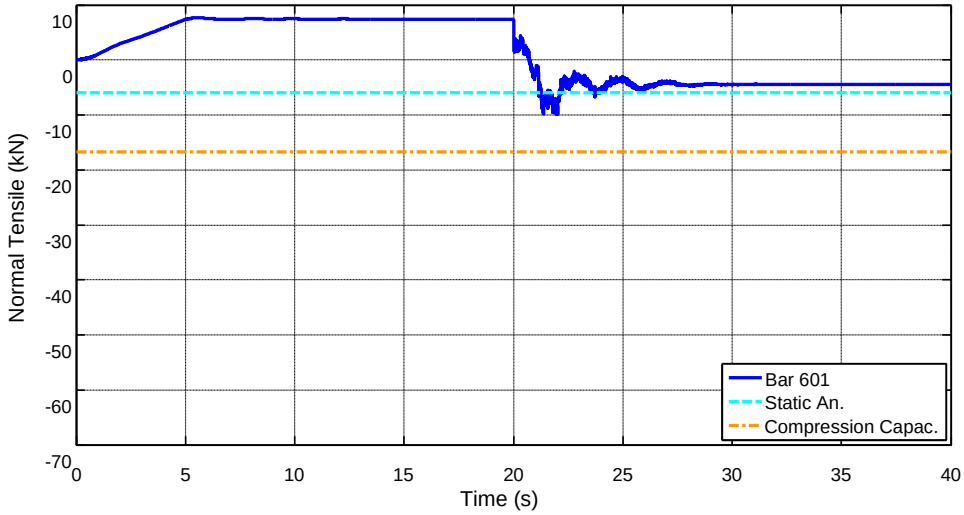


Figure 13: Normal stress in the bar 601 of the arm of conductor cable 1 in tower 2, for the conductor cable 1 breakage hypothesis.

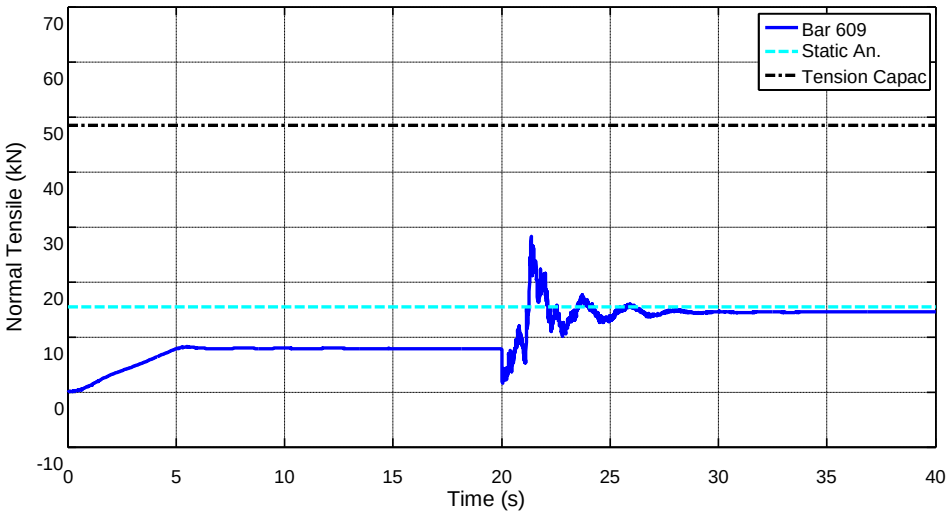


Figure 14. Normal stress in the bar 609 of the arm of conductor cable 1 in tower 2, for the conductor cable 1 breakage hypothesis

The correct functioning of the routine during the gradual application of own weight loads and prestressing on the stay cables (t = 0 to 5 s), time to damp possible vibrations (t = 5 20 s), the instant of breakage of the selected cable element (t = 20 s) and finally, during the interval for obtaining the dynamic response of the structure at the time (t = 20 to 40 s) can be observed by the results presented.

5.4.3 Post-processing

After the results of the dynamic analysis of the complete TL excerpt are obtained, the results output files generated in the routine in FORTRAN in terms of nodal displacements over the time of analysis, can be seen in post-processing programs, making it possible to see the movement of the structure over time. Fig. 15 shows the nodal coordinates of the complete TL excerpt for all time instants determined along the analysis (0 to 40 seconds).

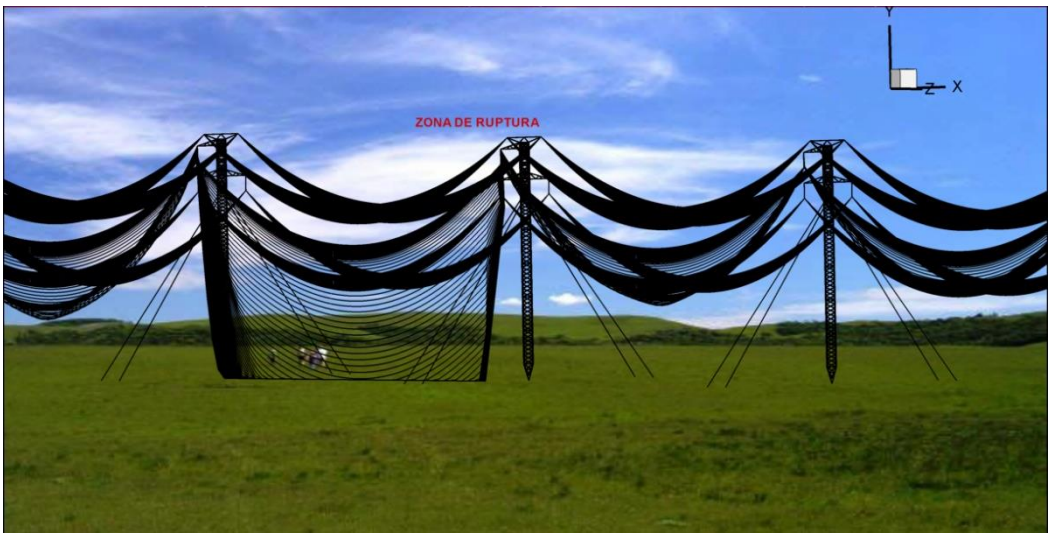


Figure 15: Animation tables of the numerical model of the complete TL excerpt, submitted to conductor cable breakage (illustration background)

From the set of nodal coordinates of all the TL excerpt, generated at certain instants of time, for the conductor cable breakage hypothesis, the post-processing tool also allows the development of video simulation, from start to finish of analysis (0 to 40 s), where the behavior of the tower adjacent to the breaking point can be observed, as well as the deflection variations in the cables adjacent spans.

6 CONCLUSIONS

This work presents a computational routine developed in FORTRAN language able to generate and analyze a complete segment of a TL, consisting of metal guyed towers and including all components subjected to dynamic actions.

The results showed that the routine is very efficient and easy to use, and can become an excellent tool for designers to obtain the dynamic response of TL structures, in order to verify the obtained sizing with a conventional static analysis, especially when new types of towers and cable arrangement is proposed.

In validation processes by vibration frequencies comparison and normal stresses on the elements by impulsive loads, it was found that the numeric routine used for the dynamic solution for the cables breakage problem, using the DIM, is accurate for the analysis, once these processes converged. Moreover, the accuracy of the results was also tested using two different integration times (Δt_1 and Δt_2), which indicated the stability of direct integration method.

Because of the large amount of elements (4012) in the dynamic model of the TL excerpt, with an integration interval (Δt) of 4,0E-06 s, the time invested in the processing of dynamic cable breakage analysis was relatively long, a bit over seven hours, considering a PC with Intel Core i5 and 4GB of RAM processor.

One should take into account the simplifications considered in programming and that their responses come from an approximate numerical solution, since this is an issue with many variables to be further investigated. It is necessary for the program development to be continued so that it can be applied in a more complete program that mainly allows the use of gantry elements plan, simplifying and increasing the accuracy of the models in analysis.

Finally, it is essential that more studies on this subject are conducted with models increasingly realistic and considering other types of guyed towers, besides conducting experimental tests to confirm the accuracy of the results.

7 ACKNOWLEDGMENTS

The authors acknowledge the financial support of CAPES, Brazil.

REFERENCES

[1] Menezes, R. C. R.; Rocha, M. M.; Riera, J. D. The use of explicit methods in the analysis of transmission line towers. LDEC, UFRGS, 1998.

[2] Kaminski Jr., J.; Miguel, L.F.F.; Menezes, R.C.R.; Miguel, L.F.F.; Martins, E. G. *Rotina computacional para análise dinâmica de estruturas de linhas de transmissão*. XIII Encontro Regional Iberoamericano do CIGRÉ (XIII ERIAC). Puerto Iguazú, Argentina. 2009.

[3] Bathe, K. J. *Finite element procedures in engineering analysis*. Englewood Cliffs, New Jersey: Prentice-Hall, 1996.

[4] Humar, J. L. *Dynamics of structures*. 2nd edition. Leiden: A. A. Balkema Publishers, 2005.

[5] Groehs, A. G. *Mecânica vibratória*. São Leopoldo, RS: Editora Unisinos, 2a edição, 2005.

[6] Kaminski Jr., J. *Incertezas de modelo na análise de torres metálicas treliçadas de linhas de transmissão*. 2007. Tese de Doutorado em Engenharia Civil. Universidade Federal do Rio Grande do Sul. Escola de Engenharia. 2007.

[7] Kaminski Jr. J, Miguel L.F.F., Menezes R.C.R. *Aspectos relevantes na análise dinâmica de torres de LT submetidas à ruptura de cabos*. Seminário Nacional de Produção e Transmissão de Energia Elétrica. Curitiba, Brasil. 2005.

[8] Carlos, T. B. *Análise dinâmica de torres estaiadas de linhas de transmissão submetidas à ruptura de cabo*. Dissertação de Mestrado em Engenharia Civil. Universidade Federal de Santa Maria, Centro de Tecnologia. 2015.

[9] Thrash, R. *Overhead conductor manual*. Southwire Company, One Southwire Drive. Caroolton, GA, 1994.

[10] Electric Power Research Institute (EPRI). TLOP/SAGT manual, *TL Workstation code*, EL-4540, v. 6, Palo Alto, CA, 1988.

[11] Associação Brasileira de Normas Técnicas (ABNT). *NBR 5422: Projeto de linhas aéreas de transmissão de energia elétrica*. Rio de Janeiro, 1985.

[12] Irvine, H. M., Caughey, T. K. *The linear theory of free vibrations of a suspended cable*. Proceedings of the Royal Society of London, Vol. A341, 299-315. Londres, Inglaterra, 1974.

[13] CSA-S37-01. Antennas, towers, and antenna-supporting structures. Canadá, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

INVARIANCE OF EIGENFREQUENCIES AND EIGENMODES UNDER GEOMETRIC TRANSFORMATION IN ELONGATED ELASTIC STRUCTURES

Maryam Morvaridi[1], **Michele Brun**[2]

[1]Dipartimento di Ingegneria Civile ed Architettura, Università di Cagliari
address
Piazza d'Armi, 09123 Cagliari, Italy morvaridi@unica.it

[2] Dipartimento di Ingegneria Meccanica, Chimica e dei Materiali, Università di Cagliari,
address
Piazza d'Armi, 09123 Cagliari, Italy mbrun@unica.it

Keywords: Flexural Waves, Transformation Optics, Cloaking, Harmonic Analysis, Eigenfrequencies.

Abstract. *The problem of transformation optics for longitudinal and flexural waves in monodimensional elastic systems is analyzed. System of finite dimensions are considered and it is shown that, under appropriate interface conditions, eigenfrequencies in finite systems remain unchanged while eigenmodes can be tuned depending on the applied geometric transformation. Eigenfrequency analysis can be used in cloaking problem in order to demonstrate the quality of the cloak.*

1 INTRODUCTION

Transformation optics has been applied for the design of cloaking models for acoustic and electromagnetic waves [1, 2, 3, 4, 5, 6, 7, 8, 9].

Cloaking for elastic waves brings new challenges regarding the physical interpretation of equations and boundary conditions in the cloaking region. In particular, the governing equations for elastic waves in the vector problem of elasticity and for flexural waves governed by fourth-order differential equations, are not invariant under transformation. The vector problem of elasticity have been analyzed in [10, 11, 12, 13] and it has been shown that in order to enforce invariance of transformed equation, polar materials involving non symmetric Cauchy stress are needed.

The theoretical analysis, the design and the physical interpretation for the cloaking of flexural waves has been addressed in [14, 15, 16, 17, 18, 19]. In particular, [17, 18] have shown that transformed equation for Kirchhoff plates corresponds to an inhomogeneous anisotropic material in presence of pre-stress. The results has also been obtained following an asymptotic procedure from three-dimensional vector elasticity. In [19] the effect of internal boundary conditions for nearly perfect cloak has been detailed.

In general, the quality of the cloak is addressed qualitatively by implementing numerically or experimentally a plane wave, or a wave generated by a point source in an infinite medium, and checking on the shadow generated by a cloaked defect. In [20] the quantitative scattering measure has been introduce in order to evaluate the effectiveness of the cloak.

Here, we suggest the analysis of eigenfrequencies and eigenmodes as an advanced and simplified tools in order to demonstrate the quality of the cloaking and of the transformation. Such a vision drastically simplify the experimental and numerical implementation avoiding the necessity to introduce Perfectly Matched Layers, which can only been obtained approximatively for flexural waves. In particular, we consider a simple monodimensional Euler-Bernoulli beam subjected to longitudinal and flexural waves and we demonstrate that under transformation eigenfrequencies remain unchanged and eigenmodes can be perturbed simply changing the adopted transformation. Again, two possible quantitative measures can be considered in order to check the quality of the effect. The perturbation of the field in the untransformed domain as in the scattering measure or the values of the eigenfrequencies after transformation.

Also, the analysis within a finite domain is important because it enhances the contribution of the non-propagating part of the waves which is often disregarded in cloaking problems. Flexural waves are governed by fourth-order differential equations, ordinary for beam and partial for plates. The solution is the superposition of propagating waves, solution of Helmholtz equation and non propagating waves, solution of the modified Helmholtz equation [21, 22, 23, 24, 25]. The boundary conditions in finite system couple the solutions of the Helmholtz and modified Helmholtz equations.

The paper is organized as follows, In Section 2 we report the time-harmonic equations of motion for longitudinal and flexural waves, in Section 3 we report the transformed equations and we discuss the conditions for the automatic satisfaction of interface conditions. In Section 4 we report the anlytical and numerical eigenfrequency analysis for longitudinal waves in a rod and flexural waves in a beam. Different transformations are discussed and detailed.

2 EQUATIONS OF MOTION IN ELONGATED BEAM STRUCTURES

We apply the transformation to one-dimensional elongated structures. We restrict the attention to time-harmonic regime. We consider slender structure such that the behavior can be

described within the classical Euler-Bernoulli beam model. The longitudinal displacement is $U(X)$ and the transverse displacements are $V(X)$ and $W(X)$. The structure has Young's modulus E, density ρ, cross sectional area A, second moments of inertia J_Y and J_Z.

Time-harmonic equation of motion for longitudinal waves in a rod is the following Helmholtz equation for the longitudinal displacement $U(X)$

$$[EA\,U'(X)]' + \rho A\,\omega^2 U(X) = 0, \tag{1}$$

where ω is the radian frequency.

The time-harmonic equations of motion for flexural waves are the following equations for the transverse displacements $V(X)$ and $W(X)$

$$\begin{aligned}[EJ_Z\,V''(X)]'' - \rho A\,\omega^2 V(X) = -T_Y'(X) - \rho A\,\omega^2 V(X) = 0,\\ [EJ_Y\,W''(X)]'' - \rho A\,\omega^2 W(X) = -T_Z'(X) - \rho A\,\omega^2 W(X) = 0.\end{aligned} \tag{2}$$

Note that the transverse shears are given by $T_Y(X) = -\,[EJ_Z\,V''(X)]'$ for the component along Y and $T_Z(X) = -\,[EJ_Y\,W''(X)]'$ for the shear component along Z.

3 TRANSFORMED EQUATION

In order to transform the domains we apply a coordinate transformation $x = G(X)$, having inverse $X = g(x)$. Capital letters indicate quantities defined in the untransformed domain, lower case letters refer to quantities in the transformed domain. Implementation of coordinate transformation within the equations of motion (1) and (2) lead to the transformed equations of motion.

The transformed equation of motion for longitudinal waves has the form

$$\left[\overline{EA}u'(x)\right]' + \overline{\rho A}\,\omega^2 u(x) = 0. \tag{3}$$

In Eq. (3) $u(x)$ is the transformed longitudinal displacement. We assume $u(x) = U(X)$. The transformed Eq. (3) involves non-homogeneous longitudinal stiffness $\overline{EA} = EA/g'(x)$ and non-homogeneous linear density $\overline{\rho A} = g'(x)\rho A$. The axial force $N(X) = EAU'(X)$ transforms to $n(x) = \overline{EA}u'(x)$, which results to be equal to $N(X)$.

The transformed equations of motion for flexural waves have the form

$$\begin{aligned}[t_y(x) + EJ_Z\,n(x)v'(x)]' + \overline{\rho A}(x)\,\omega^2 v(x) = 0,\\ [t_z(x) + EJ_Y\,n(x)w'(x)]' + \overline{\rho A}(x)\,\omega^2 w(x) = 0.\end{aligned} \tag{4}$$

In Eqs. (4) $v(x)$ and $w(x)$ are the transformed transverse displacements, that we assume such that $v(x) = V(X)$ and $w(x) = W(X)$. The transformed shear forces and axial force have the form

$$\begin{aligned}t_y(x) = m_z'(x) = -\left[\overline{EJ_Z}(x)v''(x)\right]',\\ t_z(x) = m_y'(x) = -\left[\overline{EJ_Y}(x)w''(x)\right]',\\ n(x) = \frac{3(g''(x))^2 - g'''(x)g'(x)}{(g'(x))^5},\end{aligned} \tag{5}$$

respectively. The transformed bending moments are $m_z(x) = -\overline{EJ_Z}(x)v''(x)$ and $m_y(x) = -\overline{EJ_Y}(x)w''(x)$. Eqs. (4) involve non-homogeneous bending stiffnesses $\overline{EJ_Z} = EJ_Z/(g'(x))^3$ and $\overline{EJ_Y} = EJ_Y/(g'(x))^3$ and non-homogeneous linear density $\overline{\rho A} = g'(x)\rho A$.

3.1 Interface boundary conditions

Transformation affects also boundary conditions, which have to be checked in order to avoid any perturbation of the fields corresponding to the homogeneous problems.

At interface points X_i between untransformed and transformed domains the condition $X_i = g(x_i) = x_i$ assures that the same interface point is shared between the different domains.

Concerning longitudinal waves the essential interface condition on displacement $U(X_i) = u(x_i)$ and the natural condition on axial force $N(X_i) = n(x_i)$ are automatically satisfied.

In relation to flexural waves essential interface conditions $V(X_i) = v(x_i)$ and $V'(X_i) = v'(x_i)$ are automatically satisfied if $g'(x_i) = 1$; while natural conditions $M_Z(X_i) = m_z(x_i)$ and $T_Y(X_i) = r_y(x_i)$, with M_Z the bending moment in the untransformed domain and $r_y(x) = t_y(x) + EJ_Z\, n(x)v'(x)$, require the additional conditions $g''(x_i) = 0$ and $g'''(x_i) = 0$ in order to be automatically satisfied. Therefore, in every point x_j in which we want to identify the condition in the transformed domain with a condition in the untransformed domain in X_j the relations $g(x_j) = X_j$ plus $g'(x_j) = 1$, $g''(x_j) = 0$, $g'''(x_j) = 0$ have to be imposed to the transformation. The same conditions hold for the flexural waves in the other direction involving displacement $w(x)$, rotation $w'(x)$, bending moment $m_y(x)$ and force component along z, $r_z(x)$.

4 EIGENFREQUENCY ANALYSIS

In this Section we compute eigenfrequencies and eigenmodes for longitudinal and flexural waves in a domain in which we introduce a transformation, and we compare the results with the eigenfrequencies and eigenmodes in a homogeneous system in absence of the transformation.

4.1 Longitudinal waves

We consider an homogenous rod of length $2L$ fixed at his ends. The solution of the Helmholtz equation (1) is

$$U(X) = A_1 e^{i\alpha X} + A_2 e^{-i\alpha X}, \tag{6}$$

where $\alpha = \omega\sqrt{E/\rho}$.

For fixed-fixed boundary conditions $U(\pm L) = 0$ the eigenfrequency need to satisfy the condition $\sin(2\alpha L) = 0$ leading to the well-known results $\omega = \sqrt{\rho/E}(n\pi)/(2L)$, with n positive integer. The corresponding eigenmodes are $\sin[n\pi(X + L)/(2L)]$.

Now, we introduce a second structure having the same homogeneous properties for $-L \leq X \leq 0$, while the right half $0 \leq X \leq L$ is transformed into the domain $0 \leq x \leq l$ by generic transformation $G(X)$ with inverse $g(x)$. The transformation $g(x)$ has to satisfy the conditions $g(0) = 0$ and $g(l) = L$. The problem is solved by

$$\begin{cases} U(X) = B_1 e^{i\alpha X} + B_2 e^{-i\alpha X}, & \text{for } -L \leq X \leq 0, \\ u(x) = B_3 e^{i\alpha g(x)} + B_4 e^{-i\alpha g(x)}, & \text{for } 0 \leq x \leq l. \end{cases} \tag{7}$$

The solution is found by satysfying the boundary conditions

$$U(-L) = u(l) = 0 \tag{8}$$

and the interface conditions

$$U(0) = u(0), \qquad EA\, U'(0) = \overline{EA}(0)u'(0). \tag{9}$$

The system of boundary and interface conditions has the form

$$\begin{bmatrix} e^{-i\alpha L} & e^{i\alpha L} & 0 & 0 \\ 1 & 1 & -1 & -1 \\ i\alpha & -i\alpha & -i\alpha & i\alpha \\ 0 & 0 & e^{i\alpha g(l)} & e^{-i\alpha g(l)} \end{bmatrix} \begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{10}$$

The condition of determinant equal to zero for the matrix in Eq. (10) is

$$-4\alpha \sin(2\alpha L) = 0, \tag{11}$$

which gives exactly the same eigenfrequencies as in the homogeneous case.

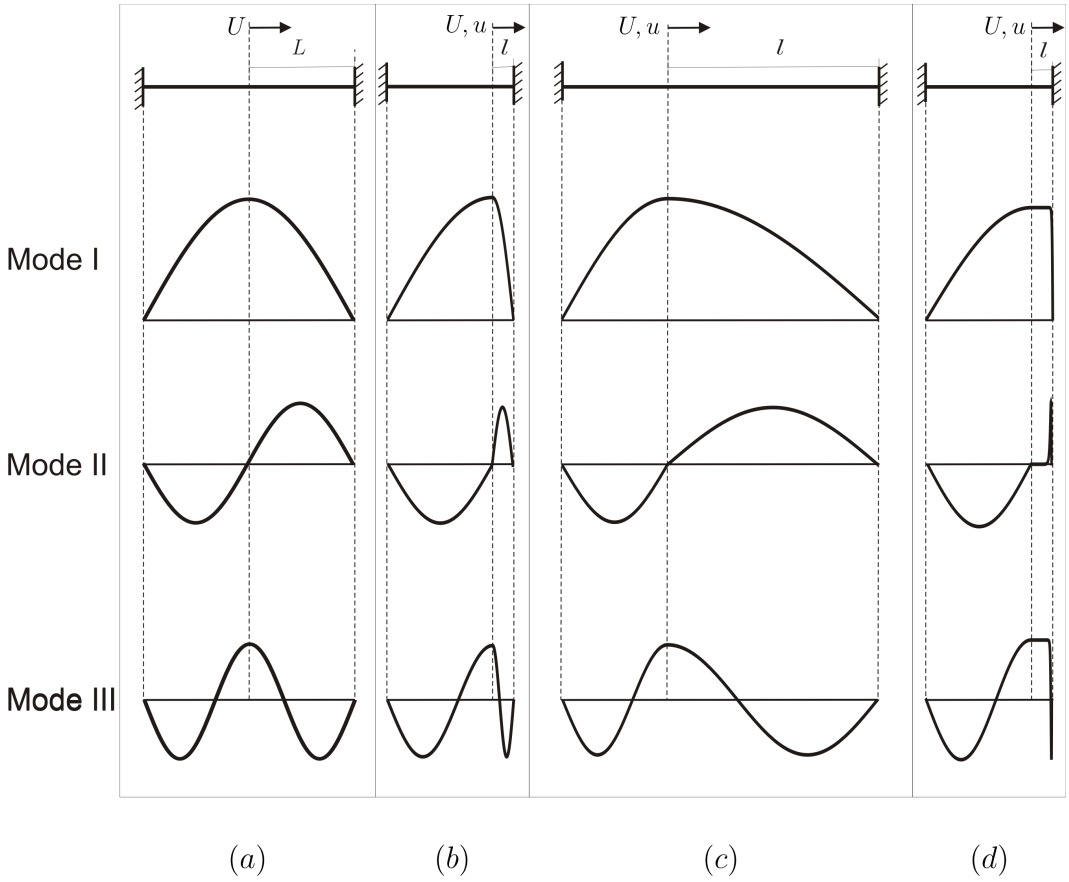


Figure 1: Eigenmodes for longitudinal waves in a rod. (a) Homogeneous rod of length $2L$. (b-d) The rods are subjected to a geometric transformation in the domain $0 \leq X \leq L$. (b) Rod subjected to the linear transformation (12) with $l/L = 0.2$. (c) Rod subjected to the linear transformation (12) with $l/L = 2$. (d) Rod subjected to the non linear transformation (13) with $l/L = 0.2$.

The eigenmodes for the homogeneous problem and the problem with transformation are given in Figure 1. We consider different transformation. We start with two linear transformations given by

$$g_a(x) = \frac{L}{l}x, \tag{12}$$

where we consider the two cases $l/L = 0.2$ and $l/L = 2$. Then, we show the results for the nonlinear transformation

$$g_b(x) = \frac{e^{-100x} - 1}{e^{-100l} - 1}L, \tag{13}$$

From the comparative analysis in Figure 1 it is shown that the eigenmodes are the same in the homogenous domain while $u(x) = u[G(X)] = U(X)$ in the non-homogenous ones. The eigenmodes in the transformed domains can be tuned by changing the transformation $g(x)$, which can be linear or not.

4.2 Flexural waves

Referring to the homogeneous problem as in Section 4.1, the solution of the equation of motion (2)a has the form

$$V(X) = C_1 e^{i\beta X} + C_2 e^{-\beta X} + C_3 e^{-i\beta X} + C_4 e^{\beta X}, \tag{14}$$

in which $\beta = [(\rho A)/(EJ)]^{1/4}\sqrt{\omega}$ is the frequency parameter. If we consider a simply supported beam the boundary conditions $V(-L) = V(L) = 0$ and $M_Z(-L) = M_Z(L) = 0$ lead to the eigenfrequencies

$$\omega = \left(\frac{n\pi}{2L}\right)^2 \sqrt{\frac{EJ_Z}{\rho A}}, \qquad (n \text{ positive integer number}). \tag{15}$$

The corresponding eigenmodes are $V(X) = \sin[n\pi(X + L)/(2L)]$.

We consider now the second structure having the same homogeneous properties for $-L \leq X \leq 0$, and the transformation $X \in [0, L] \rightarrow x \in [0, l]$. The transformation $g(x)$ has to satisfy the conditions $g(0) = 0$, $g(l) = L$, $g'(0) = g'(l) = 1$, $g''(0) = g''(l) = 0$ and $g'''(0) = g'''(l) = 0$. In order to satisfy the 8 conditions a polynomial of degree 7 has been implemented.

The problem is solved by

$$\begin{cases} V(X) = D_1 e^{i\beta X} + D_2 e^{-\beta X} + D_3 e^{-i\beta X} + D_4 e^{\beta X}, & \text{for } -L \leq X \leq 0, \\ v(x) = D_5 e^{i\beta g(x)} + D_6 e^{-\beta g(x)} + D_7 e^{-i\beta g(x)} + D_8 e^{\beta g(x)}, & \text{for } 0 \leq x \leq l. \end{cases} \tag{16}$$

The solution is found by satysfying the boundary conditions

$$\begin{cases} V(-L) = v(l) = 0, \\ M_Z(-L) = m_z(l) = 0, \end{cases} \tag{17}$$

and the interface conditions

$$\begin{cases} V(0) = v(0), \\ V'(0) = v'(0), \\ M_Z(0) = m_z(0), \\ T_Y(0) = r_y(0). \end{cases} \tag{18}$$

The system of boundary and interface conditions takes the form

$$\left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{A}_3 & \mathbf{A}_4 \end{array}\right] \begin{pmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \\ D_6 \\ D_7 \\ D_8 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{19}$$

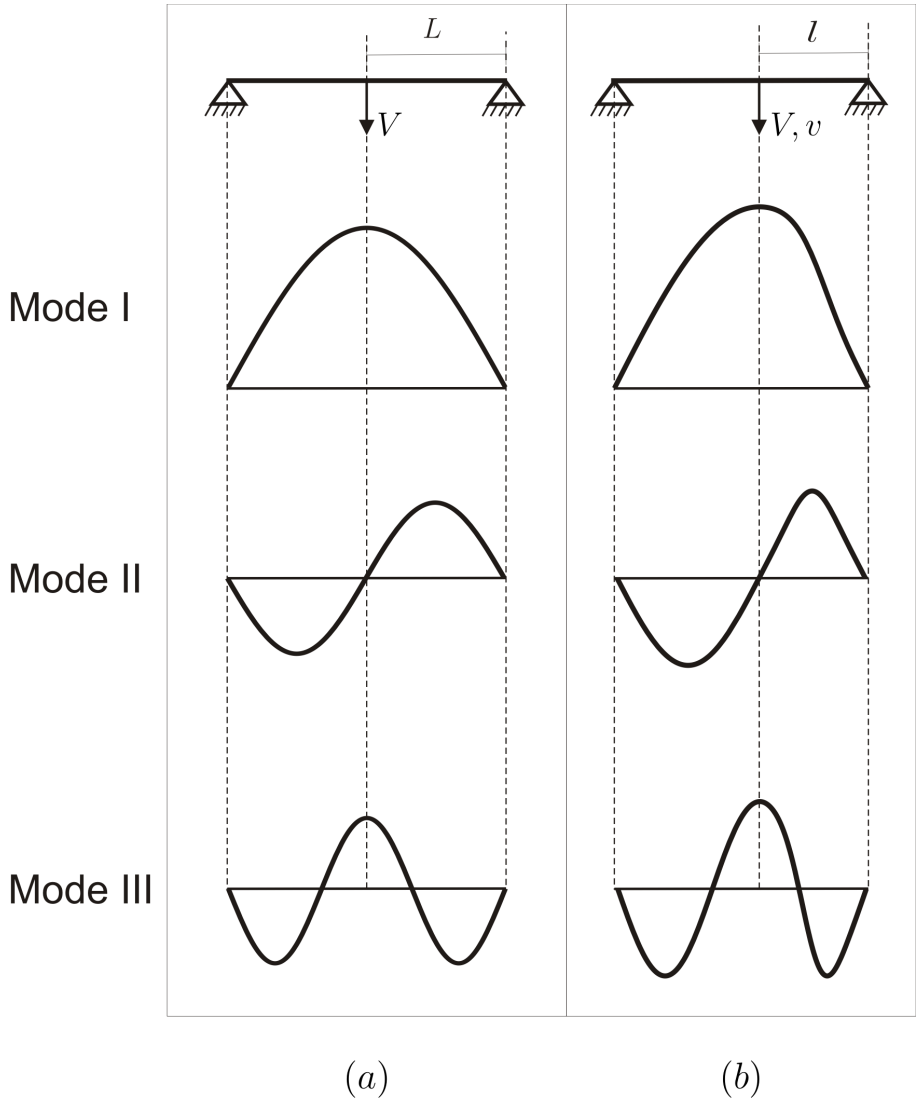


Figure 2: Eigenmodes for transverse waves in a beam. (a) Homogeneous beam of length $2L$. (b) The beam are subjected to a polynomial geometric transformation in the domain $0 \leq X \leq L$, with $l/L = 0.75$.

where

$$\mathbf{A}_1 = \begin{bmatrix} e^{-i\beta L} & e^{\beta L} & e^{i\beta L} & e^{-\beta L} \\ -\beta^2 e^{-i\beta L} & \beta^2 e^{\beta L} & -\beta^2 e^{i\beta L} & \beta^2 e^{-\beta L} \\ 1 & 1 & 1 & 1 \\ i\beta & -\beta & -i\beta & \beta \end{bmatrix} \tag{20}$$

$$\mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 \\ -i\beta & \beta & i\beta & -\beta \end{bmatrix} \tag{21}$$

$$\mathbf{A}_3 = \begin{bmatrix} EJ_Z\beta^2 & -EJ_Z\beta^2 & EJ_Z\beta^2 & -EJ_Z\beta^2 \\ iEJ_Z\beta^3 & EJ_Z\beta^3 & -iEJ_Z\beta^3 & -EJ_Z\beta^3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{22}$$

$$\mathbf{A}_4 = \begin{bmatrix} -EJ_Z\beta^2 & EJ_Z\beta^2 & -EJ_Z\beta^2 & EJ_Z\beta^2 \\ -iEJ_Z\beta^3 & -EJ_Z\beta^3 & iEJ_Z\beta^3 & EJ_Z\beta^3 \\ e^{i\beta g(l)} & e^{-\beta g(l)} & e^{-i\beta g(l)} & e^{\beta g(l)} \\ EJ_Z\beta^2 e^{i\beta g(l)} & -EJ_Z\beta^2 e^{-\beta g(l)} & EJ_Z\beta^2 e^{-i\beta g(l)} & EJ_Z\beta^2 e^{-\beta g(l)} \end{bmatrix} \tag{23}$$

Then, setting to 0 the determinant of the matrix in Eqs. (18), we obtain

$$256(EJ)^2\beta^{10}\sinh(2\beta L)\sin(2\beta L) = 0. \tag{24}$$

Eq. (24) gives exactly the same eigenfrequencies of the homogeneous system.

The eigenmodes for an homogeneous beam and a not homogenous beam are given in Figure 2 parts (a) and (b), respectively. Again, the displacement in the transformed domain are $v(x) = v[G(X)] = V(X)$.

5 CONCLUSIONS

Eigenfrequencies and eigenmodes for longitudinal and flexural waves in slender beam structures subjected to geometric transformation are detailed. It is shown that eigenfrequencies are invariant under transformation, while eigenmodes depend on the applied transformation law. Eigenfrequency analysis in finite system can help in the design and implementation of cloaking devices with particular attention to elastic waves.

REFERENCES

[1] J.B. Pendry, D. Schurig, D. Smith, Controlling electromagnetic fields. *Science*, **312 (5781)**, 1780-1782, 2006.

[2] D. Schurig, J.J. Mock, B.J. Justice, S.A. Cummer, J.B. Pendry, A.F. Starr, D.R. Smith, Metamaterial electromagnetic cloak at microwave frequencies. *Science*, **314(5801)**, 977–980, 2006.

[3] U. Leonhardt, Optical conformal mapping. *Science*, **312 (5781)**, 1777–1780, 2006.

[4] S.A. Cummer, D. Schurig, One path to acoustic cloaking. *New Journal of Physics*, **9(45)**, 45, 2007.

[5] A.N. Norris, Acoustic cloaking theory, *Proceedings of the Royal Society A*, **464 (2097)**, 2411–2434, 2008.

[6] H. Chen, C.T. Chan, Acoustic cloaking and transformation acoustics. *Journal of Physics D: Applied Physics*, **43(11)**, 113001, 2010.

[7] S. Guenneau, R.C. McPhedran, S. Enoch, A.B. Movchan, M. Farhat, N.P. Nicorovici, The colours of cloaks. *Journal of Optics*, **13 (2)**, 024014, 2011.

[8] M. Kadic, T. Bckmann, R. Schittny, M. Wegener, Metamaterials beyond electromagnetism. *Reports on Progress in Physics*, **76 (12)**, 126501, 2013.

[9] L. Zigoneanu, B.I. Popa, S.A. Cummer, Three-dimensional broadband omnidirectional acoustic ground cloak. *Nature Materials*, **13(4)**, 352–355, 2014.

[10] G.W. Milton, M. Briane, J.R. Willis, On cloaking for elasticity and physical equations with a transformation invariant form. *New Journal of Physics*, **8(10)**, 248, 2006.

[11] M. Brun, S. Guenneau, A.B. Movchan, Achieving control of in-plane elastic waves. *Applied Physics Letters*, **94 (6)**, 061903, 2009.

[12] A.N. Norris, A.L. Shuvalov, Elastic cloaking theory. *Wave Motion*, **48(6)**, 525–538, 2011.

[13] A.N. Norris, W.J. Parnell, Hyperelastic cloaking theory: transformation elasticity with pre-stressed solids. *Proceedings of the Royal Society A*, **468**, 2881–2903, 2012.

[14] M. Farhat, S. Guenneau, S. Enoch, Ultrabroadband elastic cloaking in thin plates. *Physical Review Letters*, **103 (2)**, 024301, 2009.

[15] M. Farhat, S. Guenneau, S. Enoch, A.B. Movchan, Cloaking bending waves propagating in thin elastic plates. *Physical Review B*, **79(3)**, 033102, 2009.

[16] N. Stenger, M. Wilhelm, M. Wegener, Experiments on elastic cloaking in thin plates. *Physical Review Lettres*, **108(1)**, 014301, 2012.

[17] M. Brun, D. Colquitt, I.S. Jones, A.B. Movchan, N.V. Movchan, Transformation cloaking and radial approximations for flexural waves in elastic plates. *New Journal of Physics*, **16**, 093020, 2014.

[18] D. Colquitt, M. Brun, M. Gei, A.B. Movchan, N.V. Movchan, I.S. Jones, Transformation elastodynamics and cloaking for flexural waves. *Journal of the Mechanics and Physics of Solids*, **72**, 131–143, 2014.

[19] I.S. Jones, M. Brun, N.V. Movchan, A.B. Movchan, Singular perturbations and cloaking illusions for elastic waves. in membranes and Kirchhoff plates *International Journal of Solids and Structures*, **69-70**, 498–506, 2015.

[20] D.J. Colquitt, I.S. Jones, N.V. Movchan, A.B. Movchan, M. Brun, R.C. McPhedran, Making waves round a structured cloak: lattices, negative refraction and fringes. *Proceedings of the Royal Society A*, **469(2157)**, 20130218, 2013.

[21] R.C. McPhedran, A.B. Movchan, N.V. Movchan, Platonic crystals: bloch bands, neutrality and defects. *Mechanics of Materials*, **41(4)**, 356–363, 2009.

[22] T. Antonakakis, R.V. Craster, High-frequency asymptotics for microstructured thin elastic plates and platonics. *Proceedings of the Royal Society A*, **468(2141)**, 1408–1427, 2012.

[23] M. Brun, G.F. Giaccu, A.B. Movchan, N.V. Movchan, Asymptotics of eigenfrequencies in the dynamic response of elongated multi-structures. *Proceedings of the Royal Society A*, **468(2138)**, 378–394, 2012.

[24] G. Carta, M. Brun, A.B. Movchan, Dynamic response and localization in strongly damaged waveguides. *Proceedings of the Royal Society A*, **470(2167)**, 20140136, 2014.

[25] R.C. McPhedran, A.B. Movchan, N.V. Movchan, M. Brun, M.J.A. Smith, Trapped Modes and Steered Dirac Cones in Platonic Crystals. *Proceedings of the Royal Society A*, **471(2177)**, 20140746, 2015.

ECCOMAS Congress 2016

VII European Congress on Computational Methods in Applied Sciences and Engineering

ECCOMAS Congress 2016

Proceedings of the VII European Congress on Computational Methods in Applied Sciences and Engineering

M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds)

First Edition, September 2016

ISBN: 978-618-82844-0-1

Institute of Structural Analysis and Antiseismic Research
National Technical University of Athens, Greece